


‘Toxic’ memes: A survey of computational perspectives on the detection and explanation of meme toxicities

Delfina S. Martinez Pandiani ^{b,a}, , Erik Tjong Kim Sang ^c, Davide Ceolin ^a

^a Centrum Wiskunde & Informatica, Amsterdam, 1098 XG, Netherlands

^b Institute for Logic, Language, & Computation, University of Amsterdam, 1098 XG, Netherlands

^c Netherlands eScience center, Amsterdam, 1098 XH, Netherlands

ARTICLE INFO

Keywords:

Internet memes
Toxicity
Information quality
Multimodal discourse

ABSTRACT

Internet memes are multimodal, highly shareable cultural units that condense complex messages into compact forms of communication, making them a powerful vehicle for information spread. Increasingly, they are used to propagate hateful, extremist, or otherwise ‘toxic’ narratives, symbols, and messages. Research on computational methods for meme toxicity analysis has expanded significantly over the past five years. However, existing surveys cover only studies published until 2022, resulting in inconsistent terminology and overlooked trends. This survey bridges that gap by systematically reviewing content-based computational approaches to toxic meme analysis, incorporating key developments up to early 2024. Using the PRISMA methodology, we extend the scope of prior analyses, resulting in a threefold increase in the number of reviewed works. This study makes four key contributions. First, we expand the coverage of computational research on toxic memes, reviewing 158 content-based studies, including 119 newly analyzed papers, and identifying over 30 datasets while examining their labeling methodologies. Second, we address the lack of clear definitions of meme toxicity in computational research by introducing a new taxonomy that categorizes different toxicity types, providing a more structured foundation for future studies. Third, we observe that existing content-based studies implicitly focus on three key dimensions of meme toxicity—target, intent, and conveyance tactics. We formalize this perspective by introducing a structured framework that models how these dimensions are computationally analyzed across studies. Finally, we examine emerging trends and challenges, including advancements in cross-modal reasoning, the integration of expert and cultural knowledge, the increasing demand for automatic toxicity explanations, the challenges of handling meme toxicity in low-resource languages, and the rising role of generative AI in both analyzing and generating ‘toxic’ memes.

1. Introduction

Memes, a concept introduced by Richard Dawkins in 1976, function as cultural replicators, akin to genes, rapidly transmitting ideas between human minds and shaping collective consciousness [1]. In contemporary discourse, the term “memes” has become synonymous with “internet memes” or “image memes”, which typically involve text–image pairs that spread quickly across digital platforms [2]. These internet memes are powerful tools for communication, enabling the rapid exchange of ideas, emotions, and cultural references within online communities. While often humorous and lighthearted, they have a significant impact on public and political discourse. For instance, alt-right ideologies on platforms like 4chan¹ and Encyclopedia Dramatica² have contributed to the formation of a subcultural, memetic language

that is closely linked to violent right-wing activism [3]. On the other hand, progressive leftist meme creators use memes as a tool for counter-narratives, dialectical engagement, and building solidarity within leftist communities [4]. The pervasive presence of memes on social media platforms has sparked interest and concern, as they play a key role in shaping digital culture and reflecting the collective consciousness of online societies.

In this context, memes can profoundly impact individuals and society, acting as insidious *flukes* that can “hijack and infect the brain” through repeated exposure to harmful ideas, particularly affecting those vulnerable to such influence [5]. This is especially true for internet memes, which, due to their accessibility and virality, contribute significantly to the spread of toxic ideologies and narratives and contribute to a climate of violence in public discourse [6]. The risks associated with toxic memes are extensive (see Table 1). Internet

* Corresponding author at: Institute for Logic, Language, & Computation, University of Amsterdam, 1098 XG, Netherlands.

E-mail address: d.s.martinezpandiani@uva.nl (D.S. Martinez Pandiani).

¹ <https://www.4chan.org>

² <https://encyclopedia.dramatica.online>

Table 1
Potential risks associated with toxic memes, identified in literature across various disciplines including computer science, human-computer interaction, internet pragmatics, multimedia, cybernetics, visual and media studies, and more.

General risk	Details	Citations
Violence in Public Discourse	Conveying toxic representations and messages, perpetuating harmful stereotypes via hate speech, harm, abuse, cyberbullying, offensiveness, and various other forms of toxicity.	[6,8–17]
Opinion Manipulation	Serving as potent tools for the dissemination of disinformation, propaganda, and trolling, leading to polarization and misunderstanding.	[18–24]
Psychological Impacts	Promoting groupthink and deindividuation, destructive thoughts and behaviors, desensitizing individuals to tragic news, fostering apathy, and exacerbating psychological distress.	[25–28]
Material-World Effects	Exerting tangible impacts of disinformation in election outcomes and instances of physical violence, contributing to the normalization of extremist behaviors.	[29–31]

memes can serve as conduits of behaviors that can often cross into illegal territory, exhibiting characteristics of incitement to violence [7], unlawful hate speech [8–10], harm [11–13], abuse [14], cyberbullying [15], offensiveness [16,17], opinion manipulation via dissemination of disinformation [18], propaganda [19–21], trolling [22,23], and the oversimplification of complex issues, leading to polarization and misunderstanding by presenting news as “open text”, open to multiple interpretations [24]. Moreover, exposure to toxic memes can have significant psychological effects, including fostering groupthink and deindividuation [25], promoting destructive behaviors [26], causing psychological distress through the normalization of dark humor, self-deprecating jokes, and derogatory slang [27], and desensitizing individuals to tragic news, leading to apathy toward important issues [28]. By trivializing violence and desensitizing individuals, toxic memes can normalize extremist behaviors, which may manifest in real-world actions, such as wearing meme-inspired clothing at extremist rallies or committing violent acts, like the Allen, Texas mass shooting [29]. Additionally, the spread of toxic memes influences not only individual behaviors but also broader societal events, including election outcomes [30] and instances of physical violence.

Addressing the proliferation of toxic meme culture is critical for online safety, requiring detection, analysis, and moderation strategies that can manage both the sheer volume and complexity of content. The multimodal nature of memes, along with fuzzy decision boundaries influenced by cultural context, biases, and the subtleties of harmful language, makes their assessment particularly challenging [32]. Currently, much of the moderation work is done by human crowdworkers in the Global South, especially in countries like the Philippines, India, and Kenya, who face mental health risks from prolonged exposure to toxic content [33–37]. This scale of content and the vulnerability of these workers highlight the growing need for automated solutions. Reflecting this shift, recent years have witnessed a surge in computational research on toxic memes, as indicated by the growing number of Scopus-indexed publications in computer science on this topic (Fig. 1).

Despite the rapid expansion of the field, systematic reviews of computational research on meme toxicities remain scarce. Prior surveys [11,38,39], have attempted to synthesize findings. Specifically, Afridi et al. (2020) [38] conducted the first comprehensive survey of automatic meme understanding, identifying challenges like defining hate in memes, distinguishing humor from hate, and categorizing memes into meaningful subtypes. They focused primarily on multimodal classification architectures and highlighted the lack of meme-specific datasets. Building on this foundation, Sharma et al. (2022) [11] introduced a typology for harmful memes, emphasizing categories such as hate, offensive, propaganda, and self-inflicted harm, stressing the need for cross-cultural research and broader dataset coverage. Most recently, Hermida and dos Santos (2023) [39] surveyed methodologies for detecting hateful memes, presenting a taxonomy based on machine

learning techniques, and underscored challenges like dataset biases and small sample sizes. While these surveys provide a valuable foundation, they do not evaluate the consistency between dataset labels and proposed taxonomies. Additionally, the field has grown significantly, with over 100 new Scopus-indexed studies published by early 2024, none of which have been systematically reviewed.

Building upon previous surveys, we systematically extend their scope using the PRISMA methodology, achieving a threefold result. Our review encompasses 158 research articles published between 2019 and March 2024, focusing on the automatic analysis of meme toxicity. We explicitly exclude studies on meme propagation and social network dynamics [40,41], as they fall outside the scope of this survey. Cross-referencing our findings with [11,38,39] revealed that prior surveys covered only 26 of the 158 papers identified, meaning 84% (132 out of 158) had not been previously reviewed in the context of toxic meme analysis (see Table 2; detailed tracking is available on the project’s GitHub page). As such, while these prior surveys laid an important foundation, they do not reflect the field’s rapid expansion, necessitating a more comprehensive and up-to-date review. Given the rapid growth of the field since the last surveys, our systematic review offers a more complete and up-to-date understanding of meme toxicity research. This review provides the following contributions:

- **Updated catalog of ‘toxic’ meme datasets**, labels, and research tasks (Section 4).
- **Standardized taxonomy of meme toxicity types** through the harmonization of meme toxicity labels used in the field (Section 5).
- **Multi-dimensionality framework** for describing key dimensions of meme toxicity, including intent, target, and conveyance tactics (Section 6).
- **Identification of research challenges and novel trends**, such as discourse-intensive multimodal reasoning, knowledge integration, and the use of generative AI (Section 7).
- **Pathways for advancement** of research in automatic meme toxicity analysis (Section 8).

This survey is structured as follows. In Section 2, we provide related background and an overview and definitions of (internet) memes and online (multimodal) toxicity. Section 3 outlines our methodological approach to selecting the papers surveyed. In Section 4, we provide a catalog of toxic meme datasets and associated label and task definitions. Section 5 introduces a novel taxonomy for categorizing the types of meme toxicity. Section 6 outlines a framework illustrating the relationships between dimensions of toxicity. Section 7 explores common challenges and recent trends in computational approaches to detect and interpret meme toxicities. Lastly, in Section 8, we identify key future research directions and we conclude in Section 9.

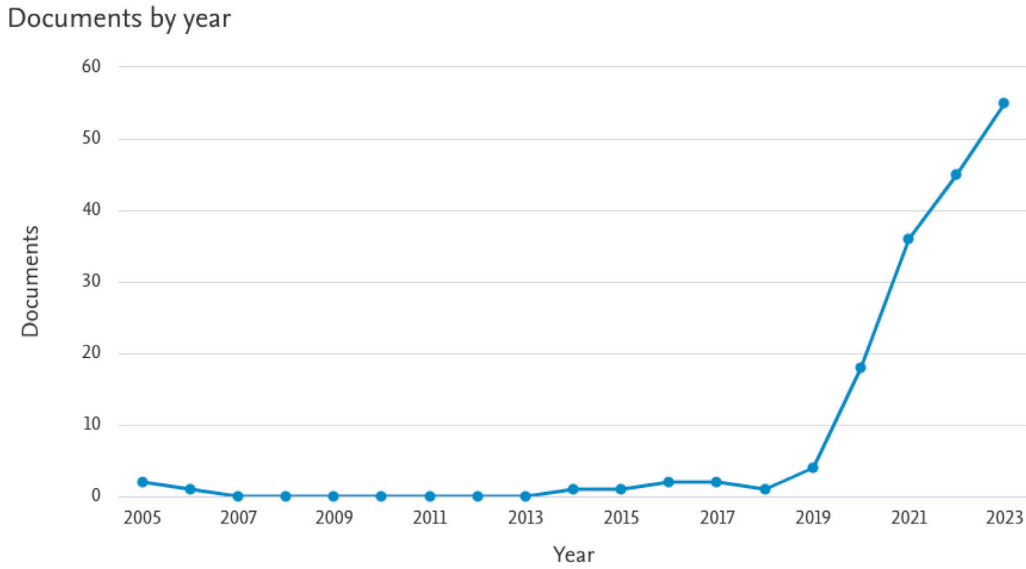


Fig. 1. Graph depicting the exponential increase in publications within the field of computer science, as indexed by SCOPUS, focusing on research related to toxic memes. The data was gathered using a query targeting specific keywords associated with meme toxicities [TITLE-ABS-KEY ((toxic OR harmful OR hateful OR unethical OR malicious OR malevolent OR offensive OR propaganda) AND meme) AND (LIMIT-TO (SUBJAREA, "COMP"))], as reported in Section 3.

Table 2

Number of papers out of the 158 identified manuscripts included in surveys about automatic detection of (toxic) memes.

Survey	Papers	Preprints	Total	Novel papers reviewed
Afridi et al. (2021) [38]	1	2	3	3
Sharma et al. (2022) [11]	13	6	19	16
Hermida & dos Santos (2023) [39]	3	10	13	7
This survey (2024)	119	39	158	132

2. Background

2.1. Defining (internet) memes

The concept of *memes* draws parallels to biological *genes* and evolution, describing ideas that self-replicate, evolve, and respond to selective Darwinian pressure, ultimately entering culture in ways similar to biological genes and potentially modifying human behavior [1]. However, the advent of the digital era has led to a resemantization of the term [42], with internet memes being intentional modifications of Dawkins' original concept, characterized by creative alterations rather than random mutation [43]. In the digital realm, memes are often defined as digital artifacts sharing common traits of content, form, or perspective, created by users and disseminated, imitated, or transformed via the Internet [44]. While a 'memetic construct' is the foundational structure comprising form, content, and perspective, a meme represents a specific instance or manifestation of a memetic construct, embodying a singular multimodal expression of a widespread cultural idea [45]. Thus, in this survey, we use the term 'meme' to specifically denote what others commonly refer to as 'internet' memes [46], 'visual memes' [47], or 'Image With Text' (IWT) memes [48].

Multimodality stands as a defining feature of memes, as memes rely on a combination of text and images to convey complex messages [45]. Typically, memes consist of an image paired with short text, allowing for easy sharing on social media [49]. They blend visual and verbal elements to convey humor, irony, or sarcasm, often referencing cultural symbols or events [42,50]. Studies computationally operationalizing the term 'meme' have primarily utilized such visuo-linguistic association analysis to distinguish memes from non-meme images and relied on implicit judgement or crowdsourced annotations for differentiation [51]. Recent research [52] employs multi-channel convolutional neural networks to distinguish memes from not only

Table 3

Features of (internet) memes, along with their corresponding descriptions and relevant citations.

Feature	Description
Multimodal	Combine visual and language information creatively [42,46]
Succinct	Spread complex messages with a minimal information unit that connects virtual circumstances to real ones [46]
Fluid	Subject to variations and alterations [46]
Anomalous Juxtaposition/Incongruity	Leverage lack of relevance in the arrangement of textual and/or visual constituents to produce unexpected outcomes [50]
Intertextual	Reference popular culture, symbols, artifacts, or events that hold meaning within the community of references [46,50]
Relatable/Tacit Background	Rely on viewer's familiarity with certain contextualized aspects of the world [53], shared knowledge, and implicit cultural references [54]

photographs but also other image-with-text (IWT) formats, such as advertisements, movie posters, online news articles, or screenshots of posts, commonly circulated online. Table 3 summarizes key features of internet memes and associated citations.

2.2. Defining online toxicity

Online toxicity typically refers to negative behaviors on the internet that damage others' or even one's own self-image, hindering personal growth [55]. This definition is adopted by researchers studying text-image combinations [56] and those exploring 'toxic memes' [3].

Understanding online toxicity is challenging due to its multifaceted nature, including fine-grained categories and overlapping terms, requiring models capable of recognizing various aspects of toxic behavior [57]. Defining and detecting online toxicity, particularly from a computer science perspective, is further complicated by challenges in annotating data and developing machine learning models to identify toxicity types and relationships [58]. For instance, sharing misinformation on social media is associated with harmful language, highlighting the importance of integrating research on two types of toxicity (misinformation and harmful language) usually studied separately [59].

2.2.1. Textual toxicity

Much of the research on online toxicity has focused on textual data, with extensive studies on detecting toxicity in texts [60] and tasks such as toxic comment classification, hate speech detection, and identification of offensive language. Different taxonomies have been proposed to categorize abusive language, distinguishing between abusive content directed at individuals or groups, and between explicit and implicit abusive content [61–63]. Even within ‘hate’ speech, there are various definitions and fine-grained labels, and thus a need for comprehensive datasets to train robust models for combating hate speech effectively [64]. Some works follow a three-level taxonomy considering the type and target of offense [63], while other toxic comment detection systems follow a multi-label classification framework, with comments labeled as ‘toxic,’ ‘severe toxic,’ ‘insult,’ ‘threat,’ ‘obscene,’ and ‘identity hate’ [65]. Other datasets use labels like ‘hate speech,’ ‘offensive but not hate speech,’ and ‘neither offensive nor hate speech,’ highlighting biases and challenges in classifying offensive language in short-form content like tweets [66].

The differentiation of concepts associated with toxic speech presents complexities, often characterized by contentious definitions [67]. Efforts to delineate the hierarchy of hate speech concepts within computer science literature reveal fuzzy boundaries between toxicity-related concepts [68], and instances of interchangeability and even conflicting hierarchical relations. For instance, some perspectives consider toxicity as a subset of hate speech [69]. In response to these challenges, some initiatives aim to harmonize toxicity labels for textual content. For instance, in [70], researchers analyzed six publicly available datasets related to hate speech, aggression, and toxicity in text. Their objective was to standardize categories to ensure consistency and comparability across datasets. This involved merging related categories, clarifying ambiguous labels, and aligning similar concepts under common headings. The resulting taxonomy identifies various types of toxicity and recommends considering unique definitions and contexts before merging terms.

2.2.2. Multimodal toxicity

Exploring multimodal toxicity from a computational perspective has not received as much attention as textual (unimodal) toxicity, yet understanding its dynamics is increasingly crucial [71]. Recent studies have begun to bridge this gap, offering varied perspectives on toxic multimodal content found online. This section provides an overview of these works, each presenting diverse taxonomies summarized in Table 4, with varying levels of granularity and emphasis on different facets of harmful content. For instance, Banko et al. [72] provide a taxonomy covering hate, harassment, self-inflicted harm, ideological harm, and exploitation, while Nakov et al. [73] offer a simpler list of harmful categories. Halevy et al. [74] focus on violations for malicious purposes, including misinformation and community standard violations such as hate speech and crimes. Pramanick et al. [49] distinguish between hateful, offensive, and generally harmful memes, while Sharma et al. [11] present a taxonomy for internet memes with categories like hateful, offensive, propaganda, harassment/cyberbullying, violence, self-inflicted harm, and exploitation. As seen in Table 4, there are commonalities and differences across these taxonomies. For instance, both [72,73] cover hate speech, harassment, and violence. However, [72] offers more

detailed subcategories like doxing and identity attack, whereas [73] includes a broader range, such as dangerous organizations/people and glorifying crime.

To clarify the taxonomical relationships among the toxicities in Table 4, we constructed a Venn diagram Appendix illustrating their intersections and semantic overlaps. This analysis revealed the nuanced complexity of defining toxic or harmful multimodal content. For instance, many behaviors classified as misbehavior by platforms – such as nudity – are restricted despite not being inherently toxic or harmful. We identified a range of harmful content types, including ideological harm, hatefulness toward protected groups, and harassment, underscoring the multifaceted nature of online toxicity. Consistent with [68] on textual toxicities, our findings highlight the difficulty of drawing clear boundaries between toxicity categories. Additionally, this exploration uncovers emerging toxic content specific to multimodal contexts and harmful behaviors that may not necessarily depend on multimodality.

2.3. Related surveys on multimodal and textual toxicity

Other surveys on multimodal disinformation and hate, though not specifically focused on memes, offer valuable insights into broader issues of multimodal toxicity. For instance, [75] review computational approaches to multimodal disinformation and harm across various content types—text, speech, images, videos, and network data. Their findings highlight the importance of explainability in model interpretation, the need to account for cultural and personal factors beyond content and network signals, and the potential of knowledge-based methods for factuality checking. Similarly, [69] examine hate speech detection in multimodal and multilingual contexts, covering diverse content types – including text, images, and videos – across platforms such as private messages, comments, ads, and user profiles. They emphasize proactive moderation strategies, advocating for blocking or reporting mechanisms, responsible content sharing, and stronger policy frameworks to curb abusive behavior. A more recent survey, [76], primarily focuses on textual hate speech but underscores the role of multimodal features. It critiques the prevalent reliance on surface-level cues such as word frequency, punctuation, capitalization, and specific keywords, arguing that such approaches often fail to capture context and meaning. The study also points out the scarcity of comparative analyses and stresses the importance of open-source code and dataset availability for evaluation. Though not addressing multimodal aspects, two additional surveys on textual toxicity are noteworthy. [77] introduce a hierarchical taxonomy of offensive language, distinguishing between explicit and implicit expressions. Meanwhile, [78] explore the subjectivity of toxicity detection, biases in datasets, and the influence of content source and topic on dataset characteristics, alongside challenges in collecting toxic comments.

3. Methodology

Selection of databases. We used Scopus³ and Web of Science (WOS)⁴ for our literature search because they are two of the largest and most reputable databases, indexing a wide range of high-quality journals across various disciplines and now including preprints.

Inclusion of preprints. Acknowledging the novelty of this research field and the rapid pace of publication, we included preprints in our methodology. This decision aligns with the approach taken by the three previous surveys on this topic, emphasizing the importance of capturing the latest developments in this rapidly evolving field.

³ <https://www.scopus.com/>

⁴ <https://www.webofscience.com/>

Table 4
Comparison of categories and subcategories in different taxonomies of toxic content.

