

Detecting and Understanding Harmful Memes: A Survey

Shivam Sharma^{1,2}, Firoj Alam³, Md. Shad Akhtar¹, Dimitar Dimitrov⁴,
Giovanni Da San Martino⁵, Hamed Firooz⁶, Alon Halevy⁶, Fabrizio Silvestri⁷,
Preslav Nakov⁸ and Tanmoy Chakraborty¹

¹IIIT-Delhi, India

²Wipro AI Labs, India

³Qatar Computing Research Institute, HBKU, Qatar

⁴Sofia University, Bulgaria

⁵University of Padova, Italy

⁶Facebook AI, USA

⁷Sapienza University of Rome, Italy

⁸Mohamed bin Zayed University of Artificial Intelligence, UAE

{shivams,shad.akhtar,tanmoy}@iiitd.ac.in, fialam@hbku.edu.qa, mitko.bg.ss@gmail.com,
dasan@math.unipd.it, {mhfiroz,ayh}@fb.com, fsilvestri@diag.uniroma1.it, preslav.nakov@mbzuai.ac.ae

Abstract

The automatic identification of harmful content online is of major concern for social media platforms, policymakers, and society. Researchers have studied textual, visual, and audio content, but typically in isolation. Yet, harmful content often combines multiple modalities, as in the case of memes. With this in mind, here we offer a comprehensive survey with a focus on *harmful memes*. Based on a systematic analysis of recent literature, we first propose a new typology of harmful memes, and then we highlight and summarize the relevant state of the art. One interesting finding is that many types of harmful memes are not really studied, *e.g.*, such featuring self-harm and extremism, partly due to the lack of suitable datasets. We further find that existing datasets mostly capture multi-class scenarios, which are not inclusive of the affective spectrum that memes can represent. Another observation is that memes can propagate globally through repackaging in different languages and that they can also be multilingual, blending different cultures. We conclude by highlighting several challenges related to multimodal semiotics, technological constraints, and non-trivial social engagement, and we present several open-ended aspects such as delineating online harm and empirically examining related frameworks and assistive interventions, which we believe will motivate and drive future research.

1 Introduction

Social media have enabled individuals to freely share content online. While this was a hugely positive development as it enabled free speech, it was also accompanied by the spread of harm and hostility [Brooke, 2019; Joksimovic *et al.*, 2019].

Hate speech [Fortuna and Nunes, 2018], offensive language [Zampieri *et al.*, 2019; Zampieri *et al.*, 2020], abusive language [Mubarak *et al.*, 2017], propaganda [Da San Martino *et al.*, 2019], cyberbullying [Van Hee *et al.*, 2015], cyber-aggression [Kumar *et al.*, 2018], and other harmful content [Pramanick *et al.*, 2021b]¹ have become prominent online. Such content can target individuals, communities, and companies. Social media have defined various categories of harmful content that they do not allow on their platforms [Halevy *et al.*, 2022; Nakov *et al.*, 2021b], and various categorizations have come from the research community [Banko *et al.*, 2020; Pramanick *et al.*, 2021a].

Social media content is often multimodal, combining text, images, and/or videos. In recent years, *Internet memes* have emerged as a popular type of content on social media. A meme can be defined as “a group of digital items sharing common characteristics of content, form, or stance, which were created by associating them and were circulated, imitated, or transformed via the Internet by many users” [Shifman, 2013]. Memes typically consist of one or more images with some text on top [Shifman, 2013; Suryawanshi *et al.*, 2020a]. The motivation and aim of memes is typically humorous, but they can also be harmful.

There has been a lot of work on detecting content that is harmful or otherwise violates the terms of service of online platforms [Alam *et al.*, 2021; Nakov *et al.*, 2021b; Pramanick *et al.*, 2021a; Pramanick *et al.*, 2021b]. This includes detecting hateful users on Twitter [Ribeiro *et al.*, 2018], understanding the virality patterns of memes [Ling *et al.*, 2021], detecting offensive and non-compliant content/logos in product images [Gandhi *et al.*, 2020], spotting hate speech in videos and other modalities [Gomez *et al.*, 2020; Wu and Bhandary, 2020], as well as detecting fine-grained propaganda techniques in memes [Dimitrov *et al.*, 2021a], among others.

¹**Disclaimer:** This paper contains content that may be disturbing to some readers.