Taxonomy	Source	Focus	Top-level	Second level	Granular
Banko et al. 2020 [72]	Research	Harmful content	Hate/ Harassment	Doxing	
				Identity Attack	
				Identity Misrepresentation	
				Insult	
				Sexual Aggression	
				Threat of Violence	
			Self-Inflicted Harm	Eating Disorder Promotion	
				Self-Harm	
			Ideological Harm	Misinformation	
				Extremism, Terror & Org. Crime	White Supremacist
Nakov et al. 2021 [73] /Arora et al. 2023 [7]	Social Media	Policy clauses	Violence	Adult Sexual Services	
				Child Sexual Abuse Materials	
				Scams	
Pramanick et al. 2021 [49]	Research	Harmful memes	Hateful		
			Offensive		
			Other (Generally) Harmful		
Halevy et al. 2022 [74]	Social Media	Violations /Malicious Purposes	Misinformation		
			Community Standards Violations	Hate Speech	
				Crimes	Selling Illegal Drugs
					Sex Trafficking
Sharma et al. 2022 [11]	Research	Harmful memes	Hateful	Child Exploitation	
				Doxxing	
				Identity Attack	
				Identity Misrepresentation	
				Insult	
				Racist	
				Misogynistic/Sexist	
				Sexual Aggression	
			Offensive	Extremism, Terrorism & Organized Crime	
			Propaganda		
			Harassment/Cyberbullying		
			Violence		
			Self-inflicted Harm	Eating Disorder Promotion	
				Self-harm	
			Exploitation	Adult Sexual Service	
				Child Sexual Abuse Material	
				Scams	

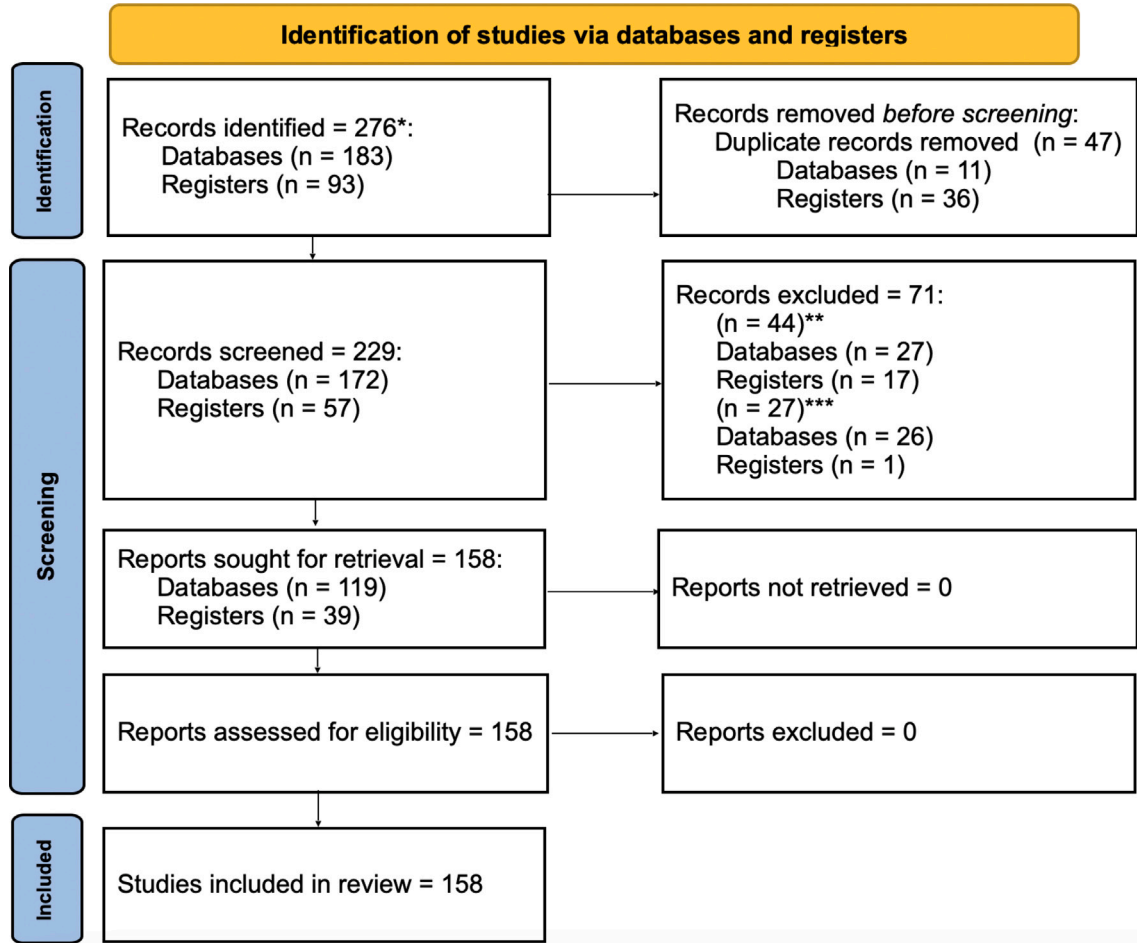


Fig. 2. PRISMA 2020 flow diagram for systematic reviews on *SCOPUS and Web of Science (WOS) databases. Registers refers to SCOPUS preprints. Records excluded due to: ** Topic Non-Relevance. *** Computational Non-Relevance.

Identification of studies to review. Inspired by PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses),⁵ we used a structured approach to select the manuscripts. Fig. 2 presents a diagram of our selection process following the PRISMA 2020 guidelines, which outlines the process of manuscript selection.

Database query. On March 6th, 2024, we queried the databases with a manually crafted key.⁶ In SCOPUS, results were confined to “Computer Science”, while in WOS, results were confined to “Computer Science Artificial Intelligence” or “Computer Science Information Systems”, yielding 276 records (183 from databases and 93 from registers).

Record screening. We removed 47 duplicate preprints, resulting in 229 manuscripts. We manually screened these based on abstracts and titles, excluding 71 irrelevant to our computational focus on toxic memes. We flagged 27 of those for discussion in the related work section, leaving 158 for further evaluation.

Retrieval assessment for eligibility. We retrieved all 158 records (119 peer-reviewed articles and 39 preprints). All 158 manuscripts were manually assessed and were considered eligible.

4. Toxic meme datasets

We systematically reviewed the selected studies, documenting the datasets used in their computational analyses. For each dataset, we

recorded key attributes, including size (number of memes), language, source (e.g., social media platforms, Google Images), computational task(s), annotated features, classification labels, task definitions (e.g., binary, single-label multi-class, multi-label multi-class), and baseline macro F1 scores (see Table 6). This analysis identified 34 datasets (see Table 5, additional details available on GitHub.⁷).

4.1. Dataset characteristics

Dataset size. Our survey reveals a diverse spectrum of dataset magnitudes, spanning from a few hundred to tens of thousands of memes. Among the smaller datasets are Derogatory Facebook-Meme [80] with 650 memes and MultiOFF [104] with 743 memes. Conversely, larger datasets include Facebook Hateful Memes (HM) [53] with 9540 memes, alongside MAMI [105], Memotion 1 [17], Memotion 2 [96], and MET-Meme [97], all hovering around the 10,000 mark. Notably, Innopolis Hateful Memes [93] boasts a substantial 23,000 memes. These findings underscore a considerable variance in dataset sizes within the domain. The distribution of dataset magnitudes demonstrates that most datasets lie within the range of 2500 to 10,000 memes (see Fig. 3, top).

Language. The majority of datasets primarily contain memes in English (see Fig. 3 (middle)), with nearly 75% exclusively featuring English-language memes [16,17,19,23,49,53,79–82,84–91,93–96,98,101,102,104]. However, we also observe a recent surge in datasets incorporating Asian languages such as Hindi [83,99,100], Bengali [14], and Chinese [97]. Notably, many of these datasets feature “code-mixed”

⁵ <https://www.prisma-statement.org/>

⁶ TITLE-ABS-KEY ((toxic OR harmful OR hateful OR unethical OR malicious OR malevolent OR offensive OR propaganda) AND meme) AND (LIMIT-TO (SUBJAREA, “COMP”))

⁷ https://github.com/delfimpandiani/toxic_memes

Table 5

Overview of the 34 datasets containing toxic memes. For each dataset, we specify the year of introduction, the corresponding manuscript, the languages included, the number of memes, the main focus/task, and the sources of the memes. Abbreviations: SE — search engines; IP — image hosting platforms; SM — social media platforms; MM — meme-specific resources.

Dataset	Year	Language	Size	Main focus	Sources			
					SE	IP	SM	MM
AOMD Gab [16]	2021	English	1965	offensive meme detection			✓	
AOMD Reddit [16]	2021	English	1094	offensive meme detection			✓	
BanglaAbuseMeme [14]	2021	Bengali, English	4043	abusive meme detection	✓		✓	
CrisisHateMM [79]	2022	English	4700	hateful meme detection			✓	
Derogatory Fb-Meme [80]	2023	Hindi, English	650	derogatory meme detection			✓	
DisinfoMeme [81]	2022	English	1170	disinformation meme detection			✓	
ELEMENT [82]	2022	English	7912	unethical meme detection	✓		✓	✓
Emoffmeme [83]	2023	Hindi	7500	offensive meme detection	✓			
Ext-Harm-P [84]	2023	English	4446	harmful reference detection	✓		✓	
Facebook Hateful [53]	2022	English	9540	hateful meme detection			✓	
FAME dataset [85]	2020	English	1000	fake meme detection	✓			
Fine grained HM [86]	2020	English	9540	fine-grained hate meme detection			✓	✓
GOAT-Bench [87]	2021	English	6626	abusive/toxic meme detection	✓	✓	✓	✓
Harm-C (HarMeme) [49]	2024	English	3544	harmful meme detection	✓		✓	
Harm-P [88]	2021	English	3552	harmful meme detection	✓		✓	
Hate Speech in Pixels [89]	2021	English	5030	hateful meme detection	✓		✓	
HatReD [90]	2019	English	3228	hateful meme explanation			✓	✓
HVVMemes [91]	2023	English	7000	entity roles in harmful memes	✓		✓	
Indian Political [92]	2022	Hindi, English	1218	hateful meme detection	✓			
Innopoliis [93]	2022	English	23 000	hateful meme detection	✓		✓	✓
KAU-Memes [94]	2022	English	2582	offensive meme detection			✓	
Meme-Merge [95]	2023	English	10 000	offensive meme detection	✓		✓	
Memotion 1 [17]	2023	English	10 000	offensive meme detection	✓			
Memotion 2 [96]	2020	English	10 000	offensive meme detection		✓	✓	
MET-Meme [97]	2022	English, Chinese	10 045	offensive meme detection	✓		✓	
Misogynistic-MEME [98]	2022	English	800	misogynous meme detection			✓	
MultiBully [99]	2022	Hindi, English	5854	cyberbullying meme detection			✓	
MultiBully-Ex [100]	2022	Hindi, English	3222	cyberbullying explanation			✓	
MAMI [101]	2024	English	10 000	misogynous meme detection		✓	✓	✓
MultiOFF [102]	2022	English	743	offensive meme detection			✓	
Pol_Off_Meme [103]	2020	Hindi, English	7500	offensive meme detection	✓			
SemEval-2021 Task 6 [19]	2024	English	950	propagandistic detection			✓	
TamilMemes [104]	2021	Tamil	2969	troll meme detection		✓	✓	
TrollsWithOpinion [23]	2020	English	8881	troll meme detection	✓			

memes, which blend these languages with English. Interestingly, beyond English and select Asian languages, no datasets have been utilized in the studies we surveyed that incorporate other widely spoken languages, such as Spanish.

Data origin. Examining the sources of meme collection revealed a diverse range of platforms, which can be categorized into four macro-categories: social media platforms (e.g., Facebook, WhatsApp), search engines (e.g., Google, Bing), image hosting platforms (e.g., Imgur, Pinterest), and dedicated meme creation and sharing resources (e.g., KnowYourMeme, 9gag) (see Fig. 3 (bottom)). Social media platforms emerged as the most prominent sources, with Facebook and Reddit being the most used sources, followed by search engines, with Google being the most utilized search engine. Additionally, platforms such as Pinterest and Imgur were used for meme acquisition. Of particular interest was the utilization of meme-specific platforms like Memegenerator, KnowYourMeme, and 9gag, highlighting the importance of dedicated meme communities in the proliferation of memes. We provide detailed source information for each of the datasets in Table A.14 in Appendix. We also observed instances where new datasets were derived by augmenting previously introduced datasets: Ext-Harm-P [84] is derived from Harm-P [88], while Fine grained HM [86] and HatReD (Hateful meme with Reasons Dataset) [90] are derived from Facebook Hateful Memes (HM) [106]. Moreover, other datasets were formed by merging existing datasets: Meme-Merge [95] is a merge of MET-Meme [97], Memotion 1 [17], and Memotion 2 [96], while GOAT-Bench [87] was created by amalgamating data from several sources

including Facebook HM [106], MAMI [101], MultiOFF [104], Harm-C [49], and Harm-P [88]. This analysis highlights the dynamic and varied meme distribution and consumption landscape across numerous online platforms. It also reveals a growing trend toward reusing existing datasets. Also, it emphasizes the evolution of dataset creation in meme research, showcasing how existing resources are being adapted to address new research questions.

4.2. Dataset target labels/annotations

We analyzed the annotation guidelines of each dataset to determine the types of labels assigned to toxic memes and their computational relevance. Table 7 summarizes the annotated features across multiple datasets, with columns representing datasets and rows representing features. Check marks indicate which features are labeled in each dataset, providing a clear comparison of coverage. Our analysis revealed significant variation in annotation methods, with datasets employing binary, single-label multi-class, or multi-label multi-class schemes, as detailed below.

Abusiveness is indicated in the BanglaAbuseMeme dataset [14] with each meme labeled as either abusive or not abusive in a binary (yes/no) format.

Aggressiveness is evaluated in the Misogynistic-MEME (MM) dataset [98] using a binary approach.

Attack types are annotated in various datasets. Fine-grained HM [86] adopts a multi-label multi-class approach, delineating different attack types within memes, including dehumanizing, inferiority, inciting

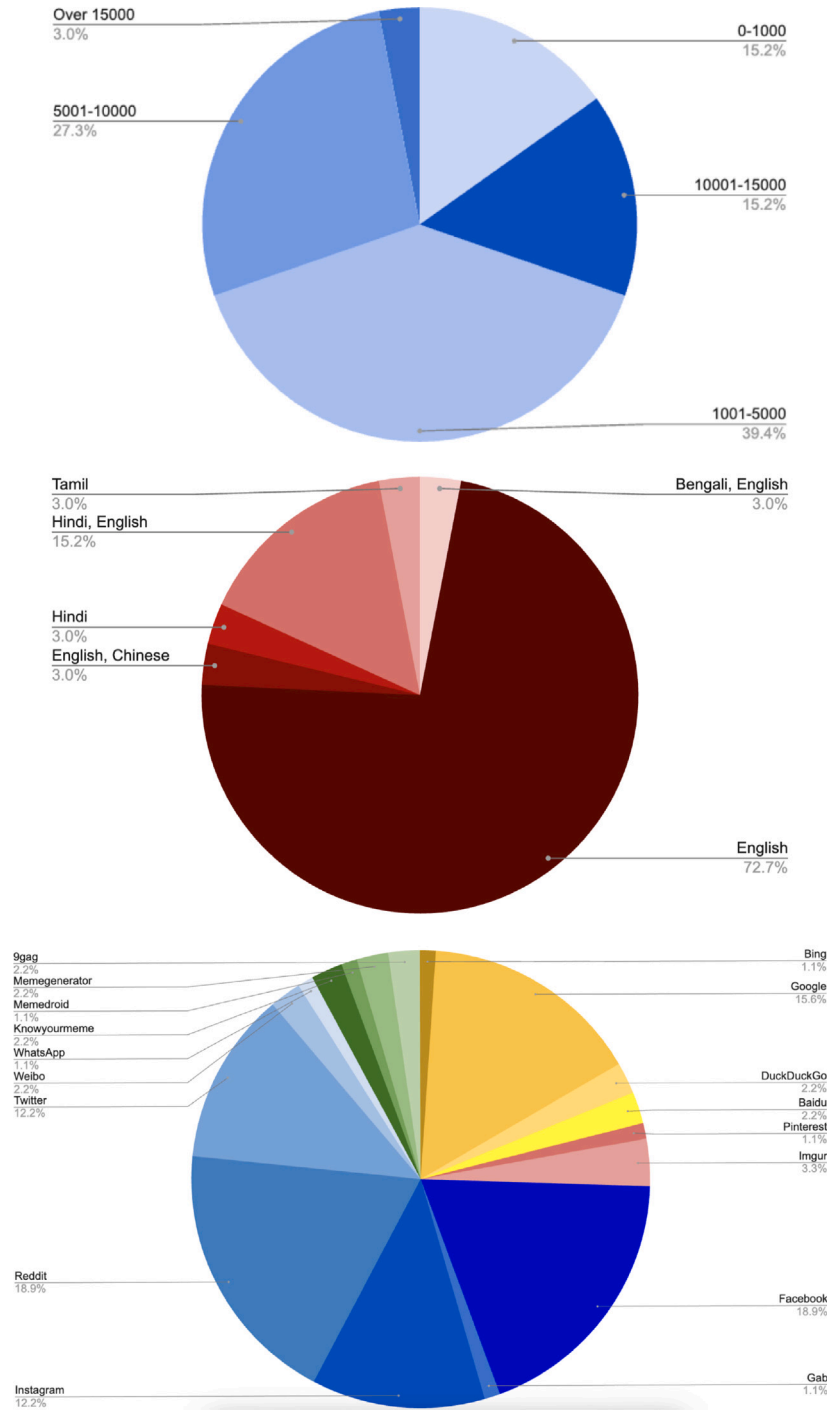


Fig. 3. Top: Dataset sizes—40% contain 1K–5K memes, 30% have 5K–10K, and under 3% exceed 15K. Middle: 75% of datasets are in English; others are in Hindi, Bengali, Tamil, or code-mixed. Bottom: Meme sources—social media dominate, followed by meme sites, search engines, and image hosts. Over half include social media content.

violence, mocking, contempt, slurs, and exclusion. Similarly, the Multimedia Misogyny Dataset (MAMI) [101] provides attack type labels specific to misogyny in a multi-label multi-class format, encompassing general misogyny, shaming, stereotype, objectification, and violence.

Bullying categorization is found in both the MultiBully dataset [99] and its extension, MultiBully-Ex [100], with memes labeled as either bully or non-bully using binary labels.

Disinformation presence is annotated in the DisinfoMeme dataset [81] through a binary approach, with memes labeled as either containing disinformation or not (yes/no).

Emotion is a feature of memes that is labeled across datasets in various ways. Some datasets such as MultiBully [99], and MultiBully-Ex [100] utilize a single-label multi-class approach, tagging memes

with emotions like joy, sadness, fear, surprise, anger, disgust, anticipation, trust, or ridicule. Similarly, MET-Meme [97] employs a single-label multi-class system, categorizing memes with emotions such as happiness, love, anger, sorrow, fear, hate, or surprise. Conversely, Emoffmeme [83] and Pol_Off_Meme [103] utilize a multi-label multi-class approach, encompassing annotations for fear, neglect, irritation, rage, disgust, nervousness, shame, disappointment, envy, suffering, sadness, joy, pride, and surprise. In contrast, Memotion 1 [17] and Memotion 2 [96] offer annotations for sarcastic, humorous, motivational, and offensive emotions within a multi-label multi-class framework. The term **Sentiment** is used to describe this same aspect of memes in Emoffmeme [83] which uses a multi-label multi-class paradigm for

including annotations for fear, neglect, irritation, rage, disgust, nervousness, shame, disappointment, envy, suffering, sadness, joy, pride, and surprise. In contrast, most other datasets utilize sentiment labels in a Likert scale, single-label multi-class approach. For example, Memotion 1 [17] and Memotion 2 [96] categorize sentiments as very positive, positive, neutral, negative, and very negative. Similarly, BanglaAbuseMeme [14], MultiBully [99], and MultiBully-Ex [100] follow a single-label multi-class scheme with simpler labels: positive, neutral, and negative. Furthermore, the Derogatory Facebook-Meme dataset [80] employs a binary annotation to indicate negative sentiment (yes/no).

Explanation annotations are present in a couple of datasets: HatRED (Hateful meme with Reasons Dataset) [90] includes human-provided textual explanations. MultiBully-Ex [100] provides both textual explanations as textual rationales (highlighted words or phrases) and visual explanations via visual masks (image segmentation) for its categorization of memes as harmful.

Fake/misattribution is annotated in the FAME dataset [85], categorizing memes as either fake or real (binary).

Harmfulness is annotated in varied ways across the datasets. The GOAT-Bench dataset [87] uses a binary system to indicate harmfulness, labeling each instance as either harmful or not (yes/no). Several other datasets, including Harm-P [88], Harm-C (also known as HarMeme) [49], MultiBully [99], and MultiBully-Ex [100], categorize the degree of harmfulness using a single-label multi-class system. These datasets label content as very harmful, partially harmful, or harmless, providing a more nuanced understanding of harmfulness. Additionally, the Ext-Harm-P dataset [84] focuses on the harmfulness of the reference to a social entity. It employs a binary system, labeling references as either harmful or harmless, thus specifically targeting the impact on social entities.

Hate speech is annotated in a varied way across different datasets. Most datasets, including Hate Speech in Pixels [89], Facebook Hateful Memes (HM) [53], Fine grained HM, [86], CrisisHateMM [79], Innopolis Hateful Memes [93], GOAT-Bench [87], and Derogatory Facebook-Meme [80], use a binary labeling system to identify hate speech, categorizing content simply as either hateful or not (yes/no). However, the Indian Political Memes (IPM) dataset [92] employs a more nuanced approach with single-label multi-class annotations among non-offensive, hate-inducing, and satirical. Additionally, CrisisHateMM [79] includes an extra dimension by labeling the direction of hate. This dataset differentiates between directed hate and undirected hate, using a binary system to specify whether the hate is aimed at a specific target or more general in nature.

Humor is assessed in Memotion 1 [17] and Memotion 2 [96] datasets on a continuum scale, spanning from not funny to hilarious, within a single-label multi-class framework.

Intention is addressed in a single-label multi-class format in the MET-Meme dataset [97], offering labels such as interactive, expressive, purely entertaining, offensive, or other.

Irony is annotated as a binary label (yes/no) in the Misogynistic-MEME dataset [98], while **Sarcasm** is similarly represented in binary format (yes/no) in BanglaAbuseMeme [14], MultiBully [99], MultiBully-Ex [100], and GOAT-Bench [87]. In contrast, sarcasm in Memotion 1 [17] and Memotion 2 [96] is classified along a continuum, with labels ranging from “not sarcastic” to “very twisted”, within a single-label multi-class framework.

Metaphors in toxic memes are addressed in the MET-Meme dataset [97] annotates metaphor, which includes binary labels for **metaphorical expression** (metaphorical/literal), and also employs a single-label multi-class approach to classify metaphor types into text dominant, image dominant, or complementary.

Misogyny is labeled in a binary manner (yes/no) across multiple datasets, including the Misogynistic-MEME (MM) dataset [98], Multimedia Misogyny Dataset (MAMI) [101], and GOAT-Bench [87].

Modality Class in Memotion 2 [96] employs a single-label multi-class approach to label memes based on their modality, allowing them to be categorized as either image and text, image only, or text only.

Motivation is annotated in Memotion 2 [96] and Memotion 1 [17] datasets, distinguishing memes as either motivational or not motivational in a binary manner.

Offensiveness. Some datasets, such as MultiOFF [102], AOMD Gab [16], AOMD Reddit [16], Pol_Off_Meme [103], EMOFFMEME [83], KAU-Memes [94], TrollsWithOpinion [23], and GOAT-Bench [87], use a binary labeling system (yes/no) to indicate offensiveness. Other datasets, such as Memotion 1 [17], Memotion 2 [96], and MET-Meme [97], assess offensiveness on a single-label multi-class scale, ranging from not offensive to hateful offensive. Meme-Merge [95] also adopts a degree-based labeling system, with offensiveness categorized from non-offensive to very offensive. Pol_Off_Meme [103], and EMOFFMEME [83] include binary labels for the explicitness of offensiveness.

Opinion manipulation is addressed in TrollsWithOpinion [23], which employs a binary labeling approach, categorizing memes as either involving opinion manipulation or not. Additionally, it provides labels for opinion manipulation types: political, product, or other, in a single-label multi-class manner.

Political attributes are identified in Pol_Off_Meme [103] through binary classification, distinguishing memes as either political or not political.

Profanity labels are included in the Derogatory Facebook-Meme dataset [80], while **vulgarity** labels are present in BanglaAbuseMeme [14]. Both are annotated in a binary framework of yes/no.

Propagandistic techniques are annotated in SemEval-2021 Task 6 [19] using a multi-label multi-class approach. This encompasses a wide range of techniques, including Loaded Language, Name Calling/Labeling, Smeared Doubt, Exaggeration/Minimization, Slogans, Appeal to Fear/Prejudice, and more. Each meme can be assigned multiple labels corresponding to the propagandistic techniques it employs.

Target of attack is commonly annotated across datasets. Derogatory Facebook-Meme [80] employs a binary label (whether a meme is targeted or not), but most datasets annotate the target in a single-label multi-class manner, with categories including individual, community, organization, and society—as in Harm-P [88], Harm-C (also known as HarMeme) [49], and CrisisHateMM [79]. Others provide more specific community targets: BanglaAbuseMeme [14] uses single-label multi-class labeling with gender, religion, national origin, individual, political, social sub-groups, and others. Fine-grained HM [86] employs a multi-label multi-class approach, including labels for protected categories such as religion, race, sex, nationality, and disability. HVMemes [91] utilizes a single-label multi-class framework to annotate each entity’s role (villain, victim, hero, or other).

Troll is annotated as a dimension in both TamilMemes [104] and TrollsWithOpinion [23] using a binary labeling.

Unethical is an aspect of memes annotated in the ELEMENT dataset [82] using a binary (yes/no) labeling system.

5. Defining meme toxicities: A taxonomy

Our review of the 158 research papers revealed a wide range of terms used to describe toxicity in memes, suggesting that multiple forms of toxicity are being explored computationally (see summary Table 9). We manually examined each paper to extract explicit labels for the types of toxicity under study.

5.1. Meme toxicities identification and definitions

We identified 12 overarching meme toxicity terms: *abusive*, *cyberbullying*, *derogatory*, *disinformation*, *fake*, *harmful*, *hateful*, *misogynous*, *offensive*, *propaganda*, *troll*, and *unethical*. We then compiled the definitions associated with each term and harmonized them into

Table 8. Our analysis of toxicity-related terms showed that some terms were used interchangeably in the literature. For example, *abusive* was used synonymously with *offensive* by [107] and with *harmful*

Table 6
Summary of Section 4, Toxic meme datasets.

Topic	Count	Most frequent	Reference
Datasets surveyed	34		Table 5
Labels/aspects	25	offensiveness (35%)	Table 7
Size groups	5	1001–5000 memes (39%)	Fig. 3 (Top)
Languages	6	English (73%)	Fig. 3 (Middle)
Data sources	17	Facebook/Reddit (19%)	Fig. 3 (Bottom)

Table 7

Overview of aspects covered in the identified datasets. Each row represents a dataset, each column an aspect, check marks indicate the presence of the corresponding aspect in each dataset.

	abusiveness	aggressiveness	attack type	bullying	disinformation	emotion/sentiment	explanation	fake/misattribution	harmfulness	hate speech	humour	intention	irony/sarcasm	metaphor	misogyny	modality-class	motivation	offensiveness	opinion manipulation	political attributes	profanity/vulgarity	propagandistic technique	target	troll	unethical
AOMD Gab [16]																		✓							
AOMD Reddit [16]																		✓							
BanglaAbuseMeme [14]	✓					✓							✓								✓		✓		
CrisisHateMM [79]																							✓		
Derogatory FB [80]						✓				✓											✓		✓		
DisinfoMeme [81]					✓																				
ELEMENT [82]																									✓
Emoffmeme [83]						✓												✓							
Ext-Harm-P [84]									✓																
Facebook Hateful [53]										✓															
FAME dataset [85]								✓																	
Fine grained HM [86]			✓							✓														✓	
GOAT-Bench [87]										✓															
Harm-C (HarMeme) [49]									✓	✓			✓		✓			✓						✓	
Harm-P [88]									✓	✓														✓	
Hate Speech in Pixels [89]										✓															
HatReD [90]							✓		✓																
HVVMemes [91]																							✓		
Indian Political [92]										✓															
Innapolis Hateful [93]										✓															
KAU-Memes [94]																									
Meme-Merge [95]																									
Memotion 1 [17]						✓												✓							
Memotion 2 [96]						✓					✓		✓				✓	✓							
MET-Meme [97]						✓					✓		✓				✓	✓							
MisogynisticMEME [98]		✓											✓		✓										
MultiBully [99]				✓		✓			✓				✓												
MultiBully-Ex [100]				✓		✓	✓		✓				✓												
MAMI [101]			✓												✓										
MultiOFF [102]																		✓							
Pol_Off_Meme [103]						✓												✓		✓					
SemEval-2021 6 [19]																						✓			
TamilMemes [104]																									✓
TrollsWithOpinion [23]																		✓	✓					✓	✓

by [14,108]. Additionally, terms like *derogatory* lacked clear definitions and were vaguely described using other toxicity descriptors. Fig. 4 shows the distribution of research focus across meme toxicity categories. Nearly half of the studies center on hateful memes, reflecting the influence of Kiela et al.'s (2020) *Hateful Memes challenge* [53] and its dataset. In contrast, categories like trolling, derogatory content, and disinformation are underexplored. Notably, no papers addressed toxicity types flagged by Sharma et al. (2022) [11], such as self-harm or exploitation, revealing key research gaps.

5.2. A taxonomy of meme toxicities

In examining toxicity term definitions, we identified explicit and implicit taxonomical relationships among certain terms. For example,

offensive, hateful, troll, and cyberbullying memes are typically defined as inherently harmful. Fake memes are seen as a subset of disinformation memes, while misogynous memes are considered a subtype of hateful memes targeting protected categories, such as sex. Recognizing the importance of clarifying these relationships, we aimed to establish a coherent taxonomy. We referred to the existing meme toxicity taxonomy by Sharma et al. (2022) [11], which included overlooked meme toxicities like self-harm and exploitation (e.g., adult sexual services, child sexual abuse, and scams). However, significant discrepancies emerged. Their taxonomy overlooked certain toxicities like disinformation, fake, and derogatory memes and lacked the concept of unethical memes, a superclass encompassing harmful memes and more. Additionally, the rationale for placing certain terms under specific macrocategories was unclear. For example, 'doxing' – defined

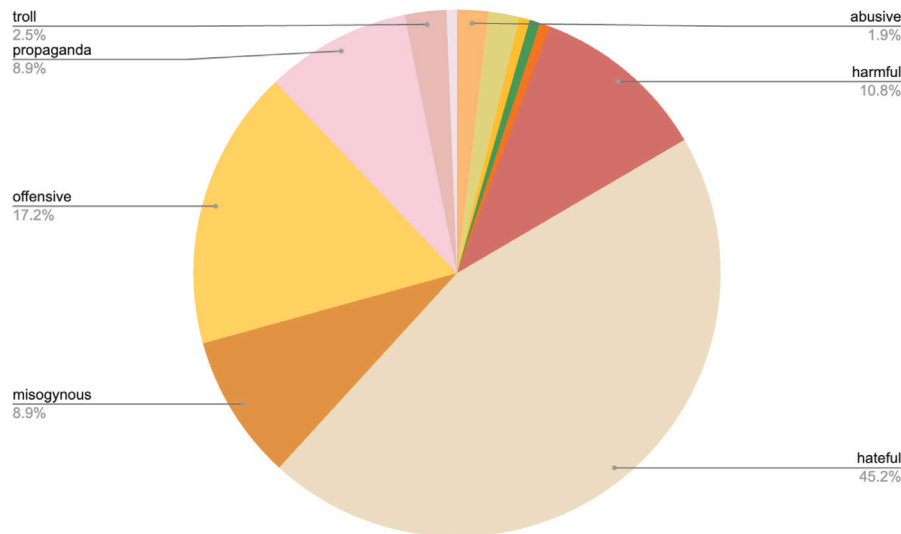


Fig. 4. Proportion of papers addressing different meme toxicity categories, based on authors' explicit labels.

as publicly publishing someone's private information as punishment or revenge – was categorized under hate rather than cyberbullying, even though it targets individuals rather than protected communities. Most critically, the taxonomy conflated two dimensions of meme toxicities: the type of toxicity (e.g., hateful, disinformation) and the techniques or attacks used (e.g., identity attack, identity misrepresentation). We believe separating these dimensions is vital, as certain techniques can be used across different toxicities (see Section 6).

Building on the taxonomy by Sharma et al. (2022) [11], we refined it to address identified discrepancies and nuances, enhancing its adaptability for future meme toxicity studies. Our revised taxonomy (Fig. 6) is grounded in literature-based definitions, organizing each toxicity type as a subcategory within a broader structure. Toxicities such as disinformation and propaganda stem from opinion manipulation, while categories like hate, harassment, and abuse – commonly studied computationally – fall under harmful/abusive content, defined by intent to demean, offend, or exploit. Our review reveals complex and overlapping relationships among these categories, which a strict taxonomy cannot fully capture. To address this, we developed a color-coded Venn diagram (Fig. 5) that visualizes their fuzzy boundaries and interconnectedness, offering a more realistic view of how toxicity types interact. While taxonomies are essential for computational analysis, the diagram reflects the real-world complexity of these relationships.

6. Dimensions of meme toxicities: Target, intent, tactic

During the refinement of the taxonomy, we found that meme toxicity definitions implicitly focused on three main aspects: the target of the toxicity, the intent behind the meme, and the rhetorical strategies used to convey the toxic message. Although terminology varied and the relationships between these aspects were not always clear, we identified them as three key dimensions of meme toxicity: intent, target, and tactic(s) (see Fig. 7). These dimensions offer a structured framework for understanding and addressing meme toxicity.

6.1. Target

The most studied dimension of toxic memes is the *intended target*. In memes, targeting refers to the harmful referencing of social entities [49,84]. This harm can take various forms, such as mental abuse, psycho-physiological injury, damage to reputation, emotional distress, or harm based on the background of the entity (e.g., bias, social or educational background) by the meme creator. However, it is important to note that referencing social entities in a meme does

not always lead to harm. Non-harmful references may include benign mentions, humor, limericks, harmless puns, or other content that does not cause distress [84]. Research has primarily focused on detecting the targets of harmful memes [49,79,84,88], categorizing them into the following types:

Individual. Toxic memes often target specific individuals, typically well-known figures such as politicians, actors, artists, scientists, or environmentalists, as noted by Sharma et al. [84]. Examples include Donald Trump, Joe Biden, Vladimir Putin, Hillary Clinton, Barack Obama, Chuck Norris, Greta Thunberg, and Michelle Obama. Some studies identify subtypes of individuals targeted by specific meme categories: **famous individuals** are often targeted in misattribution memes, with 20 notable figures frequently appearing in fake quote memes [85]. **Meme creators/posters** are implicitly targeted in self-harmful memes, while **meme viewers/receivers** are often the target of scam memes. **Children** may be targeted in memes related to child sexual abuse, and **adults** in memes involving adult sexual services [11].

Organization. Groups with specific purposes, such as businesses, government departments, companies, institutions, or associations are targeted, e.g., Facebook, the WTO, and the Democratic Party.

Community. Targets in memes often correspond to social units defined by shared personal, professional, cultural, or political attributes, such as religion, nationality, or gender identity. These communities can be geographic or virtual, formed through online platforms. A key focus within this context is **Protected Category (PC) Communities**, as highlighted in the Fine-grained version of Hateful Memes [86]. This dataset introduced detailed labels for protected categories like **race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease**, though it primarily covers race, disability, religion, nationality, and sex [122]. A similar approach was used in the BanglaAbuseMeme dataset [14], where target communities were categorized by gender, religion, nationality, political, and social sub-groups.

Society. The target is the entire societal fabric, e.g., when memes promote conspiracies, harming the general public.

6.2. Intent

Multimodal intent recognition is key to understanding human language in real-world scenes, but current methods are limited by benchmark datasets that primarily include text [233]. In memes, intent refers

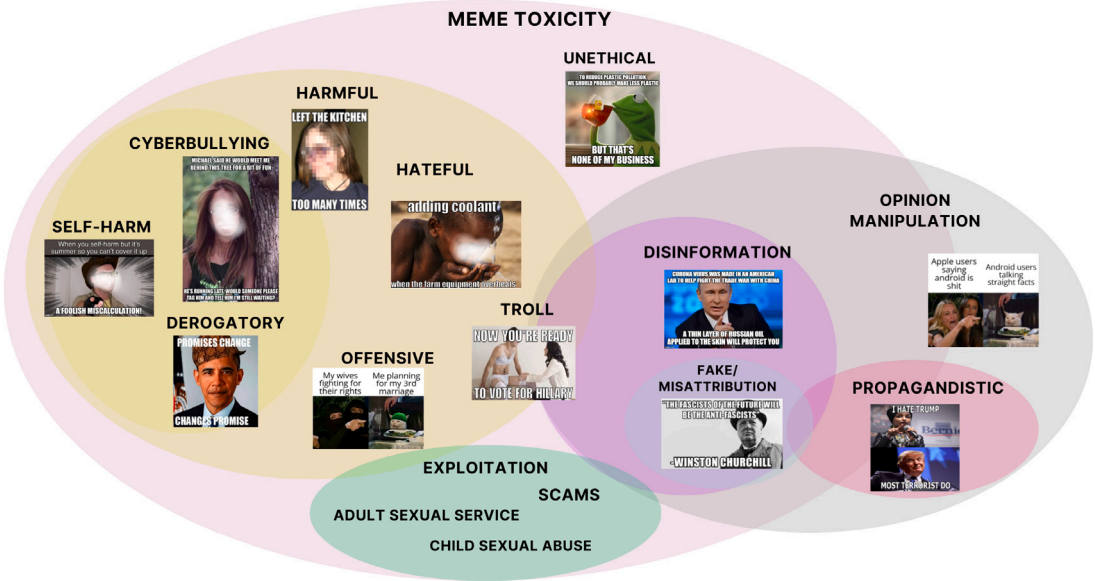


Fig. 5. Caution: Contains hateful, harmful, or otherwise toxic content, viewer discretion is advised. Venn diagram of overlapping meme toxicity categories. Exploitation is excluded due to missing dataset examples. Colors follow our taxonomy.

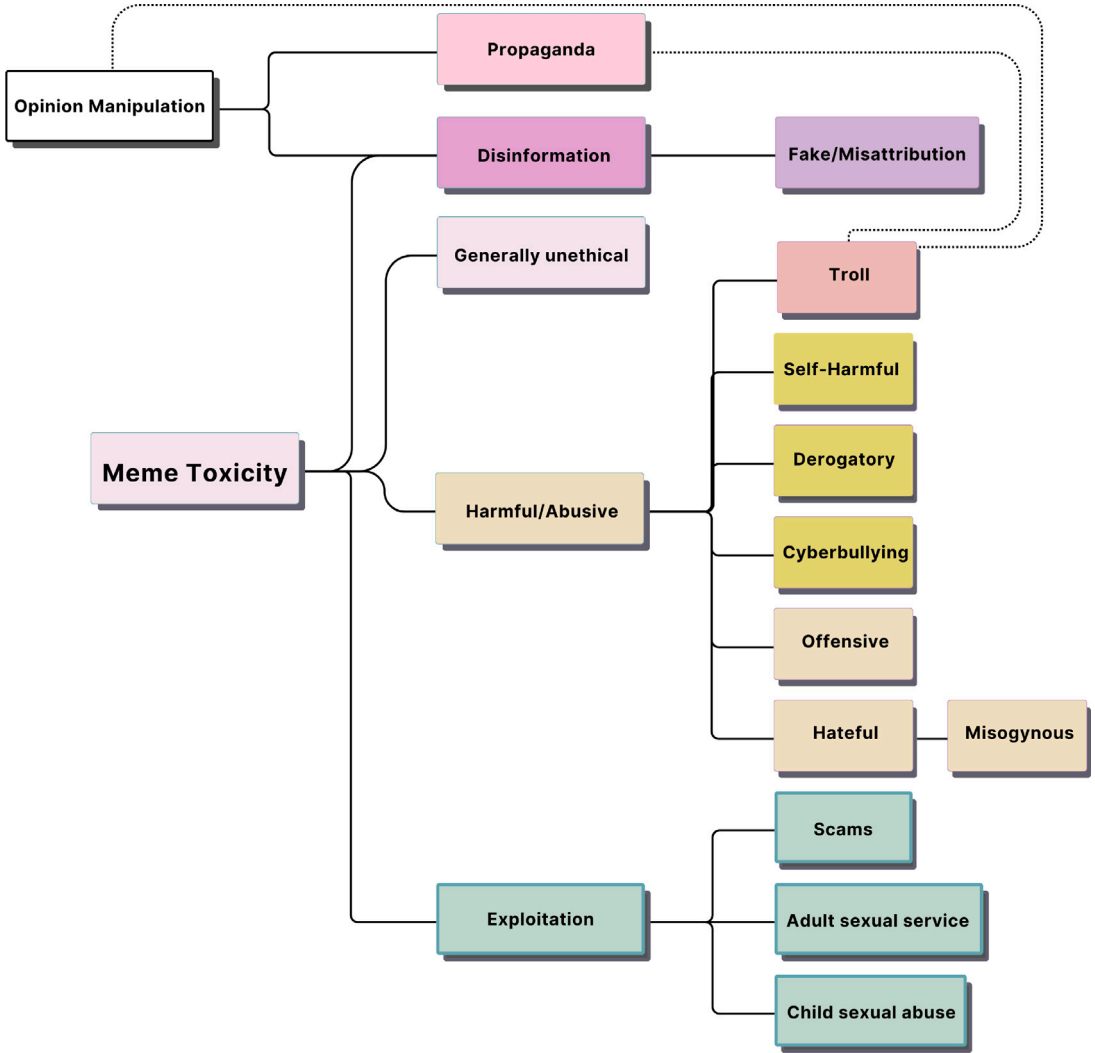


Fig. 6. The taxonomy for meme toxicities that we propose is inspired by the taxonomy presented in [11], while addressing discrepancies, enhancing taxonomical clarity, and including the most recent types of toxicities being computationally studied.

Table 8

Meme toxicity-related terms as utilized in the surveyed literature focusing on computationally addressing toxic memes. It outlines each term alongside its corresponding definition provided by the papers, with the last column indicating the surveyed papers and preprints primarily focused on that specific term..