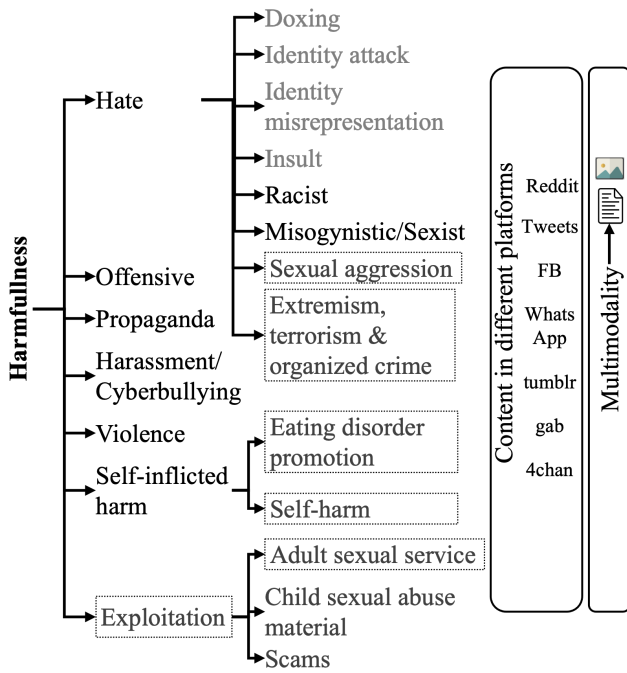


Figure 1: Typology of harmful memes. We show in *grey color* the categories for which we found no memes and no research publications; dotted boxes indicate that this type of memes exist, but we found no publications trying to detect it.

More generally, some of the latest surveys on specific aspects of violating content have been on detecting fake news [Thorne and Vlachos, 2018; Islam *et al.*, 2020; Kotonya and Toni, 2020], disinformation [Alam *et al.*, 2021; Hardalov *et al.*, 2022], misinformation [Nakov *et al.*, 2021a; Nakov *et al.*, 2021c], rumours [Bondielli and Marcelloni, 2019], propaganda [Da San Martino *et al.*, 2020], memes [Afridi *et al.*, 2021], hate speech [Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017], cyberbullying [Haidar *et al.*, 2016], and offensive content [Husain and Uzuner, 2021].

Our survey focuses on detecting and analyzing harmful memes, *i.e.*, *multimodal units consisting of an image and embedded text that has the potential to cause harm to an individual, an organization, a community, or society in general.*

Figure 1 shows our typology of harmful memes, which we defined based on an extensive literature survey; examples of different types of harmful memes are shown in Figure 2. Below, we discuss various aspects of the typology, as well as multimodality, multilinguality, cultural influences, and global propagation through repackaging. We further highlight key issues including the need for fine-grained analysis, the complex abstraction of the memes, and the challenges of the subjectivity of the annotations and of multimodal learning.

2 Harmful Memes

Figure 1 shows our new typology of harmful content on social media, with focus on memes, which is inspired, but differs, from what was proposed in previous work [Banko *et al.*, 2020; Nakov *et al.*, 2021b; Pramanick *et al.*, 2021a].

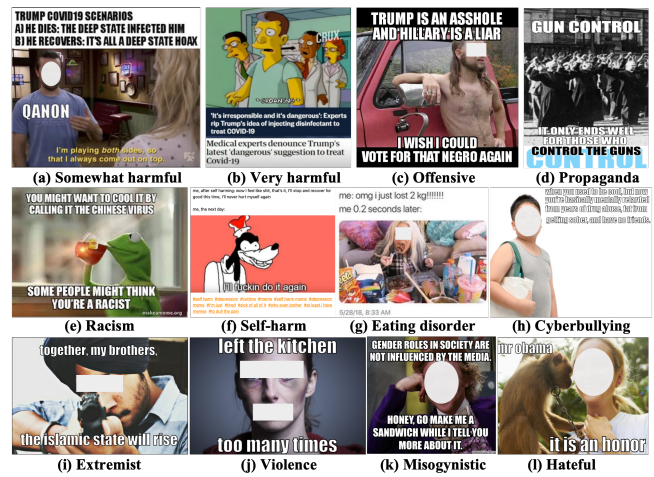


Figure 2: Examples of different types of harmful memes.

For example, [Banko *et al.*, 2020] categorized misinformation as ideological harm, which we excluded from our typology as misinformation is not always harmful. Similarly, while the intent of disinformation is harmful by definition, we do not specifically include it in our typology as most of our sub-categories (*e.g.*, *hate* and *violence*) fall under disinformation [Alam *et al.*, 2021]. Similarly, some of our sub-categories (*e.g.*, *doxing* and *identity attack*) fall under malinformation. Figure 1 highlights the categories with grey-colored text in a dotted box for which we could not find any studies, even though they are prominent in social media: for example, a query in a major search engine using the keywords from Figure 1 will return many memes expressing the respective type of harm [Sabat *et al.*, 2019].