Toxicity type	Definition	Focus of
Abusive	Used interchangeably with “harmful” by [14,108], who describe memes used by bad actors to threaten and abuse individuals or specific target communities. These memes contain words that target individuals or different protected communities, implicitly containing hateful, harmful, and antisemitic content. [107] uses it interchangeably with “offensiveness.”	[14,107,108]
Cyberbullying	Defined as any communication that disparages an individual based on characteristics such as color, gender, race, sexual orientation, ethnicity, nationality, or other features [109]. This definition, used by [100], builds upon the work by [15], who consider cyberbullying as an antisocial activity where victims are targeted with malicious behavior, including posting cruel comments and messages without fear of being identified. It appears that both [15] and [108] consider cyberbullying as a subset of harmful behavior.	[15,100,110]
Derogatory	Term used by [80] with no explicit definition. Vaguely defined as posts that convey a derogatory notion about a recognized individual (or person) of the country; malicious content about a political, spiritual, and cultural entity; and/or posts that include hateful and negative sentiments of sentences.	[80]
Disinformation	[81] defines disinformation memes as those designed to actively spread inaccurate information. They differentiate instances that criticize misinformation from those actively spreading inaccurate information.	[18]
Fake/Misattribution	Defined by [85] as a type of disinformation meme, these contain messages, fabricated or otherwise, falsely attributed to specific individuals. Such memes could be deployed against political opponents during smear campaigns.	[85]
Harmful	[49] see harmful memes as those that “have the potential to cause harm to an individual, an organization, a community, or the society more generally. Here, harm includes mental abuse, defamation, psycho-physiological injury, proprietary damage, emotional disturbance, and compensated public image. [...] Harmful is a more general term than offensive and hateful: offensive and hateful memes are harmful, but not all harmful memes are offensive or hateful” (page 4). As defined in [84], harmful memes encompass various forms of harm expressed overtly or subtly toward target entities, socio-cultural or political ideologies, beliefs, principles, or doctrines associated with them. The harm may manifest as abuse, offense, disrespect, insult, demeaning, or disregard toward the target entity or its affiliations. It can also include more subtle attacks such as mockery or ridicule of a person or an idea.	[12,13,49,84,88,91,111–121]
Hateful	In line with [106], hateful memes are characterized as direct or indirect attacks on individuals based on protected characteristics such as ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, disability, or disease. An attack is defined as containing violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. Additionally, mocking hate crimes is classified as hate speech. Exceptions to this definition include attacks on individuals/famous people not based on protected characteristics and attacks on groups perpetrating hate, such as terrorist groups. Detection of hate speech in memes often requires nuanced understanding of societal norms and context. More recently, fine-grained hateful types [122] have been studied, focusing on Protected Categories (PC) such as race, disability, religion, nationality, and sex, along with different Attack Types (AT) including contempt, mocking, statements of inferiority, slurs, exclusion, dehumanizing content, and incitement to violence.	[8–10,42,86,89,90,92,93,106,122–182]
Misogynous	Misogyny is a form of hate against women, and misogynous memes manifest different expressions of hate directed toward women, encompassing a broad spectrum [183]. These misogynous manifestations may be categorized into four main types.	[46,105,184–195]
Offensive	In some of the earliest meme analyses, “offensive” was conceived as a type of “emotion” alongside humor, sarcasm, and motivation, simply defined as something that aims to torment or disturb people [17]. The work introducing one of the most widely used datasets, multiOFF [102], follows [196]’s definition of offensive content as intending to upset or embarrass people by being rude or insulting. Offensive or abusive content on social media can be explicit or implicit [61], and could be classified as explicitly offensive or abusive if it is unambiguously identified as such. For example, it might contain racial, homophobic, or other offending slurs. In the case of implicit offensive or abusive content, the actual meaning is often obscured by the use of ambiguous terms, sarcasm, lack of profanity, hateful terms, or other means.	[16,17,83,94,103,197–218]
Propaganda	Defined by [19] as a form of communication to influence the opinions or actions of people toward a specific goal, achieved through well-defined rhetorical and psychological devices.	[19–21,219–229]
Troll	Defined by [23] as a meme that is often provocative, distractive, digressive, or off-topic with the intent to demean or offend particular people, groups, or races, containing either (I) Offensive text and non-offensive images; (II) Offensive images with non-offensive text; or (III) Sarcastically offensive text with non-offensive images, or sarcastic image with offensive text.	[22,230–232]
Unethical	As defined by [82], unethical memes are those deemed “inconsistent with human values,” failing to comply with ethical norms. They specify ethical memes as multi-modal units consisting of an image and embedded text aiming to promote fairness, justice, harmony, security, avoidance of bias, discrimination, privacy, and prevention of information leakage. While harmful and hateful memes are considered unethical, not all unethical memes are necessarily harmful or hateful. Detection of unethical content in memes is particularly challenging due to its often deeply implicit nature. Ethical memes focus not only on interpersonal principles but also on societal, human, and environmental relationships.	[82]

Table 9

Summary of Section 5, Defining meme toxicities: A taxonomy.

Topic	Count	Most frequent	Reference
Papers surveyed	158		Table 8
Toxicity types in papers	12	hateful (45%)	Table 8, Fig. 4
Labels in literature but not in datasets	3	self-harm, scams, exploitation	Figs. 5 and 6

to the underlying purpose of creation or dissemination, often annotated as interactive, expressive, entertaining, offensive, or other [97]. Within these categories, only the offensive intent, defined as the intent for

discrimination, satire, and abuse directed toward others’ occupation, gender, appearance, or personality [97] seems directly linked to toxicity [97]. While it is generally assumed that toxic memes have unethical

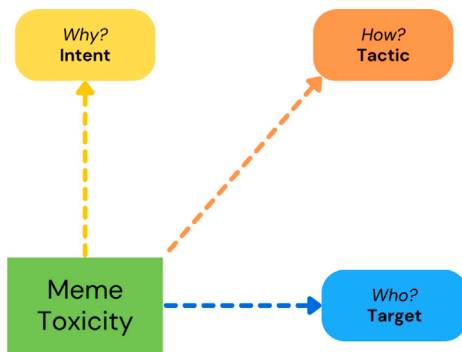


Fig. 7. The toxicity of memes is a complex phenomenon with multiple dimensions, represented by dotted edges. In this survey, we have identified three content-based dimensions of meme toxicity: target, intent, and tactic. These dimensions address the fundamental questions: *who* is the toxicity directed toward, *why* the toxic meme is shared (i.e., the underlying goal), and *how* the toxicity is manifested and conveyed. Other dimensions also exist and should be studied, potentially answering questions like *where* the meme was posted and *when* it was shared.

intent, such as conveying messages that violate societal values like fairness and privacy [82], no research has yet focused on specific toxic intents in memes. Comprehensive research comparing intents is still lacking but, nevertheless, based on the definitions of meme toxicities that we collected above, we identified at least three subtypes of intent in toxic memes:

Harm/abuse. The intent to harm, demean, abuse, or insult social entities [88].

Disinform. The intent to manipulate opinions by actively spreading false or inaccurate information [18,85].

Exploit. This intent is mentioned, but not explicitly defined, by Sharma et al. (2022) [11], who identified memes attempting to exploit people, including children for sexual abuse, adults for sexual services, or viewers for scams.

6.3. Tactic(s)

We also noticed a growing emphasis on automatically identifying the rhetorical techniques or tactics used by memes to convey toxicity. This focus is justified as memes employ a wide range of rhetorical strategies, persuasion techniques, and tactics to convey messages within their narratives [19]. This aspect, explored through attack types [86, 105,122], persuasion/propagandistic techniques [19], or entities' roles in harmful memes [115], represents a third dimension implicitly or explicitly mentioned in many papers, but these terms have not, to our knowledge, been collectively examined.

Attack types. delineate how memes, particularly those with hateful connotations, target individuals, groups, or ideas, and serve to categorize how memes express their message [86,122]. Attack types were defined within the context of the WOA5 shared task on fine-grained hateful memes detection [86]. Attack types have also been specialized to address specific forms of hateful messaging in the realm of misogynous memes [105]:

- **Dehumanization:** Explicitly or implicitly describing or presenting a group as subhuman [86]. It includes **misogynous objectification**, i.e., treating or regarding women as objects, devoid of agency/humanity [105].
- **Statements of Inferiority:** Claiming that a group is inferior, less worthy or less important than either society in general or another group [86].

- **Violence:** Explicitly or implicitly calling for harm to be inflicted on a group, including physical attacks [86]. It includes **misogynous violence**, i.e., the incitation or direct expression of acts of violence against women [105].
- **Mocking/Shaming:** Making jokes about, undermining, belittling, or disparaging a group [86]. It includes **misogynous shaming**, defined as the practice of criticizing women who violate expectations of behavior and appearance regarding issues related to gender typology (such as “slut shaming”) or related to physical appearance (such as “body shaming”). This category focuses on content that seeks to insult and offend women because of some characteristics of their body or personality [101].
- **Expression of Contempt:** Expressing intensely negative feelings or emotions about a group [86].
- **Slurs:** Using prejudicial terms to refer to, describe or characterize a group [86].
- **Calls for Exclusion:** Advocating, planning or justifying the exclusion or segregation of a group from all of society or certain parts [86].
- **Misogynous Stereotyping:** Propagating generalized beliefs about women across various contexts, including societal roles, personalities, and behaviors [105]. Stereotypes are fixed, conventional ideas or set of characteristics assigned to a woman, a meme can use an image of a woman according to her role in the society (role stereotyping), or according to her personality traits and domestic behaviors (gender stereotyping) [101].

Persuasion or propagandistic techniques. encompass various methods employed to influence the perception of a meme's audience. These methods include selective editing of images or text, framing narratives to elicit specific emotional responses, and employing symbols and motifs associated with particular ideologies or agendas. One significant study [19] identifies key propaganda techniques that serve as shortcuts in the argumentation process of memes specifically, leveraging audience emotions or logical fallacies to influence perception. Notably, the presence of these techniques does not inherently categorize a meme as propagandistic; rather, memes are only annotated based on the propaganda techniques they contain.

- **Loaded language:** Using emotionally charged words and phrases to influence the audience.
- **Name calling or labeling:** Assigning labels to entities that evoke strong reactions from the target audience.
- **Doubt:** Questioning the credibility of someone or something.
- **Exaggeration/Minimization:** Representing something in an excessive or diminished manner.
- **Appeal to fear/prejudices:** Instilling anxiety or panic to build support for an idea.
- **Slogans:** Brief and striking phrases acting as emotional appeals.
- **Whataboutism:** Discrediting an opponent's argument by charging them with hypocrisy.
- **Flag-waving:** Appealing to strong national or group sentiments to justify or promote an action or idea.
- **Straw man:** Substituting an opponent's proposition with a similar one, which is then refuted.
- **Causal oversimplification:** Assuming a single cause or reason when multiple causes exist for an issue.
- **Appeal to authority:** Stating that a claim is true simply because an authority on the issue said it was true.
- **Thought-terminating cliché:** Phrases discouraging critical thought and meaningful discussion on a given topic.
- **Black-and-white fallacy or dictatorship:** Presenting two alternative options as the only possibilities.

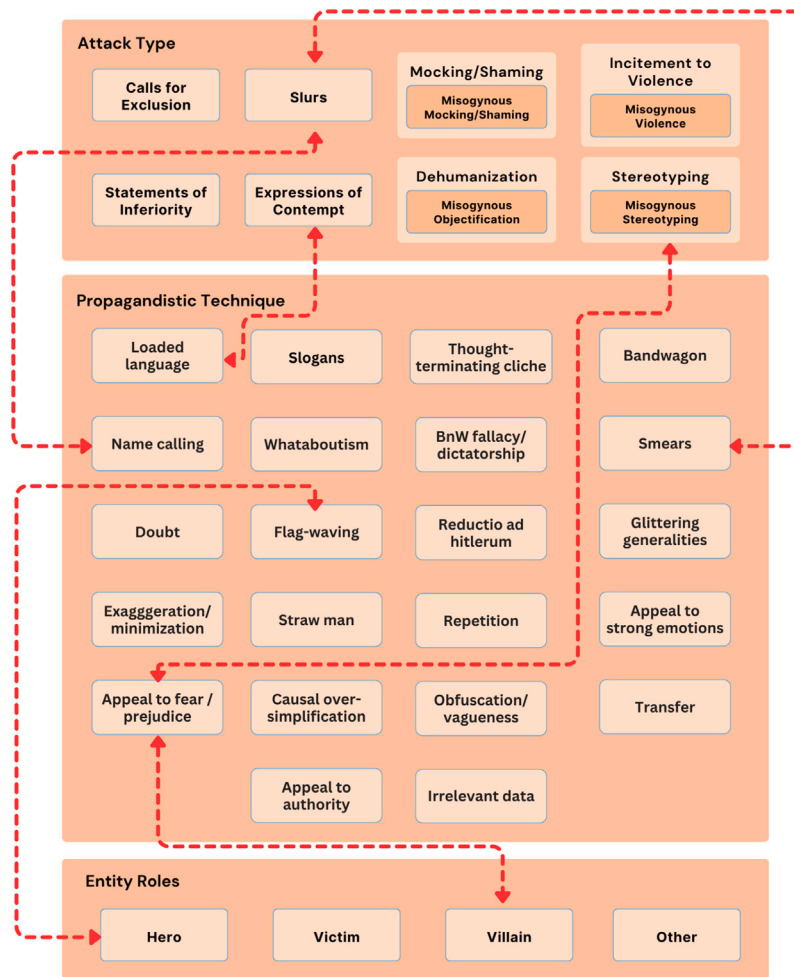


Fig. 8. Although the relationships between attack types, persuasion/propagandistic techniques, and entity roles have not been extensively studied, their definitions provide insights that suggest potential connections, equivalences, and other relationships (indicated by red dotted lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- **Reductio ad Hitlerum:** Disapproving an action or idea by suggesting it is popular with hated groups.
- **Repetition:** Repeating the same message to influence acceptance.
- **Obfuscation/Intentional vagueness/Confusion:** Using deliberately unclear words.
- **Presenting irrelevant data (Red Herring):** Introducing irrelevant material to divert attention.
- **Bandwagon:** Persuading the target audience to join a course of action because “everyone else is doing it”.
- **Smears:** Attempting to damage someone’s reputation through negative propaganda.
- **Glittering generalities:** Using words or symbols that produce a positive image when attached to a person/issue.
- **Appeal to (strong) emotions:** Using emotionally charged images to influence the audience.
- **Transfer:** Evoking an emotional response by projecting qualities of a person, entity, or value onto another.

Entity roles. A key feature in the rhetorical dimension of meme toxicity, particularly in harmful memes, involves identifying which entities are glorified, vilified, or victimized. The goal is to classify, from the perspective of the meme’s author, for a given meme and entity whether the entity is portrayed as a Hero, Villain, Victim, or another category [115]:

- **Hero:** Entities portrayed in a positive light, often glorified for their actions or inferred from context.

- **Villain:** Entities depicted negatively, associated with adverse traits like wickedness, cruelty, or hypocrisy.
- **Victim:** Entities shown suffering from the negative consequences of someone else’s actions or conveyed implicitly.
- **Other:** Entities that do not fit the categorization of hero, villain, or victim within the context of the meme.

Overall, these three aspects of memes – attack types, persuasion techniques, and entity roles – are all related to the rhetorical strategies used to convey meme toxicity, yet they have been studied separately thus far. However, more relationships may exist between specific attack types, persuasion techniques, and entity roles than previously recognized (see Fig. 8). For instance, the attack type *slurs* may be equivalent or strongly related to the persuasion techniques of *name calling* and *smears*, as they all involve the use of prejudicial terms to influence perception. Similarly, the attack type of *contempt* may be equivalent or related to the persuasion technique of *loaded language*, as they both entail the use of strongly charged emotional language. Additionally, there may be specific attack types or persuasion techniques employed to portray certain entities as heroes or villains. Furthermore, as noted in Section 4.2, some toxic meme datasets include labels for additional features that may be part of the rhetorical strategy for conveying toxicity in memes: emotion [17,83,96,97,99,100], sentiment [14,17,96,99], humor [51,96], irony/sarcsam [14,17,87,96,98,99], metaphor [97], profanity/vulgarity [14,80], and more. These connections remain largely unexplored, underscoring the need for a comprehensive examination of these aspects within the rhetorical dimension.

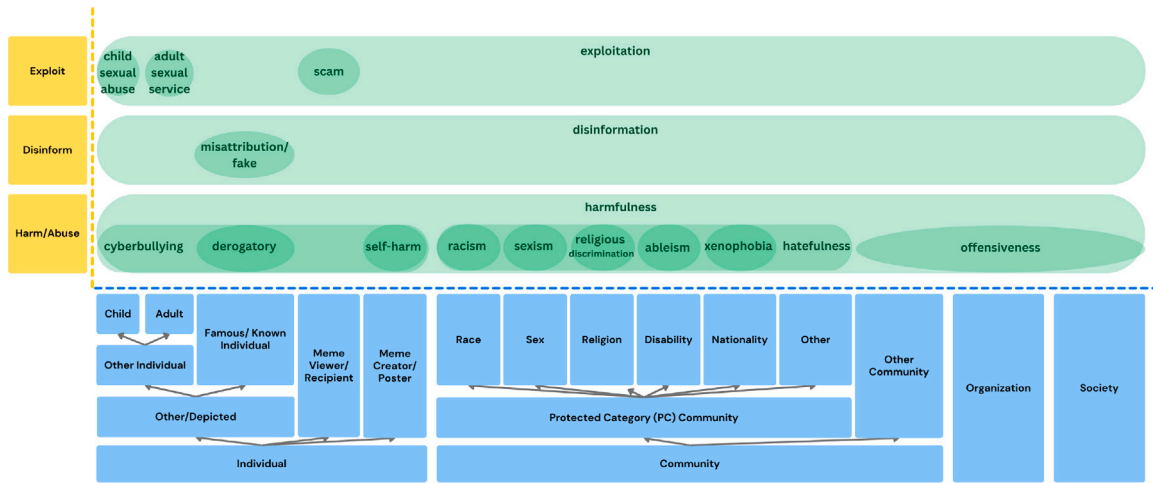


Fig. 9. Intersections between intent and target dimensions hint at specific types of meme toxicity. The y-axis, highlighted in yellow, categorically presents the three identified intents. Meanwhile, the x-axis, depicted in blue, categorizes target types with varying levels of specificity (from less specific at the bottom to more specific at the top). The green plane illustrates the toxicity landscape observed through the lens of intent and target dimensions, with greener areas indicating more specific toxicity types determined by particular intersections of intent and target. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 10
Summary of Section 6, Dimensions of meme toxicities: Target, intent, tactic.

Dimension	Types	Reference
Target	Individual, Organization, Community, Society	Section 6.1
Intent	Harm/Abuse, Disinform, Exploit	Section 6.2
Tactic	8 attack types, 22 propagandistic techniques, 4 entity types	Section 6.3, Fig. 8
Combinations	14 combinations of Target and Intent	Fig. 9

6.4. Dimensions' intersections

Intent \times target. We propose that analyzing multiple dimensions of toxicity simultaneously offers untapped potential, especially for tasks like automatic detection and explanation of toxicity. For example, identifying both the intent and target dimensions concurrently can provide insights into the toxicity types in memes. Fig. 9 illustrates how different intersections of targets and intents may lead to specific types of toxicity. For instance, a harmful intent targeted at people with disabilities can result in ableism. While the toxicities discussed in the literature are included here, this representation is not exhaustive—there could be other combinations of intents and targets not yet categorized.

Tactics and dimensions. It is important to note that the tactic used to convey a toxic message is likely independent of both the target and the intent. The same target can be attacked with the same intent through various tactics. Some tactics may not be applicable to certain combinations of target and intent, while others may be more suitable. However, in principle, it is not possible to establish a direct relationship between tactic, intent, and target since they are separate dimensions of toxic memes. Future research should explore whether these dimensions correlate in any meaningful way (see Table 10).

7. Research gaps and recent trends

7.1. Trend 1: Multimodal and cross-modal analysis

Until recently, AI models for detecting toxic or hateful speech primarily relied on unimodal classifiers and text-based datasets (e.g., [61, 66]). However, social science research has long advocated for a multimodal critical discourse approach, emphasizing the role of multiple sign systems – language, visuals, and context – in meaning construction [42,234,235]. Memes, which derive their meaning from the interplay between text and image, pose unique challenges for machine interpretation due to the complexity of multimodal understanding [44, 50,236].

The Hateful Memes Challenge at NeurIPS 2020 [106] exposed significant limitations in AI models that lacked holistic, multimodal reasoning. This prompted increased research into multimodal deep learning, where specialized neural networks analyze individual modalities before integrating their outputs through fusion techniques [39]. While these advancements have improved toxic meme detection, current models remain imperfect, as many toxic memes continue to evade automated filters due to their nuanced, multimodal nature [41]. A key work by Polli et al. (2023) [42] highlights the persistent challenges in automated toxic meme detection, demonstrating that the interplay between textual and visual elements often exceeds the capabilities of current computational models. Drawing from sociosemiotics and critical multimodal studies, they argue that meaning in hateful memes cannot be reduced to unimodal analysis or basic multimodal fusion. Their analysis illustrates how seemingly innocuous components can combine to produce toxic outcomes. For example, in Fig. 10, while both memes feature a non-hateful image, the second becomes toxic due to text that implicitly objectifies and dehumanizes the child, reinforcing racist stereotypes. This case underscores the need for cross-modal reasoning: since the text lacks explicit racial references, language-based AI classifiers fail to recognize its hateful intent. These findings expose the limitations of models relying solely on unimodal features. As a solution, Polli et al. advocate for computational approaches informed by semiotics and multimodal analysis, emphasizing the multiplicative nature of meaning-making [237].

A key trend emerging from this survey is thus the growing focus on sophisticated multimodal approaches to capture the complex interplay between different modalities in memes. Researchers are refining methods to assess the contribution of each modality, integrating unimodal and multimodal learning strategies. For instance, [147] employs a multi-task learning approach for hateful meme detection, combining a primary multimodal task with two unimodal auxiliary tasks. Similarly, [159] investigates the distinct roles of textual and visual components in hate detection, highlighting nuanced cross-modal interactions. Meanwhile, [13] proposes a representation framework



Fig. 10. **Caution: Contains hateful, harmful, or otherwise toxic content, viewer discretion is advised.** Illustration of how altering text modifies an image's interpretation. The left meme addresses water crisis awareness, while the right meme reframes the child as 'farm equipment,' perpetuating racist dehumanization. The child's face is blurred for privacy protection.

that facilitates inter-modal interaction and dynamically balances inter- and intra-modal relationships, systematically disentangling memes into modality-invariant and modality-specific spaces.

Attention-based architectures are becoming central to cross-modal meme analysis. For example, [209] utilizes an inter-modal attention framework to detect offensive memes by fusing visual and textual data, while [146] employs a cross-attention network to explore interactions between image objects and text features extracted via OCR. Similarly, [228] introduces a Multimodal Visual-Textual Object Graph Attention Network (MVITO-GAT) to model semantic and positional relationships between visual and textual elements, outperforming state-of-the-art baselines. Many recent studies leverage CLIP embeddings, which provide meaningful joint representations of text and images [238]. For instance, [100] integrates CLIP embeddings with modality-specific gating mechanisms to regulate interactions between text and image features. The HateCLIPper model [10] explicitly models cross-modal interactions by using CLIP encoders and a feature interaction matrix to capture meaningful semantic relationships. Additionally, [151] systematically examines semantic regularities in CLIP-generated embeddings, enabling the study of how hateful memes evolve by recombining visual and textual elements.

Building on and beyond fusion techniques, we observe a growing research focus on capturing nuanced cross-modal relationships and modeling discourse-intensive linguistic phenomena. In particular, perhaps in response to the challenges outlined by Polli et al. [42], recent studies have increasingly addressed complex analogies and figurative language in memes, leading to the emergence of several new approaches. For example, [121] introduces MemeFier, a dual-stage modality fusion approach leveraging Transformer encoders to learn token-level inter-modality correlations. [156] develops the Topology-Aware Optimal Transport (TOT) framework, which formulates optimal transportation plans and incorporates topology information to enhance representation alignment. Similarly, [16] presents the Analogy-aware Offensive Meme Detection (AOMD) framework, designed to detect offensive analogy memes by aligning implicit cross-modal analogies. Addressing figurative and politically charged language, [225] introduces FigMemes, a dataset for figurative language classification in politically opinionated memes, providing benchmark results for unimodal and multimodal models. Additionally, [239] explores brain-inspired perceptual integration to infer the subtle metaphors embedded in memes, exemplifying a broader shift toward discourse-intensive cross-modal reasoning.

7.2. Trend 2: Background knowledge integration

Memes are inherently intertextual, referencing elements from popular culture, symbols, events, or artifacts that hold significance within specific communities or affinity spaces [50]. This intertextuality makes memes highly contextual, requiring background knowledge in areas

such as politics, current events, and cultural references for accurate interpretation. Such knowledge, often termed "meme literacy" or "prior knowledge", is essential for understanding the nuanced meanings embedded in memes [131]. In the case of *toxic* memes, contextualization becomes even more crucial: an image that appears harmless can acquire a harmful meaning simply by shifting context, altering its interpretation and impact.⁸ As a result, integrating external knowledge into meme classification is emerging as a promising strategy to improve harmful meme detection and enhance real-world applicability [12].

This survey highlights the integration of background knowledge as a key trend in meme classification. A notable approach combines named entity recognition (NER) with knowledge base (KB) linking. For example, [216] incorporate Probase [240], a large-scale knowledge base with isA semantic relations, to enhance semantic representation in their MeBERT model. This enables the retrieval of relevant concepts from meme text, linking them to corresponding image regions via a concept-image attention module, improving meme classification. However, challenges remain for text-only memes and long-tail entities with limited knowledge base coverage.

Further advances are seen in MemeGraphs [144], which enhances meme classification by applying NER and KB augmentation to both meme text and scene graphs. By retrieving background knowledge from Wikidata and augmenting the meme text, MemeGraphs improves performance over models relying solely on learned representations.

Other works, like KnowMeme [213], utilize ConceptNet to capture implicit meanings and cross-modal relations within memes, achieving significant improvements in harmful meme detection. Similarly, KERMIT [12] enriches meme graphs by integrating internal meme entities with external knowledge from ConceptNet. KERMIT employs dynamic learning and memory-augmented neural networks to focus on the most informative segments of the enriched knowledge network for accurate classification.

A key recent development is the creation of meme-specific external knowledge bases. The Image Meme Knowledge Graph (IMKG) [241] is a structured repository of meme-related semantics, integrating data from sources like Wikidata and KnowYourMeme (KYM). IMKG enhances textual and visual data through entity extraction and semantic links, supporting applications in hate speech detection. The KYMKB [242], a knowledge base focused on meme templates, provides detailed information on 54,000 meme-related images, improving offensive meme detection by incorporating template meanings.

A key trend is the development of meme-specific external knowledge bases. The Image Meme Knowledge Graph (IMKG), introduced

⁸ For example, the Nesquik bunny, originally used in public information campaigns, was recontextualized into a disparaging meme mocking the African water crisis [42].

by [241], offers a structured repository of meme-related semantics across text, visuals, and metadata. It integrates diverse data sources, enriches information through entity extraction and semantic links, and adheres to Semantic Web principles. Using Wikidata [243], the authors extract background knowledge about meme seeds and entities, while KnowYourMeme (KYM) contributes lore, interpretations, and historical context. Textual enrichment is done with DBpedia Spotlight [244], which extracts entities from KYM content and ImgFlip captions, linking them to Wikidata entries. For visual enrichment, the Google Vision API detects objects and links them to Freebase [245]. While IMKG lacks image data and is not used for downstream tasks, the authors suggest incorporating it to improve the accuracy and explainability of neuro-symbolic methods, such as hate speech detection. Another development, KYMB [242], provides a knowledge base with 54,000 meme-related images, focusing on meme templates. It distinguishes between templatic and non-templatic memes, providing detailed data such as title, meaning, and origin. This database has proven valuable for tasks like offensive meme detection, showing promise in improving accuracy by leveraging template-specific information.

7.3. Trend 3: Interpretability and explanations

There has been a growing emphasis on interpretability in toxic meme detection, with many of the studies we surveyed recognizing it as a crucial aspect [12,91,126,144,197]. This trend is part of a broader movement toward developing complementary measures alongside algorithms for the automatic detection and removal of harmful content [246]. Specifically, there is an emerging focus on providing comprehensive rationales or explanations to help users and content moderators understand the nuances of toxic memes [90]. This shift reflects a growing recognition of the importance of enhancing users' understanding of harmful content, particularly in the context of memes, to improve media literacy [131]. While straightforward indicators like nudges, labels, and red flags have been suggested for identifying hateful memes, their efficacy remains largely untested [247].

To enhance the interpretability of meme toxicity, recent trends include the explicit identification of targeted entities, exploration of underlying themes and similarities, and the use of pre-trained language models for enhanced analysis and explanation. The explicit identification of entities targeted by toxicity is increasingly approached automatically, as exemplified in DisMultiHate [145], which disentangles target entities within multimodal memes to improve the classification and explainability of hateful content, and in DISARM [84], which employs named entity recognition and person identification for this purpose. Recent research also aims to provide insights into underlying themes and similarities, such as [46], which proposes modular and explainable architectures for meme understanding using example- and prototype-based reasoning, and the Hate-CLIPper approach [10], which identifies underlying themes and similarities across modalities. Furthermore, a recent trend in works addressing toxic memes involves enhancing interpretability by leveraging pre-trained language models, such as BERT and RoBERTa, to analyze linguistic patterns and contextual cues, generating insights into why certain content is classified as sexist [189], and using abductive reasoning to explore the interplay of multimodal information in memes [112].

Critically, recent datasets include ground-truth explanation annotations for memes, as noted in Section 4.2. For example, [131] propose quality-controlled crowdsourcing as an effective strategy for offering explanations and background knowledge for hateful memes through a Generate-Annotate-Revise workflow. The MultiBully-Ex dataset [100] provides multimodal explanations, combining visual cues like image segmentation with textual cues such as words relevant to or explaining the cyberbullying. Their experimental results demonstrate that training with multimodal explanations improves performance in generating textual justifications and accurately identifying visual evidence. Additionally, the HatReD dataset [90] annotates hateful memes with textual explanations, and models are trained to decode the meaning of multimodal hateful memes and provide explicit explanations for their classifications.

7.4. Trend 4: Low-resource languages

As social media continues to expand globally, the spread of harmful content, including toxic memes, transcends linguistic and cultural boundaries. While considerable attention has been given to detecting such content in English, there is a growing recognition of the urgency to address this issue in low-resource languages and contexts. Researchers have proposed integrating state-of-the-art deep learning models, such as BERT and Electra, for multilingual text analysis, alongside face recognition and optical character recognition models for understanding meme images [80]. Additionally, Bengali BERT models have been deployed for automated Bengali abusive text classification, aiming to streamline the hate speech filtering process in resource-constrained languages, achieving notable accuracy and performance [107]. In low-resource settings, multi-modal prompt tuning has emerged as an effective approach for detecting propaganda techniques in memes. This method incorporates visual cues into language models through prompt-based multi-modal fine-tuning, demonstrating efficacy in resource-limited scenarios [226]. Furthermore, transfer learning techniques have been explored to extend abusive meme detection to multiple languages. By leveraging model transfer techniques, researchers aim to bridge the language gap and establish baseline models for detecting abusive memes [108]. Dataset creation is also a significant effort, particularly in Asian languages such as Hindi, Bengali, Tamil, and Chinese. Efforts to address the lack of benchmark datasets for specific languages have led to the creation of resources like the BanglaAbuseMeme dataset for Bengali abusive meme classification. These datasets facilitate the development and evaluation of models for detecting abusive content in low-resource languages [14].

7.5. Trend 5: Generative AI and Large Language Models (LLMs)

The rapid development of Large Language Models (LLMs) and generative artificial intelligence (AI) represents a pivotal trend in the detection and interpretation of toxic memes. These advanced models enable deeper understanding of implicit meanings and context within multimodal content, offering promising solutions to address the challenges posed by harmful memes.

As of March 2024, the integration of LLMs with multimodal analysis has emerged as a significant approach to detecting and explaining toxic memes. Notably, LLMs are being employed for abductive reasoning, as seen in [112], where models distill multimodal reasoning knowledge and undergo lightweight fine-tuning to improve the detection of toxic content. Another innovative approach, explored in [111], leverages multimodal debates between LLMs to enhance the explainability of harmful meme detection. This approach facilitates a deeper understanding of the implicit meanings and context within memes, yielding more interpretable results. The use of LLMs also extends to the correction of harmful content, particularly hate speech within multimodal memes. [172] demonstrates how a large visual language model can effectively detect and correct hate speech, contributing to the mitigation of online toxicity. Additionally, LLMs are being explored for emotional analysis in memes, as exemplified by [174], which examines how GPT models can identify and interpret the emotions conveyed in memes, revealing the potential of LLMs for emotion detection in multimodal content.

Further advancing this trend, [133] introduces PromptHate, a novel prompt-based model that utilizes pre-trained language models for the classification of hateful memes. By constructing tailored prompts and providing contextual examples, PromptHate exploits the implicit knowledge embedded in pre-trained models, achieving high classification accuracy and demonstrating the value of leveraging external knowledge for meme detection. Additionally, a unified Multimodal Generative framework (MGex), proposed by [15], reframes meme detection as a multimodal text-to-text generation task. This innovative

Table 11
Summary of Section 7, Research gaps and recent trends.

Topic	Research gaps/challenges	Recent approaches	Ref.
Multimodal and Cross-Modal Analysis	<ul style="list-style-type: none"> Detection evasion due to text-image nuances. Lack of modality-specific role assessments. Struggles with unseen multimodal combinations. 	<ul style="list-style-type: none"> CLIP-based embeddings, attention models. Adaptive fusion beyond early/late methods. Discourse-aware detection for figurative language. 	Section 7.1
Background Knowledge Integration	<ul style="list-style-type: none"> Need for dynamic context-aware meme interpretation. Limited coverage of emerging meme entities. Difficulty linking diverse knowledge to text/visuals. 	<ul style="list-style-type: none"> Named Entity Recognition and linking (Probase, ConceptNet, Wikidata). Memory-augmented networks like KERMIT. Meme-specific knowledge bases (e.g., IMKG, KYMKB). 	Section 7.2
Interpretability and Explanations	<ul style="list-style-type: none"> Lack of clear explanations for toxic meme classification. Need for context-specific rationale for harmful content. Lack of tools bridging detection and user understanding. 	<ul style="list-style-type: none"> Explanation models (e.g., DisMultiHate, DISARM). Example/prototype-based reasoning for meme toxicity (e.g., Hate-CLIPper). Ground truth datasets for meme explanations (e.g., MultiBully-Ex). 	Section 7.3
Low-Resource Languages	<ul style="list-style-type: none"> Scarcity of datasets for low-resource languages. Adapting models trained on high-resource languages. Ensuring cross-linguistic and cultural detection. 	<ul style="list-style-type: none"> Multilingual models (e.g., BERT, Electra). Transfer learning to bridge language gaps. Benchmark datasets for languages like Hindi, Bengali. 	Section 7.4
Generative AI and LLMs	<ul style="list-style-type: none"> Early stage integration of LLMs for multimodal meme interpretation. Challenges with safety guardrails and misuse concerns. 	<ul style="list-style-type: none"> LLMs for abductive reasoning, emotional analysis, debate. Prompt-based models for text-to-text meme generation. GPT models for emotion analysis in memes. 	Section 7.5

approach demonstrates competitive performance against baselines and state-of-the-art models in detecting cyberbullying in memes.

While the applications of LLMs in toxic meme detection are promising, the growing implementation of safety guardrails within LLMs presents a key challenge. These guardrails are designed to prevent the generation and analysis of unsafe content, which may limit the models' ability to effectively process and interpret harmful datasets in the future [248].

In parallel, the rise of text-to-image models presents new challenges and opportunities in the generation of unsafe images and hateful memes. Studies such as [152] focus on understanding the potential of advanced text-to-image models to generate unsafe images, including hateful meme variants, and highlight the associated risks of model misuse. Moreover, [175] examines proactive approaches to generating unsafe images using benign prompts, underscoring the importance of implementing robust safety measures to mitigate the risks of generating harmful content.

Overall, while the use of generative AI and LLMs in meme detection shows significant potential, the future of this trend will be shaped by the evolving challenges of model safety, the need for further research into multimodal techniques, and the ongoing development of strategies to mitigate misuse.

8. Discussion and future directions

This survey offers a roadmap for researchers exploring computational approaches to toxic internet memes, with a focus on multimodal analysis and the complex taxonomies underlying harmful online content (see Section 2). Our review of existing work revealed the need for an up-to-date synthesis that addresses critical gaps in the literature and reflects emerging research directions. Following PRISMA 2020 guidelines (see Section 3), we analyzed 158 papers published between 2019 and 2024, offering a current and comprehensive perspective on the field.

8.1. Survey contributions and implications

We advance the study of toxic memes through five key contributions (see Table 12). First, Section 4 provides an in-depth overview of over 30 datasets used in toxic meme research, ranging from small collections to large-scale corpora. While most datasets are in English, there is a growing presence of non-English resources, particularly in Asian languages. We also examine the diverse data sources – social media platforms, image-hosting sites, search engines, and meme communities – and identify patterns in dataset reuse and adaptation. We also analyze the annotation practices used across datasets, offering a snapshot of labeling schemes. This synthesis enables researchers to compare approaches, evaluate labeling effectiveness, and support future dataset integration and methodological refinements in meme toxicity analysis.

Theoretical contributions are outlined in Sections 5 and 6. In Section 5, we unify 12 commonly used toxicity terms, revealing a research focus on hateful memes and a lack of attention to categories such as trolling, derogatory content, and disinformation. To support consistency and adaptability in future work, we propose a taxonomy that clarifies the relationships among toxicity categories, resolving inconsistencies found in earlier frameworks. Section 6 introduces a multidimensional framework grounded in three core elements of meme toxicity: the target, the intent behind creation or dissemination, and the rhetorical strategies employed. These dimensions are often examined separately in computational studies, but we argue for their integration to support more nuanced and systematic analyses. This framework provides a practical foundation for both manual and automated classification efforts.

Finally, Section 7 distills emerging trends and research directions, summarized in Table 11. These include tackling cross-modal entailment via deep learning and attention mechanisms, enhancing contextual and cultural awareness through entity recognition and external knowledge bases, and advancing interpretability and explainability. Additional trends point to the growing importance of supporting low-resource languages and harnessing generative AI and Large Language Models (LLMs) for more sophisticated analysis.

Table 12
Summary of SubSection 8.1 — Survey contributions and implications.

Contribution	Ref.
• Surveyed and cataloged 34 datasets containing labeled toxic memes	Section 4
• Defined, harmonized, and classified 12 toxicity terms into a comprehensive taxonomy	Section 5
• Developed a framework for meme toxicity dimensions, with an in-depth analysis of 3 key dimensions (intent, target, tactics)	Section 6
• Identified 5 critical research gaps in toxic meme analysis and current approaches	Section 7
• Proposed 6 actionable future research directions for advancing the field	Section 8

8.2. Future research directions

The trends identified in the previous section reflect a growing emphasis on capturing the multimodal, cultural, and intertextual complexity of memes—not only by moving beyond binary toxic/non-toxic labels, but also through culturally and linguistically informed analyses. There is a notable shift from simply detecting toxicity to explaining it, aligning with broader goals of transparency and accountability in AI systems. However, the field still lacks clear distinctions between computational transparency (explaining how a system works) and interpretability (explaining why a meme is labeled toxic). Encouragingly, we observe a shift toward the latter: providing reasoned, context-aware justifications for toxicity labels to foster trust and accountability in automated moderation. Based on these insights, we propose six future research directions, summarized in Table 13:

Semiotics-informed multimodal reasoning. Integrating semiotic insights into AI models can significantly enhance meme analysis by improving cross-modal understanding. Collaboration with semiotics experts can guide the integration of these insights, helping models better capture how meaning is constructed across text, image, and other modalities. Future work should explore feature engineering and fusion techniques – such as modality-specific gating and attention mechanisms – to better balance inter- and intra-modal relationships. Advancing cross-modal reasoning models will support more accurate toxicity detection and improve interpretability, making automated content moderation more transparent, helping users better understand how toxic meanings are constructed and disseminated.

Toxicity-specific knowledge injection. While resources like ConceptNet, WikiData, and Probase have improved meme interpretation, there remains a gap in databases focused on toxicity-related cultural icons. Integrating culturally specific knowledge databases that focus on toxicity is thus essential. Databases like the Global Extremist Symbols⁹ or other cultural references could help models recognize and interpret symbols or phrases that may be considered toxic in specific contexts. Research should explore how to effectively inject this type of knowledge into machine learning models to make them more culturally aware and sensitive to subtle forms of online toxicity. Furthermore, the integration of these knowledge sources should be dynamic, continuously updated to reflect emerging forms of toxicity. This approach would enable systems to better recognize evolving toxic content and provide more accurate and contextually appropriate labels, improving both the accuracy and cultural sensitivity of meme toxicity detection systems.

Multi-dimensional exploration of meme toxicity. Recognizing that, for AI to detect multimodal toxicity, “it must learn to understand content the way that people do: holistically” [86, p. 201], future research should focus on concurrently examining multiple dimensions of meme toxicity. We anticipate that the three dimensions of toxic meme classification (target, intent, tactic) may exhibit interdependencies, such as certain

tactics being more effective when targeting specific groups. Investigating these interrelations could provide valuable insights to refine and improve meme analysis in the future. We especially advocate for a multidimensional approach that incorporates evidential reasoning based on decision theory to guide judgements through the accumulation and scrutiny of evidence. Additionally, analyzing the context of posting (e.g., user, forum, platform) and propagation features [249] is essential for a deeper understanding of meme toxicity dynamics.

Legal, ethical, and collateral impacts. Exploring the legal and ethical implications of automatic moderation is crucial. Systems must differentiate between illegal and harmful yet legal content, raising complex ethical dilemmas. Further research is needed to address collateral damage in moderation algorithms, including the potential for false positives and unintended consequences, such as the censorship of marginalized voices or suppression of resistance narratives. Additional ethical challenges in meme analysis include managing subjectivity, preventing AI from generating toxic content, and navigating the complexities of toxic positivity.

Embracing linguistic diversity in toxic meme detection. Expanding toxic meme detection beyond English is essential for improving cultural relevance and detection accuracy. Annotating datasets in diverse languages and cultural contexts will ensure models are effective across different linguistic backgrounds. Future research should adapt existing models to these varied contexts and explore new approaches that address the unique characteristics of different languages. Embracing linguistic diversity will create more inclusive and robust frameworks for toxic meme detection.

Leveraging LLMs with RAG. Recent research has shown the potential of LLMs and generative AI for detecting and interpreting toxic memes, but more investigation is needed to assess their ability to distinguish between harmful and harmless content and identify cultural or modal biases. Future work should focus on evaluating LLMs’ cross-modal analysis capabilities, especially in complex multimodal data, and exploring methods like Retrieval-Augmented Generation (RAG) to integrate background knowledge. It is also crucial to examine how LLMs can adapt to evolving cultural references and ensure effective safeguards in toxic meme detection, while addressing ethical concerns.

9. Conclusion

‘Toxic’ memes encompass a broad spectrum of harmful or malevolent multimodal content circulating across online platforms, often with the intent to promote harm, spread hate, disseminate disinformation, or radicalize individuals. To address this growing challenge, there has been an exponential increase in computer science work aimed at automatically analyzing memes and their toxicities. This survey provides a comprehensive review of the automatic analysis of meme toxicities from a computational, content-based perspective, offering an in-depth examination of key developments in the field up to early 2024.

We cataloged 34 datasets containing toxic memes, detailing their labels, sources, and key attributes. We identified inconsistencies in

⁹ <https://globalextrémism.org/global-extremist-symbols-database/>

Table 13
Summary of SubSection 8.2 — Future research directions for meme toxicity analysis.

#	Future research direction	Focus areas
1	Semiotics-informed multimodality	Integrating semiotics for cross-modal reasoning, feature engineering, sophisticated fusion techniques
2	Toxicity-specific knowledge injection	Integrating cultural knowledge databases that include information about online toxicity
3	Multi-dimensional analyses of toxicity	Analyzing interdependencies between target, intent, tactic, context, and propagation
4	Legal, ethical, and collateral impacts	Addressing ethical dilemmas, collateral damage, subjectivity, and AI content generation
5	Embracing linguistic diversity	Expanding detection to more languages and cultural contexts; leveraging multilingual knowledge
6	Leveraging LLMs with RAG	Exploring LLMs with Retrieval-Augmented Generation for cross-modal analysis and cultural understanding

the terminology used to describe and reference meme toxicities and highlighted the need for frameworks and tools to detect fine-grained toxicity types beyond a simple toxic/non-toxic classification. To address this, we proposed a harmonized set of definitions, introduced a novel taxonomy, and presented a framework capturing the multi-dimensional nature of meme toxicity, including intent, target, and conveyance tactics. We also explored how these dimensions interact and encouraged further research into these relationships.

Importantly, we identify five key research gaps and trends in the field, proposing specific future research directions that will help guide further progress. By analyzing the prevalent challenges in the computational detection of toxic memes and identifying emerging trends, we aim to contribute to the development of more effective solutions for automatic meme interpretation. In summary, this survey offers valuable insights into the current state of the field, provides critical resources for future research, and emphasizes the need for interdisciplinary collaboration to foster a more nuanced and effective approach to meme toxicity detection and moderation.

CRediT authorship contribution statement

Delfina S. Martinez Pandiani: Conceptualization, Writing – original draft, Writing – review & editing, Resources, Methodology, Investigation, Data curation. **Erik Tjong Kim Sang:** Writing – review & editing, Validation, Investigation, Data curation. **Davide Ceolin:** Conceptualization, Writing – review & editing, Methodology, Supervision, Project administration, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

Statement: During the preparation of this work, the authors used generative AI-based tools (i.e., Grammarly and ChatGPT) solely during the final copyediting stage to enhance readability and language of

the manuscript. All content edited with the assistance of these tools was thoroughly reviewed and edited by the authors, who take full responsibility for the final content of the published article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Davide Ceolin reports financial support was provided by Netherlands eScience Center. Davide Ceolin reports a relationship with Netherlands eScience Center that includes: funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This publication has been supported by the Netherlands eScience Center project “The Eye of the Beholder” (project nr. 027.020.G15) and it is part of the AI, Media & Democracy Lab (Dutch Research Council project number: NWA.1332.20.009). For more information about the lab and its further activities, visit <https://www.aim4dem.nl/>.

This publication has been supported by the Netherlands eScience Center project “The Eye of the Beholder” (project nr. 027.020.G15) and it is part of the AI, Media & Democracy Lab (Dutch Research Council project number: NWA.1332.20.009). For more information about the lab and its further activities, visit <https://www.aim4dem.nl/>. However, now this content was repeated both in the Declaration of competing interest and in the Acknowledgements. I took out the one from the Declaration

Appendix

See Fig. A.11 and Table A.14.

Table A.14

Sources of meme acquisition, as documented in the papers introducing the datasets. Rows are datasets, while columns are sources categorized into search engines, image hosting platforms, social media platforms, or meme creation and sharing platforms. Abbreviations used: B: Bing, G: Google, D: DuckDuckGo, B: Baidu, Pin: Pinterest, Img: Image Hosting Platforms, Img: Imgur, FB: Facebook, IG: Instagram, Meme Platforms: Meme Creation and Sharing Platforms, RD: Reddit, TW: Twitter, Wei: Weibo, WA: WhatsApp, KYM: KnowYourMeme, MD: Memedroid, MG: MemeGenerator.

Dataset	Search engines				Img platforms		Social media platforms								Meme platforms			
	B	G	D	B	Pin	Img	FB	Gab	IG	RD	TW	Wei	WA	KYM	MD	MG	9gag	
AOMD Gab								✓										
AOMD Reddit										✓								
BanglaAbuseMeme	✓	✓					✓		✓									
CrisisHateMM							✓			✓	✓							
Derogatory Fb-Meme							✓											
DisinfoMeme							✓											
ELEMENT		✓								✓						✓		
Emoffmeme		✓																
Ext-Harm-P		✓					✓		✓	✓								

(continued on next page)

Table A.14 (continued).

Dataset	Search engines				Img platforms		Social media platforms							Meme platforms			
	B	G	D	B	Pin	Img	FB	Gab	IG	RD	TW	Wei	WA	KYM	MD	MG	9gag
Facebook HM							✓										
FAME dataset			✓														
Fine grained HM							✓			✓						✓	
GOAT-Bench		✓				✓	✓		✓	✓	✓			✓			✓
Harm-C		✓					✓		✓	✓							
Harm-P		✓					✓		✓	✓							
Hate Speech in Pixels		✓								✓							
HatReD							✓										
HVVMemes		✓					✓		✓	✓							
Indian Political Memes		✓															
Innapolis Hateful Memes			✓								✓				✓		
KAU-Memes							✓		✓	✓	✓						
Meme-Merge				✓							✓	✓					
Memotion 1		✓															
Memotion 2						✓	✓		✓	✓							
MET-Meme		✓		✓							✓	✓					
Misogynistic-MEME							✓		✓	✓	✓						
MultiBully										✓	✓						
MultiBully-Ex										✓	✓						
MAMI						✓				✓	✓			✓			✓
MultiOFF							✓		✓	✓	✓						
Pol_Off_Meme		✓															
SemEval-2021 Task 6							✓										
TamilMemes					✓		✓		✓				✓				
TrollsWithOpinion		✓															

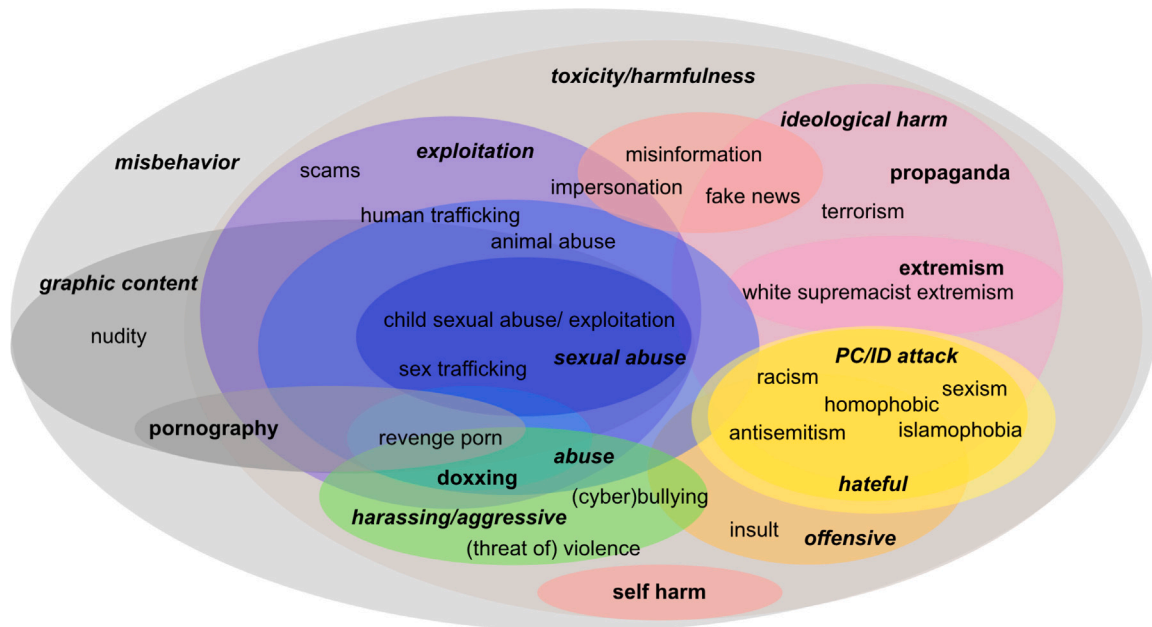


Fig. A.11. Toxicity-related terms derived from our investigation of harmfulness and toxicity in multimodal data, illustrating the complex and overlapping nature of multimodal toxicities.

Data availability

The data is available on GitHub.

References

- [1] R. Dawkins, *The Selfish Gene*, New ed., Oxford University Press, 1989.
- [2] C. Koutlis, M. Schinas, S. Papadopoulos, MemeTector: Enforcing deep focus for meme detection, *Int. J. Multimed. Inf. Retr.* 12 (1) (2023) 11.
- [3] S. Peeters, M. Tuters, T. Willaert, D. De Zeeuw, On the vernacular language games of an antagonistic online subculture, *Front. Big Data* 4 (2021) 718368.
- [4] C. Arkenbout, Political meme toolkit: leftist dutch meme makers share their trade secrets, in: *Critical Meme Reader II: Memetic Tacticality*, Institute of Network Cultures, 2022, pp. 20–31.
- [5] D. Dennet, Dangerous memes - a TED talk, 2007, Youtube, URL <https://www.youtube.com/watch?v=KzGjEkp772s>.
- [6] A. Wagener, Semiotic excess in memes: From postdigital creativity to social violence, *Internet Pragmat.* 6 (2) (2023) 239–258.
- [7] A. Arora, P. Nakov, M. Hardalov, S.M. Sarwar, V. Nayak, Y. Dinkov, D. Zlatkova, K. Dent, A. Bhatawdekar, G. Bouchard, I. Augenstein, Detecting harmful content on online platforms: What platforms need vs. where research efforts go, 2023, arXiv:2103.00153.
- [8] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, C. Fitzpatrick, P. Bull, G. Lipstein, T. Nelli, R. Zhu, N. Muennighoff, R. Veliglu, J. Rose, P. Lippe, N. Holla, S. Chandra, S. Rajamanickam, G. Antoniou, E. Shutova, H.

- Yannakoudakis, V. Sandulescu, U. Ozertem, P. Pantel, L. Specia, D. Parikh, The hateful memes challenge: Competition report, in: *Proceedings of Machine Learning Res.*, vol. 133, 2020.
- [9] H. Kirk, Y. Jun, P. Rauba, G. Wachtel, R. Li, X. Bai, N. Broestl, M. Doff-Sotta, A. Shtedritski, Y. Asano, Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset, in: *Proceedings of WOAHI 2021 - 5th Workshop on Online Abuse and Harms*, vol. 26, 2021.
- [10] G. Kumar, K. Nandakumar, Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features, in: *NLP4PI 2022 - 2nd Workshop on NLP for Positive Impact*, *Proceedings of the Workshop*, vol. 171, 2022.
- [11] S. Sharma, F. Alam, M.S. Akhtar, D. Dimitrov, G. Da San Martino, H. Firooz, A. Halevy, F. Silvestri, P. Nakov, T. Chakraborty, Detecting and understanding harmful memes: A survey, 1 (1), 2022, pp. 1–9, arXiv preprint arXiv:2205.04274.
- [12] B. Grasso, V. La Gatta, V. Moscato, G. Sperli, KERMIT: Knowledge-Empowered model in harmful meme detection, *Inf. Fusion* 106 (2024) <http://dx.doi.org/10.1016/j.inffus.2024.102269>.
- [13] C. Yang, F. Zhu, J. Han, S. Hu, Invariant meets specific: A scalable harmful memes detection framework, in: *MM 2023 - Proceedings of the 31st ACM International Conference on Multimedia*, vol. 4788, 2023, <http://dx.doi.org/10.1145/3581783.3611761>.
- [14] M. Das, A. Mukherjee, BanglaAbuseMeme: A dataset for bengali abusive meme classification, 2023, pp. 15498–15512, arXiv preprint arXiv:2310.11748.
- [15] R. Jain, K. Maity, P. Jha, S. Saha, Generative models vs discriminative models: Which performs better in detecting cyberbullying in memes? in: *Proceedings of the International Joint Conference on Neural Networks*, 2023, <http://dx.doi.org/10.1109/ijcnn54540.2023.10191363>.
- [16] L. Shang, Y. Zhang, Y. Zha, Y. Chen, C. Youn, D. Wang, AOMD: An analogy-aware approach to offensive meme detection on social media, *Inf. Process. Manage.* (2021) <http://dx.doi.org/10.1016/j.ipm.2021.102664>.
- [17] C. Sharma, D. Bhageria, W. Scott, S. Pykl, A. Das, T. Chakraborty, V. Pulabaigari, B. Gambak, SemEval-2020 task 8: Memotion analysis—the visuo-lingual metaphor!, in: *Proceedings of SemEval 2020 – 14th International Workshops on Semantic Evaluation*, 2020, pp. 759–773, arXiv preprint arXiv:2008.03781.
- [18] A. Williams, M. Dupuis, I don't always spread disinformation on the web, but when I do I like to use memes: An examination of memes in the spread of disinformation, in: *Proceedings of the 11th International Multi-Conferences on Complexity, Informatics and Cybernetics: IMCIC*, 2020, pp. 165–172.
- [19] D. Dimitrov, B.B. Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, G.D.S. Martino, Detecting propaganda techniques in memes, in: *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, *Proceedings of the Conference*, 2021, pp. 6603–6617, arXiv preprint arXiv:2109.08013.
- [20] M. Abdullah, D. Abujaber, A. Al-Qarqaz, R. Abbott, M. Hadzikadic, Combating propaganda texts using transfer learning, *IAES Int. J. Artif. Intell.* 12 (2) (2023) 956–965, <http://dx.doi.org/10.11591/ijai.v12.i2.pp956-965>.
- [21] D. Rodríguez, P. Nakov, V. Dankers, E. Shutova, Paper bullets: Modeling propaganda with the help of metaphor, in: *European Chapter of the Association for Computational Linguistics, Findings of EACL 2023*, 2023, pp. 472–489.
- [22] M.G. Shridara, D. Hlášek, M. Pleva, R. Haluška, Identification of trolling in memes using convolutional neural networks, in: *2023 33rd International Conference Radioelektronika, RADIOELEKTRONIKA, IEEE*, 2023, pp. 1–6.
- [23] S. Suryawanshi, B.R. Chakravarthi, M. Arcan, P. Buitelaar, TrollsWithOpinion: A taxonomy and dataset for predicting domain-specific opinion manipulation in troll memes, *Multimedia Tools Appl.* 82 (6) (2023) 9137–9171, <http://dx.doi.org/10.1007/s11042-022-13796-x>.
- [24] D. Bebić, M. Volarevic, Do not mess with a meme: the use of viral content in communicating politics, *Commun. Soc.* 31 (3) (2018) 43–56.
- [25] G. Mazambani, M.A. Carlson, S. Reysen, C.F. Hempelmann, Impact of status and meme content on the spread of memes in virtual communities, *Hum. Technol.: An Interdiscip. J. Humans ICT Environ.* 11 (2) (2015) 148–164.
- [26] Netanyahu, The dangers of memes — netanyataitague, 2019, <https://medium.com/@netanyataitague/the-dangers-of-memes-b1bb67e10083>. (Accessed 17 April 2024).
- [27] M. Vang, Is meme culture problematic - the current — thecurrentmsu.com, 2021, <https://thecurrentmsu.com/2021/08/07/is-meme-culture-problematic/>. (Accessed 17 April 2024).
- [28] K. Rojas, The toxicity of online meme culture: When is it too far? — studybreaks.com, 2022, <https://studybreaks.com/thoughts/meme-culture-2/>. (Accessed 17 April 2024).
- [29] Z.D. Roberts, How the 'free helicopter rides' meme went viral — progressive.org, 2023, <https://progressive.org/magazine/how-the-free-helicopter-rides-meme-went-viral-roberts-20230907/>. (Accessed 17 April 2024).
- [30] F.J.A. Serna, Los memes como símbolos del discurso de odio: La influencia del humor gráfico en la libertad de expresión y la política, *VISUAL REVIEW. Int. Vis. Cult. Review/ Rev. Int. de Cult. Vis.* 16 (2) (2024) 241–253.
- [31] L. Needham, How the toxic went mainstream — pursuit.unimelb.edu.au, 2019, <https://pursuit.unimelb.edu.au/articles/how-the-toxic-went-mainstream>. (Accessed 17 April 2024).
- [32] K.M. Duchscherer, J.F. Dovidio, When memes are mean: Appraisals of and objections to stereotypic memes, *Transl. Issues Psychol. Sci.* 2 (3) (2016) 335–345.
- [33] P.M. Bennet, Who Moderates the Social Media Giants? A Call to End Outsourcing, Technical Report, NYU STERN, Center for Business and Human Rights, 2020.
- [34] N. Nondo, Facing disturbing content daily, online moderators in africa want better protections and a fair wage, 2023, CBC Radio.
- [35] B. Perrigo, Inside facebook's african sweatshop, *Time* (2022).
- [36] N. Rowe, "It's destroyed me completely": Kenyan moderators decry toll of training of AI models, *Guardian* (2023).
- [37] N. Mbagathi, In africa, taking on viral hate, *Open Soc. Found.* (2023).
- [38] T.H. Afridi, A. Alam, M.N. Khan, J. Khan, Y.-K. Lee, A multimodal memes classification: A survey and open research issues, in: *Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications*, Springer, 2021, pp. 1451–1466.
- [39] P.C.d.Q. Hermida, E.M.D. Santos, Detecting hate speech in memes: a review, *Artif. Intell. Rev.* 56 (2023) 1–19, <http://dx.doi.org/10.1007/s10462-023-10459-7>.
- [40] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, G. Suarez-Tangil, On the origins of memes by means of fringe web communities, in: *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 188–202.
- [41] T. Chakraborty, S. Masud, Nipping in the bud: detection, diffusion and mitigation of hate speech on social media, *ACM SIGWEB Newsl.* 2022 (Winter) (2022) 1–9.
- [42] C. Polli, M.G. Sindoni, Multimodal computation or interpretation? Automatic vs. Critical understanding of text-image relations in racist memes, 2023, Ssrn.
- [43] O. Solon, Richard dawkins on the internet's hijacking of the word 'meme', 20, 2013, *Wired UK*.
- [44] L. Shifman, Memes in digital culture, in: *MIT Press Essential Knowledge Series*, MIT Press, 2013.
- [45] M. Dynel, The life of COVID-19 mask memes: a diachronic study of the pandemic memescape, *Comunicar* 30 (72) (2022) 73–85.
- [46] A. Thakur, F. Ilievski, H.-A. Sandlin, Z. Sourati, L. Luceri, R. Tommasini, A. Mermoud, Explainable classification of internet memes, in: *CEUR Workshop Proceedings*, vol. 3432, 2023, pp. 395–409.
- [47] L. Xie, A. Natsev, J.R. Kender, M. Hill, J.R. Smith, Visual memes in social media: tracking real-world news in youtube videos, in: *Proceedings of the 19th ACM International Conference on Multimedia*, 2011, pp. 53–62.
- [48] Y. Du, M.A. Masood, K. Joseph, Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter, in: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, Aaai, 2020, pp. 153–164.
- [49] S. Pramanick, D. Dimitrov, R. Mukherjee, S. Sharma, M.S. Akhtar, P. Nakov, T. Chakraborty, Detecting harmful memes and their targets, 2021, arXiv preprint arXiv:2110.00413.
- [50] M. Knobel, C. Lankshear, Online memes, affinities, and cultural production, *A New Literacies Sampl.* 29 (2007) 199–227.
- [51] C. Sharma, V. Pulabaigari, A. Das, Meme vs. Non-meme classification using visuo-linguistic association., in: *WEBIST*, 2020, pp. 353–360.
- [52] V. Sherratt, K. Pimblett, N. Dethlefs, Multi-channel convolutional neural network for precise meme classification, in: *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 2023, pp. 190–198.
- [53] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 2611–2624.
- [54] B. Kostadinovska-Stojchevska, E. Shalevska, Internet memes and their socio-linguistic features, *Eur. J. Lit. Lang. Linguist. Stud.* 2 (4) (2018).
- [55] N. Lapidot-Leffer, A. Barak, Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition, *Comput. Hum. Behav.* 28 (2) (2012) 434–443.
- [56] D. Gordeev, V. Potapov, Toxicity in texts and images on the internet, in: *International Conference on Speech and Computer*, Springer, 2020, pp. 156–165.
- [57] A. Sheth, V.L. Shalin, U. Kursuncu, Defining and detecting toxicity on social media: context and knowledge are key, *Neurocomputing* 490 (2022) 312–318.
- [58] N. Carlisle, Toxicity, memes and raids - power of zero — powerof0.org, 2022, <https://powerof0.org/toxicity-memes-and-raids/>. (Accessed 17 April 2024).
- [59] M. Mosleh, R. Cole, D.G. Rand, Misinformation and harmful language are interconnected, rather than distinct, challenges, *PNAS Nexus* (2024) pgae111.
- [60] S. Ghosh, S. Lepcha, S. Sakshi, R.R. Shah, S. Umesh, Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances, 2021, arXiv preprint arXiv:2110.07592.
- [61] Z. Waseem, T. Davidson, D. Warmesley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, 2017, arXiv preprint arXiv:1705.09899.

- [62] M. Wiegand, M. Siegel, J. Ruppenhofer, Overview of the GermEval 2018 shared task on the identification of offensive language, in: Proceedings of GermEval 2018, 14th Conference on Natural Language Processing, KONVENS 2018, Austrian Academy of Sciences, 2019, pp. 1–10, URL <https://nbn-resolving.org/urn:nbn:de:bsz:mh39-84935>.
- [63] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), 2019, arXiv preprint [arXiv:1903.08983](https://arxiv.org/abs/1903.08983).
- [64] P. Piot, P. Martín-Rodilla, J. Parapar, MetaHate: A dataset for unifying efforts on hate speech detection, 2024, arXiv preprint [arXiv:2401.06526](https://arxiv.org/abs/2401.06526).
- [65] C. Adams, J. Sorensen, J. Elliott, L. Dixon, M. McDonald, n, W. Cukierski, Toxic comment classification challenge, 2017, URL.
- [66] T. Davidson, D. Warmesley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, 2017, pp. 512–515.
- [67] F.E. Ayo, O. Folorunso, F.T. Ibharalu, I.A. Osinuga, Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions, *Comput. Sci. Rev.* 38 (2020) 100311.
- [68] F. Alkomah, X. Ma, A literature review of textual hate speech detection methods and datasets, *Information* 13 (6) (2022) 273.
- [69] A. Chhabra, D.K. Vishwakarma, A literature survey on multimodal and multilingual automatic hate speech identification, *Multimedia Syst.* 29 (3) (2023) 1203–1230.
- [70] P. Fortuna, J. Soler, L. Wanner, Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 6786–6794.
- [71] M. Yankoski, W. Scheirer, T. Weninger, Meme warfare: AI countermeasures to disinformation should focus on popular, not perfect, fakes, *Bull. At. Sci.* 77 (3) (2021) 119–123.
- [72] M. Banko, B. MacKeen, L. Ray, A unified taxonomy of harmful content, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, 2020, pp. 125–137.
- [73] P. Nakov, V. Nayak, K. Dent, A. Bhatawdekar, S.M. Sarwar, M. Hardalov, Y. Dinkov, D. Zlatkova, G. Bouchard, I. Augenstein, Detecting abusive language on online platforms: A critical analysis, 2021, arXiv preprint [arXiv:2103.00153](https://arxiv.org/abs/2103.00153).
- [74] A. Halevy, C. Canton-Ferrer, H. Ma, U. Ozertem, P. Pantel, M. Saeidi, F. Silvestri, V. Stoyanov, Preserving integrity in online social networks, *Commun. ACM* 65 (2) (2022) 92–98.
- [75] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G.D.S. Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, 2021, arXiv preprint [arXiv:2103.12541](https://arxiv.org/abs/2103.12541).
- [76] Anjum, R. Katarya, Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities, *Int. J. Inf. Secur.* 23 (1) (2024) 577–608.
- [77] B. Lewandowska-Tomaszczyk, A. Bączkowska, C. Liebeskind, G. Valunaite O., S. Žitnik, An integrated explicit and implicit offensive language taxonomy, *Lodz Pap. Pragmat.* 19 (1) (2023) 7–48.
- [78] T. Garg, S. Masud, T. Suresh, T. Chakraborty, Handling bias in toxic speech detection: A survey, *ACM Comput. Surv.* 55 (13s) (2023) 1–32.
- [79] A. Bhandari, S.B. Shah, S. Thapa, U. Naseem, M. Nasim, CrisisHateMM: Multimodal analysis of directed and undirected hate speech in text-embedded images from Russia-Ukraine conflict, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1993–2002.
- [80] R. Bhowmick, I. Ganguli, J. Paul, J. Sil, A multimodal deep framework for derogatory social media post identification of a recognized person, *ACM Trans. Asian Low- Resour. Lang. Inf. Process.* 21 (1) (2022) 3447651, <http://dx.doi.org/10.1145/3447651>.
- [81] J. Qu, L.H. Li, J. Zhao, S. Dev, K.-W. Chang, Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation, 2022, arXiv preprint [arXiv:2205.12617](https://arxiv.org/abs/2205.12617).
- [82] N. Zhang, X. Feng, T. Gu, L. Chang, MVLP: Multi-perspective vision-language pre-training model for ethically aligned meme detection, *Authorea Prepr.* (2023).
- [83] G. Kumari, D. Bandyopadhyay, A. Ekbal, EmoffMeme: identifying offensive memes by leveraging underlying emotions, in: *Multimedia Tools and Applications*, vol. 82, 2023, pp. 45061–45096, <http://dx.doi.org/10.1007/s11042-023-14807-1>.
- [84] S. Sharma, M.S. Akhtar, P. Nakov, T. Chakraborty, DISARM: Detecting the victims targeted by harmful memes, in: Findings of the Association for Computational Linguistics: NAACL 2022 - Findings, 2022, pp. 1572–1588, arXiv preprint [arXiv:2205.05738](https://arxiv.org/abs/2205.05738).
- [85] B. Jabiyev, J. Onaolapo, G. Stringhini, E. Kirda, E-game of FAME: Automatic detection of fake memes., in: *TTO*, 2021, pp. 1–11.
- [86] L. Mathias, S. Nie, A.M. Davani, D. Kiela, V. Prabhakaran, B. Vidgen, Z. Waseem, Findings of the WOAHS 5 shared task on fine grained hateful memes detection, in: Proceedings of the 5th Workshop on Online Abuse and Harms, WOAHS 2021, 2021, pp. 201–206.
- [87] H. Lin, Z. Luo, B. Wang, R. Yang, J. Ma, GOAT-bench: Safety insights to large multimodal models through meme-based social abuse, 2024, arXiv preprint [arXiv:2401.01523](https://arxiv.org/abs/2401.01523).
- [88] S. Pramanick, S. Sharma, D. Dimitrov, P. Nakov, T. Chakraborty, MOMENTA: A multimodal framework for detecting harmful memes and their targets, 2021, arXiv.
- [89] B.O. Sabat, C.C. Ferrer, X. Giro-I-Nieto, Hate speech in pixels: Detection of offensive memes towards automatic moderation, 2019, arXiv.
- [90] M. Hee, W.-H. Chong, R.-W. Lee, Decoding the underlying meaning of multimodal hateful memes, in: IJCAI International Joint Conference on Artificial Intelligence, vol. 2023-August, 2023.
- [91] S. Sharma, T. Suresh, A. Kulkarni, H. Mathur, P. Nakov, M.S. Akhtar, T. Chakraborty, Findings of the constraint 2022 shared task on detecting the hero, the villain, and the victim in memes, in: Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situations, Constraint, 2022, pp. 1–11.
- [92] K. Rajput, R. Kapoor, K.K. Rai, P. Kaur, Hate me not: Detecting hate inducing memes in code switched languages, 2022, arXiv.
- [93] J. Badour, J. Brown, Hateful memes classification using machine learning, in: 2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 - Proceedings, vol. 2, 2021, <http://dx.doi.org/10.1109/ssci50451.2021.9659896>.
- [94] J. Bacha, F. Ullah, J. Khan, A. Sardar, S. Lee, A deep learning-based framework for offensive text detection in unstructured data for heterogeneous social media, in: *IEEE Access*, vol. 11, 2023, pp. 124484–124498, <http://dx.doi.org/10.1109/access.2023.3330081>.
- [95] A. Alzu'bi, L. Bani Younis, A. Abuarqoub, M. Hammoudeh, Multimodal deep learning with discriminant descriptors for offensive memes detection, *ACM J. Data Inf. Qual.* 15 (3) (2023) 1–16.
- [96] S. Ramamoorthy, N. Gunti, S. Mishra, S. Suryavardan, A. Reganti, P. Patwa, A. DaS, T. Chakraborty, A. Sheth, A. Ekbal, C. A. Memotion 2: Dataset on sentiment and emotion analysis of memes, in: Proceedings of de-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.
- [97] B. Xu, T. Li, J. Zheng, M. Naseriparsa, Z. Zhao, H. Lin, F. Xia, MET-meme: A multimodal meme dataset rich in metaphors, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2887–2899.
- [98] F. Gasparini, G. Rizzi, A. Saibene, E. Fersini, Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content, *Data Brief* 44 (2022) 108526.
- [99] K. Maity, P. Jha, S. Saha, P. Bhattacharyya, A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multimodal code-mixed memes, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1739–1749.
- [100] P. Jha, K. Maity, R. Jain, A. Verma, S. Saha, P. Bhattacharyya, Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations, 2024, arXiv preprint [arXiv:2401.09899](https://arxiv.org/abs/2401.09899).
- [101] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval-2022, 2022, pp. 533–549.
- [102] S. Suryawanshi, B.R. Chakravarthi, M. Arcan, P. Buitelaar, Multimodal meme dataset (multioff) for identifying offensive content in image and text, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 32–41.
- [103] G. Kumari, A. Sinha, A. Ekbal, A. Chatterjee, B. Vinutha, Enhancing the fairness of offensive memes detection models by mitigating unintended political bias, *J. Intell. Inf. Syst.* (2024) 1, <http://dx.doi.org/10.1007/s10844-023-00834-9>.
- [104] S. Suryawanshi, B.R. Chakravarthi, P. Verma, M. Arcan, J.P. McCrae, P. Buitelaar, A dataset for troll classification of TamilMemes, in: Proceedings of the WILDRE5-5th Workshop on Indian Language Data: Resources and Evaluation, 2020, pp. 7–13.
- [105] E. Fersini, G. Rizzi, A. Saibene, F. Gasparini, Misogynous MEME recognition: A preliminary study, in: Lecture Notes in Computer Science, in: Lecture Notes in Computer Science, vol. 13196, 2022, pp. 279–293, http://dx.doi.org/10.1007/978-3-031-08421-8_19.
- [106] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: *Advances in Neural Information Processing Systems*, vol. 2020-December, 2020.
- [107] S.R. Titli, S. Paul, Automated bengali abusive text classification: Using deep learning techniques, in: 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems, ICAECIS, Ieee, 2023, pp. 1–6.
- [108] M. Das, A. Mukherjee, Transfer learning for multilingual abusive meme detection, in: *ACM International Conference Proceeding Series*, 2023, pp. 245–250, <http://dx.doi.org/10.1145/3578503.3583607>.
- [109] P.K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, N. Tippet, Cyberbullying: Its nature and impact in secondary school pupils, *J. Child Psychol. Psychiatry* 49 (4) (2008) 376–385.
- [110] M.N. Kumar, D.M. Ahmed, J. Prashanth, V. Vinaykumar, J.A. Babu, T.K. Kumar, An efficient deep learning approach to deal with cyberbullying, in: 2023 2nd International Conference on Computational Modelling, Simulation and Optimization, ICCMSO, IEEE, 2023, pp. 253–258.

- [111] H. Lin, Z. Luo, W. Gao, J. Ma, B. Wang, R. Yang, Towards explainable harmful meme detection through multimodal debate between large language models, 2024, arXiv preprint [arXiv:2401.13298](https://arxiv.org/abs/2401.13298).
- [112] H. Lin, Z. Luo, J. Ma, L. Chen, Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models, in: Findings of the Association for Computational Linguistics: EMNLP 2023, vol. 9114, 2023, arXiv preprint [arXiv:2312.05434](https://arxiv.org/abs/2312.05434).
- [113] J. Ji, W. Ren, U. Naseem, Identifying creative harmful memes via prompt based approach, in: ACM Web Conference 2023 - Proceedings of the World Wide Web Conference, WWW 2023, vol. 3868, 2023.
- [114] S. Farook, S. Ahmed, G. Rithika, S. Budde, S. Saumya, S. Biradar, Are you a hero or a villain? A semantic role labelling approach for detecting harmful memes, in: CONSTRAINT 2022 - 2nd Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situation, Proceedings of the Workshop, Constraint, 2022, pp. 19–23.
- [115] T. Chakraborty, M.S. Akhtar, K. Shu, H.R. Bernard, M. Liakata, P. Nakov (Eds.), Constraint 2022 - 2nd workshop on combating online hostile posts in regional languages during emergency situation, proceedings of the workshop, 2022, p. 112.
- [116] R. Nandi, F. Alam, P. Nakov, Detecting the role of an entity in harmful memes: Techniques and their limitations, in: CONSTRAINT 2022 - 2nd Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, Proceedings of the Workshop, Constraint, 2022, pp. 43–54.
- [117] S. Sharma, A. Kulkarni, T. Suresh, H. Mathur, P. Nakov, M. Akhtar, T. Chakraborty, Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? in: EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, 2023, pp. 2141–2155.
- [118] P. Singh, A. Maladry, E. Lefever, Combining language models and linguistic information to label entities in memes, in: CONSTRAINT 2022 - 2nd Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situation, Proceedings of the Workshop, Constraint, 2022, pp. 35–42.
- [119] Z. Zhou, H. Zhao, J. Dong, J. Gao, X. Liu, DD-TIG at constraint2022: Multimodal understanding and reasoning for role labeling of entities in hateful memes, in: Proceedings of CONSTRAINT 2022 - 2nd Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situation, Constraint, 2022, pp. 12–18.
- [120] S. Sharma, M.K. Siddiqui, M.S. Akhtar, T. Chakraborty, Domain-aware self-supervised pre-training for label-efficient meme analysis, 2022, arXiv.
- [121] C. Koutlis, M. Schinas, S. Papadopoulos, MemeFier: Dual-stage modality fusion for image meme, in: ICMR 2023 - Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, 2023, pp. 586–591, [http://dx.doi.org/10.1145/3591106.3592254](https://dx.doi.org/10.1145/3591106.3592254).
- [122] H.B. Zia, I. Castro, G. Tyson, Racist or sexist meme? classifying memes beyond hateful, in: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), vol. 215, 2021, pp. 215–219.
- [123] J. Armenta-Segura, C.-J. Núñez Prado, G. Sidorov, A. Gelbukh, R. Román-Godínez, Omteotmultimodal hate speech event detection 2023: Hate speech and text-image correlation detection in real life memes using pre-trained BERT models over text, in: CASE 2023 - Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text At RANLP, vol. 53, 2023, [http://dx.doi.org/10.26615/978-954-452-089-2_007](https://dx.doi.org/10.26615/978-954-452-089-2_007).
- [124] A. Aggarwal, V. Sharma, A. Trivedi, M. Yadav, C. Agrawal, D. Singh, V. Mishra, H. Gritli, Two-way feature extraction using sequential and multimodal approach for hateful meme classification, Complexity 2021 (2021) [http://dx.doi.org/10.1155/2021/5510253](https://dx.doi.org/10.1155/2021/5510253).
- [125] G. Arya, M. Hasan, A. Bagwari, N. Safie, S. Islam, F. Ahmed, A. De, M. Khan, T. Ghazal, Multimodal hate speech detection in memes using contrastive language-image pre-training, IEEE Access 12 (2024) 22359–22375, [http://dx.doi.org/10.1109/access.2024.3361322](https://dx.doi.org/10.1109/access.2024.3361322).
- [126] P. Aggarwal, P. Chawla, M. Das, P. Saha, B. Mathew, T. Zesch, A. Mukherjee, HateProof: Are hateful meme detection systems really robust? in: ACM Web Conference 2023 - Proceedings of the World Wide Web Conference, WWW 2023, in: vol. 3734, 2023, [http://dx.doi.org/10.1145/3543507.3583356](https://dx.doi.org/10.1145/3543507.3583356).
- [127] P. Aggarwal, M. Liman, D. Gold, T. Zesch, VL-bert+: Detecting protected groups in hateful multimodal memes, in: WOAH 2021 - 5th Workshop on Online Abuse and Harms, Proceedings of the Workshop, vol. 207, 2021.
- [128] M. Ahmed, N. Bhadani, I. Chakraborty, Hateful meme prediction model using multimodal deep learning, in: 2021 International Conference on Computing, Communication and Green Engineering, CCGE 2021, vol. 2, 2021, [http://dx.doi.org/10.1109/ccge50943.2021.9776440](https://dx.doi.org/10.1109/ccge50943.2021.9776440).
- [129] A. Bhat, V. Varshney, V. Bajlotra, V. Gupta, Detection of hatefulness in memes using unimodal and multimodal techniques, in: Proceedings - 2022 6th International Conference on Intelligent Computing and Control Systems, ICIACS 2022, vol. 65, 2022, [http://dx.doi.org/10.1109/iciacs53718.2022.9788376](https://dx.doi.org/10.1109/iciacs53718.2022.9788376).
- [130] A. Bhat, V. Vashisht, V. Sahni, S. Meena, Hate speech detection using multimodal meme analysis, in: Proceedings of the 2nd International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2023, Vol. 1137, 2023, [http://dx.doi.org/10.1109/icaaic56838.2023.10140393](https://dx.doi.org/10.1109/icaaic56838.2023.10140393).
- [131] N. Bi, Y.-C. Huang, C.-C. Han, J.-J. Hsu, You know what I meme: Enhancing People's understanding and awareness of hateful memes using crowdsourced explanations, in: Proceedings of the ACM on Human-Computer Interaction, vol. 7, 2023, [http://dx.doi.org/10.1145/3579593](https://dx.doi.org/10.1145/3579593).
- [132] E. Blaier, I. Malkiel, L. Wolf, Caption enriched samples for improving hateful memes detection, in: EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings, vol. 9350, 2021.
- [133] R. Cao, R.-W. Lee, W.-H. Chong, J. Jiang, Prompting for multimodal hateful meme classification, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Vol. 321, EMNLP 2022, 2022.
- [134] R. Cao, M. Hee, A. Kuek, W.-H. Chong, R.-W. Lee, J. Jiang, Pro-cap: Leveraging a frozen vision-language model for hateful meme detection, in: MM 2023 - Proceedings of the 31st ACM International Conference on Multimedia, vol. 5244, Association for Computing Machinery, 2023, [http://dx.doi.org/10.1145/3581783.3612498](https://dx.doi.org/10.1145/3581783.3612498).
- [135] A. Chhabra, D. Vishwakarma, Multimodal hate speech detection via multi-scale visual kernels and knowledge distillation architecture, Eng. Appl. Artif. Intell. 126 (2023) [http://dx.doi.org/10.1016/j.engappai.2023.106991](https://dx.doi.org/10.1016/j.engappai.2023.106991).
- [136] M. Constantin, D.-S. Parvu, C. Stanciu, D. Ionascu, B. Ionescu, Hateful meme detection with multimodal deep neural networks, in: ISSCS 2021 - International Symposium on Signals, Circuits and Systems, vol. 9497374, 2021, [http://dx.doi.org/10.1109/isscs52333.2021.9497374](https://dx.doi.org/10.1109/isscs52333.2021.9497374).
- [137] T. Deshpande, N. Mani, An interpretable approach to hateful meme detection, in: ICMI 2021 - Proceedings of the 2021 International Conference on Multimodal Interaction, vol. 723, 2021, [http://dx.doi.org/10.1145/3462244.3479949](https://dx.doi.org/10.1145/3462244.3479949).
- [138] H. Fang, F. Zhu, J. Han, S. Hu, Multimodal hateful memes detection via image caption supervision, in: Proceedings - 2022 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Autonomous and Trusted Vehicles, Scalable Computing and Communications, Digital Twin, Privacy Computing, Metaverse, SmartWorld/UIC/ATC/ScalCom/DigitalTwin/PriComp/Metaverse 2022, vol. 1530, 2022, [http://dx.doi.org/10.1109/SmartWorld-UIC-ATC-ScalCom-DigitalTwin-PriComp-Metaverse56740.2022.00221](https://dx.doi.org/10.1109/SmartWorld-UIC-ATC-ScalCom-DigitalTwin-PriComp-Metaverse56740.2022.00221).
- [139] B. Gaikwad, B. Kurma, M. Patwardhan, S. Karande, N. Pedanekar, Can a pretrained language model make sense with pretrained neural extractors? An application to multimodal classification, in: CEUR Workshop Proceedings, vol. 3168, 2022, [http://dx.doi.org/10.1109/iceic57457.2023.10049865](https://dx.doi.org/10.1109/iceic57457.2023.10049865).
- [140] A. Goswami, A. Rawat, S. Tongaria, S. Jhingran, Detection of hate speech in multi-modal social post, in: Artificial Intelligence, Blockchain, Computing and Security: Volume 1, vol. 1, 2023, [http://dx.doi.org/10.1201/9781003393580-50](https://dx.doi.org/10.1201/9781003393580-50).
- [141] M. Hee, R.-W. Lee, W.-H. Chong, On explaining multimodal hateful meme detection models, in: WWW 2022 - Proceedings of the ACM Web Conference 2022, vol. 3651, 2022, [http://dx.doi.org/10.1145/3485447.3512260](https://dx.doi.org/10.1145/3485447.3512260).
- [142] A. Kiran, M. Shetty, S. Shukla, V. Kerenallii, B. Das, Getting around the semantics challenge in hateful memes, in: Lecture Notes on Data Engineering and Communications Technologies, vol. 142, 2023, [http://dx.doi.org/10.1007/978-981-19-3391-2_26](https://dx.doi.org/10.1007/978-981-19-3391-2_26).
- [143] V. Kougia, J. Pavlopoulos, Multimodal or text? Retrieval or bert? Benchmarking classifiers for the shared task on hateful memes, in: WOAH 2021 - 5th Workshop on Online Abuse and Harms, Proceedings of the Workshop, vol. 220, 2021.
- [144] V. Kougia, S. Fetzl, T. Kirchmair, E. Çano, S. Baharlou, S. Sharifzadeh, B. Roth, MemeGraphs: Linking memes to knowledge graphs, in: Lecture Notes in Computer Science, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 14187, Springer, 2023, pp. 534–551, [http://dx.doi.org/10.1007/978-3-031-41676-7_31](https://dx.doi.org/10.1007/978-3-031-41676-7_31).
- [145] R.K.-W. Lee, R. Cao, Z. Fan, J. Jiang, W.-H. Chong, Disentangling hate in online memes, in: Proceedings of the 29th ACM International Conference on Multimedia, vol. 5138, Association for Computing Machinery, 2021, pp. 5138–5147, [http://dx.doi.org/10.1145/3474085.3475625](https://dx.doi.org/10.1145/3474085.3475625).
- [146] X. Liang, Y.-C. Huang, W. Liu, H. Zhu, Z. Liang, L. Chen, TRICAN: Multimodal hateful memes detection with triplet-relation information cross-attention network, in: Proceedings of the International Joint Conference on Neural Networks, vol. 2022-July, 2022, [http://dx.doi.org/10.1109/ijcnn55064.2022.9892164](https://dx.doi.org/10.1109/ijcnn55064.2022.9892164).
- [147] Z. Ma, S. Yao, L. Wu, S. Gao, Y. Zhang, Hateful memes detection based on multi-task learning, Mathematics 10 (2022) [http://dx.doi.org/10.3390/math10234525](https://dx.doi.org/10.3390/math10234525).
- [148] G. MacRayo, W. Casino, J. Dalangin, J. Gabriel Gahoy, A. Christian Reyes, C. Vitto, M. Abisado, S. Lor Huyo-A, G. Avelino Sampedro, Please be nice: A deep learning based approach to content moderation of internet memes, in: 2023 International Conference on Electronics, Information, and Communication, ICEIC 2023, Ieee, 2023, pp. 1–5, [http://dx.doi.org/10.1109/iceic57457.2023.10049865](https://dx.doi.org/10.1109/iceic57457.2023.10049865).
- [149] L. Mookdarsanit, P. Mookdarsanit, Combating the hate speech in thai textual memes, Indones. J. Electr. Eng. Comput. Sci. 21 (2021) 1493–1502, [http://dx.doi.org/10.11591/ijeecs.v21.i3.pp1493-1502](https://dx.doi.org/10.11591/ijeecs.v21.i3.pp1493-1502).

- [150] A. Nayak, A. Agrawal, Detection of hate speech in social media memes: A comparative analysis, in: Proceedings of the 2022 3rd International Conference on Intelligent Computing, Instrumentation and Control Technologies: Computational Intelligence for Smart Systems, ICICIT 2022, vol. 1179, 2022, <http://dx.doi.org/10.1109/icicict54557.2022.9917633>.
- [151] Y. Qu, X. He, S. Pierson, M. Backes, Y. Zhang, S. Zannettou, On the evolution of (hateful) memes by means of multimodal contrastive learning, in: Proceedings - IEEE Symposium on Security and Privacy, vol. 2023-May, 2023, <http://dx.doi.org/10.1109/sp46215.2023.10179315>.
- [152] Y. Qu, X. Shen, X. He, M. Backes, S. Zannettou, Y. Zhang, Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models, in: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, Ccs '23, Association for Computing Machinery, 2023, pp. 3403–3417, <http://dx.doi.org/10.1145/3576915.3616679>.
- [153] A. Sethi, U. Kuchhal, Anjum, R. Kataraya, Study of various techniques for the classification of hateful memes, in: 2021 6th International Conference on Recent Trends on Electronics, Information, Communication and Technology, RTEICT 2021, vol. 675, 2021, <http://dx.doi.org/10.1109/rteict52294.2021.9573926>.
- [154] L. Wanbo, L. Suiyng, Research on multi-modal hateful meme detection, in: ACM International Conference Proceeding Series, vol. 3470385, 2021, <http://dx.doi.org/10.1145/3469213.3470385>.
- [155] P. Wu, W. Mebane, MARMOT a deep learning framework for constructing multimodal representations for vision-and-language tasks, Comput. Commun. Res. 4 (2022) <http://dx.doi.org/10.5117/ccr2022.1.008.wu>.
- [156] L. Zhang, L. Jin, X. Sun, G. Xu, Z. Zhang, X. Li, N. Liu, Q. Liu, S. Yan, TOT: Topology-aware optimal transport for multimodal hate detection, in: Proceedings of the 37th AAAI Conference on Artificial Intelligence, Vol. 37, AAAI 2023, 2023.
- [157] Y. Zhou, Z. Chen, H. Yang, Multimodal learning for hateful memes detection, in: 2021 IEEE International Conference on Multimedia and Expo Workshops, vol. 28, 2021, <http://dx.doi.org/10.1109/icmew53276.2021.9455994>.
- [158] J. Zhu, R.-W. Lee, W. Chong, Multimodal zero-shot hateful meme detection, in: ACM International Conference Proceeding Series, vol. 382, 2022, <http://dx.doi.org/10.1145/3501247.3531557>.
- [159] P. Aggarwal, J. Mehrabian, W. Huang, Ö. Alacam, T. Zesch, Text or image? What is more important in cross-domain generalization capabilities of hate meme detection models?, 2024, arXiv.
- [160] Y. Chen, F. Pan, Multimodal detection of hateful messages using visual-linguistic pre-trained deep learning models, Res. Sq. (2022).
- [161] A. Das, J.S. Wahi, S. Li, Detecting hate speech in multi-modal memes, 2020, arXiv.
- [162] I. Evtimov, R. Howes, B. Dolhansky, H. Firooz, C.C. Ferrer, Adversarial evaluation of multimodal models under realistic gray box assumptions, 2020, arXiv.
- [163] A. Gao, B. Wang, J. Yin, Y. Tian, Hateful memes challenge: An enhanced multimodal framework, 2021, arXiv.
- [164] C. Jennifer, F. Tahmasbi, J. Blackburn, S. Zannettou, E. De Cristofaro, Feels bad man: Dissecting automated hateful meme detection through the lens of facebook's challenge, 2022, arXiv.
- [165] W. Jin, L. Wilhelm, The hateful memes challenge next move, 2022, arXiv.
- [166] Y. Li, Z. Zhang, H. Huang, Enhance multimodal model performance with data augmentation: facebook hateful meme challenge solution, 2021, arXiv.
- [167] P. Lippe, N. Holla, S. Chandra, S. Rajamanickam, G. Antoniou, E. Shutova, H. Yannakoudakis, A multimodal framework for the detection of hateful memes, 2020, arXiv preprint [arXiv:2012.12871](https://arxiv.org/abs/2012.12871).
- [168] J. Mei, J. Chen, W. Lin, B. Byrne, M. Tomalin, Improving hateful memes detection via learning helpfulness-aware embedding space through retrieval-guided contrastive learning, 2023, arXiv.
- [169] Y. Miyaniishi, M. Le Nguyen, Causal intersectionality and dual form of gradient descent for multimodal analysis: a case study on hateful memes, 2023, arXiv.
- [170] N. Muennighoff, Vilio: State-of-the-art visio-linguistic models applied to hateful memes, 2020, arXiv preprint [arXiv:2012.07788](https://arxiv.org/abs/2012.07788).
- [171] V. Sandulescu, Detecting hateful memes using a multimodal deep ensemble, 2020, arXiv preprint [arXiv:2012.13235](https://arxiv.org/abs/2012.13235).
- [172] M.-H. Van, X. Wu, Detecting and correcting hate speech in multimodal memes with large visual language model, 2023, arXiv.
- [173] R. Velioglu, J. Rose, Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge, 2020, arXiv preprint [arXiv:2012.12975](https://arxiv.org/abs/2012.12975).
- [174] J. Wang, J. Luo, G. Yang, A. Hong, F. Luo, Is GPT powerful enough to analyze the emotions of memes?, 2023, arXiv preprint [arXiv:2311.00223](https://arxiv.org/abs/2311.00223).
- [175] Y. Wu, N. Yu, M. Backes, Y. Shen, Y. Zhang, On the proactive generation of unsafe images from text-to-image models using benign prompts, 2023, arXiv: [2310.16613](https://arxiv.org/abs/2310.16613).
- [176] J. Yuan, Y. Yu, G. Mittal, S. Sajeev, M. Chen, Rethinking multimodal content moderation from an asymmetric angle with mixed-modality, 2023, arXiv.
- [177] W. Zhang, G. Liu, Z. Li, F. Zhu, Hateful memes detection via complementary visual and linguistic networks, 2020, arXiv.
- [178] B. Zhao, A. Zhang, B. Watson, G. Kearney, I. Dale, A review of vision-language models and their performance on the hateful memes challenge, 2023, arXiv.
- [179] X. Zhong, Classification of multimodal hate speech -the winning solution of hateful memes challenge, 2020, arXiv.
- [180] Y. Zhou, Z. Chen, Multimodal learning for hateful memes detection, 2020, arXiv.
- [181] R. Zhu, Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution, 2020, arXiv preprint [arXiv:2012.08290](https://arxiv.org/abs/2012.08290).
- [182] G. Burbi, A. Baldrati, L. Agnolucci, M. Bertini, A. Del Bimbo, Mapping memes to words for multimodal hateful meme classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2832–2836, <http://dx.doi.org/10.1109/iccv67093.2023.00303>.
- [183] E. Fersini, F. Gasparini, S. Corchs, Detecting sexist meme on the web: A study on textual and visual cues, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW, Ieee, 2019, pp. 226–231.
- [184] G. Attanasio, D. Nozza, F. Bianchi, Milanlp at SemEval-2022 task 5: Using perceiver IO for detecting misogynous memes with text and image modalities, in: Proceedings of SemEval 2022 - 16th International Workshop on Semantic Evaluation, 2022, pp. 654–662.
- [185] M. Behzadi, A. Derakhshan, I. Harris, Mitra behzadi at SemEval-2022 task 5: Multimedia automatic misogyny identification method based on CLIP, in: Proceedings of SemEval 2022 - 16th International Workshop on Semantic Evaluation, 2022, pp. 724–727.
- [186] L. Chen, H. Chou, RIT boston at SemEval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from CLIP model and data-centric AI principle, in: Proceedings of SemEval 2022 - 16th Intl. Workshop on Semantic Evaluation, 2022, pp. 636–641.
- [187] Y. Gu, I. Castro, G. Tyson, MMVAE at SemEval-2022 task 5: A multi-modal multi-task VAE on misogynous meme detection, in: Proceedings of SemEval 2022 - 16th International Workshop on Semantic Evaluation, 2022, pp. 700–710.
- [188] M. Kalkenings, T. Mandl, University of hildesheim at SemEval-2022 task 5: Combining deep text and image models for multimedia misogyny detection, in: Proceedings of SemEval 2022 - 16th International Workshop on Semantic Evaluation, 2022, pp. 718–723.
- [189] D. Obeidat, H. Nammam, M. Abdullah, Just one at SemEval-2023 task 10: Explainable detection of online sexism (EDOS), in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 526–531.
- [190] A. Paraschiv, M. Dascalu, D.-C. Cercel, UPB at SemEval-2022 task 5: Enhancing UNITER with image sentiment and graph convolutional networks for multimedia automatic misogyny identification, in: Proceedings of SemEval 2022 - 16th International Workshop on Semantic Evaluation, 2022, pp. 618–625.
- [191] J. Ravagli, L. Vaiani, JRLV at SemEval-2022 task 5: The importance of visual elements for misogyny identification in memes, in: SemEval 2022 - 16th International Workshop on Semantic Evaluation, Proceedings of the Workshop, 2022, pp. 610–617.
- [192] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing misogynous memes: Biased models and tricky archetypes, Inf. Process. Manag. 60 (5) (2023) 103474, <http://dx.doi.org/10.1016/j.ipm.2023.103474>.
- [193] S. Singh, A. Haridasan, R. Mooney, “Female astronaut: Because sandwiches won't make themselves up there!": Towards multi-modal misogyny detection in memes, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2023, pp. 150–159.
- [194] P. Tung, N. Viet, N. Anh, P. Hung, SemiMemes: A semi-supervised learning approach for multimodal memes analysis, in: Lecture Notes in Computer Science, in: Lecture Notes in Computer Science, vol. 14162, 2023, pp. 565–577, http://dx.doi.org/10.1007/978-3-031-41456-5_43.
- [195] N.K. Singh, P. Das, A. Manderna, S. Chand, DeVi deep learning framework for misogyny identification in multimodal data, Res. Sq. (2023).
- [196] J. Drakett, B. Rickett, K. Day, K. Milnes, Old jokes, new media—online sexism and constructions of gender in internet memes, Fem. Psychol. 28 (1) (2018) 109–127.
- [197] A. Alzu'bi, L. Bani Younis, A. Abuarqoub, M. Hammoudeh, Multimodal deep learning with discriminant descriptors for offensive memes detection, 15, 2023, 3597308, <http://dx.doi.org/10.1145/3597308>.
- [198] A. Aman, G. Krishna, T. Anand, A. Lal, Identification of offensive content in memes, in: Lecture Notes in Networks and Systems, vol. 290, 2021, pp. 438–445, http://dx.doi.org/10.1007/978-981-16-4486-3_49.
- [199] A. Baruah, K. Das, F. Barbhuiya, K. Dey, IIITG-ADBU at SemEval-2020 task 8: A multimodal approach to detect offensive, sarcastic and humorous memes, in: Proceedings of SemEval 2020 - 14th International Workshops on Semantic Evaluation, 2020, pp. 885–890.
- [200] I. Bejan, Memosys at SemEval-2020 task 8: Multimodal emotion analysis in memes, in: Proceedings of SemEval 2020 - 14th International Workshops on Semantic Evaluation, 2020, pp. 1172–1178.
- [201] S. Boinepelli, M. Shrivastava, V. Varma, SISIITH at SemEval-2020 task 8: An overview of simple text classification methods for meme analysis, in: Proceedings of SemEval 2020 - 14th International Workshops on Semantic Evaluation, 2020, pp. 1190–1194.

- [202] A.-M. Bucur, A. Cosma, I.-B. Iordache, BLUE at memotion 2.0 2022: You have my image, my text and my transformer, in: CEUR Workshop Proceedings, vol. 3168, 2022.
- [203] V. Sharma, V. Kushwaha, S. Jaiswal, G. Nandi, Meme detection for sentiment analysis and human robot interactions using multiple modes, in: 9th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering, UPCon 2022, 2022, <http://dx.doi.org/10.1109/upcon56432.2022.9986453>.
- [204] G. de la Peña Sarracén, P. Rosso, A. Giachanou, PRHLT-UPV at SemEval-2020 task 8: Study of multimodal techniques for memes analysis, in: 14th International Workshops on Semantic Evaluation, SemEval 2020 - Co-located 28th International Conference on Computational Linguistics, COLING 2020, Proceedings of the Workshop, 2020, pp. 908–915.
- [205] R. Giri, S. Gupta, U. Gupta, An approach to detect offence in memes using natural language processing(NLP) and deep learning, in: 2021 International Conference on Computer Communication and Informatics, ICCCI 2021, 2021, 9402406, <http://dx.doi.org/10.1109/iccci50826.2021.9402406>.
- [206] A. Gupta, H. Kataria, S. Mishra, T. Badal, V. Mishra, BennettNLP at SemEval-2020 task 8: Multimodal sentiment classification using hybrid hierarchical classifier, in: Proceedings of SemEval 2020 - 14th International Workshops on Semantic Evaluation, 2020, pp. 1085–1093.
- [207] S. Hakimov, G. Cheema, R. Ewerth, TIB-VA at SemEval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes, in: Proceedings of SemEval 2022 - 16th International Workshop on Semantic Evaluation, 2022, pp. 756–760.
- [208] E. Hossain, O. Sharif, M. Hoque, M. Akber Dewan, N. Siddique, M. Hossain, Identification of multilingual offense and troll from social media memes using weighted ensemble of multimodal features, 34, 2022, pp. 6605–6623, <http://dx.doi.org/10.1016/j.jksuci.2022.06.010>,
- [209] E. Hossain, M. Hoque, M. Hossain, An inter-modal attention framework for multimodal offense detection, in: Lecture Notes in Networks and Systems, vol. 569 Lnn, 2023, pp. 853–862, http://dx.doi.org/10.1007/978-3-031-19958-5_81.
- [210] K. Mylvahanan, B. Shashank, T. Raj, C. Attanti, S. Sahay, A study on deep learning based classification and identification of offensive memes, in: Proceedings of the 3rd International Conference on Trends in Electronics and Informatics, ICOEI 2019, 2023, pp. 214–218, <http://dx.doi.org/10.1109/icoei.2019.8862647>.
- [211] T. Nguyen, N. Pham, N. Nguyen, H. Nguyen, L. Nguyen, Y.-G. Kim, Hcilab at memotion 2.0 2022: Analysis of sentiment, emotion and intensity of emotion classes from meme images using single and multi modalities, in: CEUR Workshop Proceedings, vol. 3168, 2022.
- [212] K. Phan, G.-S. Lee, H.-J. Yang, S.-H. Kim, Little flower at memotion 2.0 2022: Ensemble of multi-modal model using attention mechanism in MEMOTION analysis, in: CEUR Workshop Proceedings, vol. 3168, 2022.
- [213] L. Shang, Y. Zhang, Y. Zha, KnowMeme: A knowledge-enriched graph neural network solution to offensive meme detection, in: Proceedings - IEEE 17th International Conference on EScience, EScience 2021, 2021, pp. 186–195, <http://dx.doi.org/10.1109/escience51609.2021.00029>.
- [214] U. Walinska, J. Potoniec, Urszula walińska at SemEval-2020 task 8: Fusion of text and image features using LSTM and VGG16 for memotion analysis, in: Proceedings of SemEval 2020 - 14th International Workshops on Semantic Evaluation, 2020, pp. 1215–1220.
- [215] W. Yu, D. Kolossa, Wentao at memotion 3: Ensemble learning for multi-modal meme classification, in: CEUR Workshop Proceedings, vol. 3555, 2022.
- [216] Q. Zhong, Q. Wang, J. Liu, Combining knowledge and multi-modal fusion for meme classification, in: Lecture Notes in Computer Science, vol. 13141, 2022, pp. 599–611, http://dx.doi.org/10.1007/978-3-030-98358-1_47.
- [217] S. Pramanick, M.S. Akhtar, T. Chakraborty, Exercise? I thought you said 'extra fries': Leveraging sentence demarcations and multi-hop attention for meme affect analysis, 2021, arXiv.
- [218] D. Gaurav, S. Shandilya, S. Tiwari, A. Goyal, A machine learning method for recognizing invasive content in memes, in: Communications in Computer and Information Science, 2020, pp. 195–213, http://dx.doi.org/10.1007/978-3-030-65384-2_15.
- [219] S. Gundapu, R. Mamidi, Detection of propaganda techniques in visuo-lingual metaphor in memes, 2022, arXiv.
- [220] D. Abujaber, A. Qarqaz, M. Abdullah, Lecun at SemEval-2021 task 6: Detecting persuasion techniques in text using ensemble pretrained transformers and data augmentation, in: Proceedings of SemEval 2021 - 15th Intl. Workshop on Semantic Evaluation, 2021, pp. 1068–1074.
- [221] F. Alam, H. Mubarak, W. Zaghouani, G. Da San Martino, P. Nakov, Overview of the WANLP 2022 shared task on propaganda detection in arabic, in: WANLP 2022 - 7th Arabic Natural Language Processing - Proceedings of the Workshop, 2022, pp. 108–118.
- [222] J. Cui, L. Li, X. Zhang, J. Yuan, Multimodal propaganda detection via anti-persuasion prompt enhanced contrastive learning, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2023, <http://dx.doi.org/10.1109/icassp49357.2023.10096771>.
- [223] T. Hossain, J. Naim, F. Tasneem, R. Tasnia, A. Chy, CSECU-DSG at SemEval-2021 task 6: Orchestrating multimodal neural architectures for identifying persuasion techniques in texts and images, in: Proceedings of SemEval 2021 - 15th International Workshop on Semantic Evaluation, 2021, pp. 1088–1095.
- [224] P. Li, X. Li, X. Sun, 1213Li at SemEval-2021 task 6: Detection of propaganda with multi-modal attention and pre-trained models, in: SemEval 2021 - 15th International Workshop on Semantic Evaluation, Proceedings of the Workshop, 2021, pp. 1032–1036.
- [225] C. Liu, G. Geigle, R. Krebs, I. Gurevych, FigMemes: A dataset for figurative language identification in politically-opinionated memes, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 7069–7086.
- [226] H. Wu, X. Li, L. Li, Q. Wang, Propaganda techniques detection in low-resource memes with multi-modal prompt tuning, in: Proceedings - IEEE International Conference on Multimedia and Expo, 2022, <http://dx.doi.org/10.1109/icme52920.2022.9859642>.
- [227] X. Zhu, J. Wang, X. Zhang, YNU-HPCC at SemEval-2021 task 6: Combining ALBERT and text-CNN for persuasion detection in texts and images, in: SemEval 2021 - 15th International Workshop on Semantic Evaluation, Proceedings of the Workshop, 2021, pp. 1045–1050.
- [228] P. Chen, L. Zhao, Y. Piao, H. Ding, X. Cui, Multimodal visual-textual object graph attention network for propaganda detection in memes, Multimedia Tools Appl. 1 (1) (2023) 1–10, <http://dx.doi.org/10.1007/s11042-023-15272-6>.
- [229] J. Cui, L. Li, X. Tao, Be-or-not prompt enhanced hard negatives generating for memes category detection, in: Proceedings - IEEE International Conference on Multimedia and Expo, 2023.
- [230] M. Das, S. Banerjee, A. Mukherjee, Hate-alertDravidianLangTech-ACL2022: Ensembling multi-modalities for tamil TrollMeme classification, 2022, arXiv.
- [231] S.U. Hegde, A. Hande, R. Priyadarshini, B. Bharathi, B.R. Chakravarthi, Do images really do the talking? analysing the significance of images in tamil troll meme classification, 2021, arXiv.
- [232] R.N. Nandi, F. Alam, P. Nakov, TeamXDravidianLangTech-ACL2022: A comparative analysis for troll-based meme classification, 2022, arXiv.
- [233] H. Zhang, H. Xu, X. Wang, Q. Zhou, S. Zhao, J. Teng, Mintrec: A new dataset for multimodal intent recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1688–1697.
- [234] N. Mirzoeff, White Sight: Visual Politics and Practices of Whiteness, MIT Press, 2023.
- [235] S. Sekimoto, C. Brown, Race and multimodality: An introduction to the special issue, Multimodality Soc. 3 (3) (2023) 199–209.
- [236] F. Yus, Multimodality in memes: A cyberpragmatic approach, Anal. Digit. Disc.: New Insights Futur. Dir. (2019) 105–131.
- [237] J.L. Lemke, Metamedia literacy: Transforming meanings and media, in: Handbook of literacy and technology: Transformations in a post-typographic world, vol. 283301, 1998.
- [238] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
- [239] F. Wu, B. Gao, X. Pan, L. Li, Y. Ma, S. Liu, Z. Liu, Fuser: An enhanced multimodal fusion framework with congruent reinforced perceptron for hateful memes detection, Inf. Process. Manage. 61 (4) (2024) 103772.
- [240] W. Wu, H. Li, H. Wang, K.Q. Zhu, Probase: A probabilistic taxonomy for text understanding, in: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012, pp. 481–492.
- [241] R. Tommasini, F. Ilievski, T. Wijesiriwardene, IMKG: The internet meme knowledge graph, in: European Semantic Web Conference, Springer, 2023, pp. 354–371.
- [242] L. Bates, P.E. Christensen, P. Nakov, I. Gurevych, A template is all you meme, 2023, arXiv preprint arXiv:2311.06649.
- [243] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (10) (2014) 78–85.
- [244] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: International Semantic Web Conference, Springer, 2007, pp. 722–735.
- [245] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008, pp. 1247–1250.
- [246] M.K. Scheuerman, J.A. Jiang, C. Fiesler, J.R. Brubaker, A framework of severity for harmful content online, Proc. the ACM Human- Comput. Interact. 5 (CSCW2) (2021) 1–33.
- [247] F. Jahanbakhsh, A.X. Zhang, A.J. Berinsky, G. Pennycook, D.G. Rand, D.R. Karger, Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media, Proc. ACM Human- Comput. Interact. 5 (CSCW1) (2021) 1–42.
- [248] T. Kumara, A. Bhattacharjee, J. Garland, Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection, 2024, arXiv preprint arXiv:2403.08035.
- [249] D.M. Beskow, S. Kumar, K.M. Carley, The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning, Inf. Process. Manage. 57 (2) (2020) 102170.