2.1 Types of Harmful Memes

I: Hate

Studies on hate speech detection have focused primarily on textual content [Fortuna and Nunes, 2018], and less on the visual modality [Wu and Bhandary, 2020], with limited research focus on memes [Kiela *et al.*, 2020; Zhou *et al.*, 2021]. An enabling effort in this respect was the *Hateful Memes Challenge* [Kiela *et al.*, 2020], which aimed to identify the targeted protected categories (*e.g.*, *race* and *sex*) and the type of attack (*e.g.*, *contempt* and *slur*) in memes [Zia *et al.*, 2021]. The best system in the competition used different unimodal and multimodal pre-trained models such as VisualBERT [Li *et al.*, 2019], VL-BERT [Su *et al.*, 2020], UNITER [Chen *et al.*, 2020], VILLA [Gan *et al.*, 2020], and ensembles thereof [Kiela *et al.*, 2021]. Using the same dataset, [Zhou *et al.*, 2021] proposed a novel method by incorporating image captioning and data augmentation. The shared task on hateful memes at WOA2021 introduced new labels and tasks, which [Zia *et al.*, 2021] addressed using state-of-the-art pre-trained visual and textual representations along with logistic regression. There have also been efforts to detect the specific *protected categories* being targeted. Below, we elaborate on two such major protected categories: racist and misogynistic/sexist, which are most common in hateful memes on social media.

I.A: Racist: Race is one such protected category that has multi-dimensional aspects in which a systematic out-casting takes place within social, economic, and cultural ecosystems. It is defined as,² “Policies, behaviors, rules, etc. that result in a continued unfair advantage to some people and unfair or harmful treatment of others based on race.” Memetic racism mostly leverages the following:

I.A.a: Physical Appearance: Online racism through memes was found to be prominently based on physical appearance. Research studies used keyword-based scraping of memes from platforms such as Gab, Twitter, 4chan, etc., followed by an in-depth qualitative discussion of the characteristics of online discourse, and supported by thematic analysis. [Williams *et al.*, 2016] investigated the correlation between the offline racial experiences and online perception of racism, where user feedback from white people and people of color was obtained for understanding the differences in the perception of racism. Their findings suggested a higher likelihood of perceiving racism online, primarily by offline victims.

One of the classic scenarios of demeaning people of color and camouflaging systematic racism, also referred to as *color-blindness racism* [Yoon, 2016], against African-Americans is the usage of standard meme templates that primarily target black NBA athletes, whilst juxtapositioning against white men from the NFL, thereby promoting white supremacy [Dickerson, 2016]. This is also exemplified within racism by non-indigenous Australians against Aborigines, which primarily leverages skin tone, stereotypes, and phenotypical characteristics. These memes use either slur/racist words like *abo* and *abbos* or crop facial depictions of aborigines to convey white supremacy and vilification [Al-Natour, 2021].

I.A.b: Ethnicity: Ethnocultural aspects are prominent online, as users from various cultural backgrounds share a common platform for exchanging ideas. [Fairchild, 2020] presented a generic thematic analysis of nine codesets focusing on *race and ethnicity, slurs and language, stereotypes, typology, politics, and culture*, followed by a contextual analysis of the racist discourse and associated tags. [Tuters and Hagen, 2020] presented a qualitative perspective of the prevalence of *triple parenthesis* memes promoting hostility against Jews on 4chan’s /pol/. [Zannettou *et al.*, 2020] empirically analysed (i) the spread of anti-Semitic memes like the *Happy Merchant* meme via semantic embeddings, and (ii) the temporal influence that fringe online users have towards their normalization into mainstream media using the Hawkes process and change-point analysis. They highlighted the use of derogatory slang words, nationalism, conspiracy theories grounded in biblical literature, and hatred towards Jews, encoded using visual-linguistic instruments [Fairchild, 2020]. *Floating signifiers* (e.g., the *Pepe the Frog* meme) along with the adversarial language games [Tuters and Hagen, 2020], lend themselves as versatile and highly accessible platforms for malevolence. As mentioned earlier, social media platforms are instrumental in propagating various types of harmful memes. [Zannettou *et al.*, 2020] studied 4chan’s /pol/ as the major unidirectional spreader of the *Happy Merchant* meme, among many other platforms.

²<https://dictionary.cambridge.org/dictionary/english/racism>

[Fairchild, 2020] highlighted the pivotal role that gamer communities play in facilitating the spread of highly racist content against the generic ones that enable moderately racist content. From computation studies’ viewpoint, [Chandra *et al.*, 2021] emphasized optimal encoding of different modalities using models like ResNET152 [He *et al.*, 2016] and RoBERTa [Liu *et al.*, 2019], along with Multimodal Fusion Architecture Search (MFAS), yielding 0.71 and 0.90 F1 score for Twitter and Gab datasets, respectively, suggesting greater propensity for multimodality in the latter.

I.B: Misogynistic/Sexist: Misogyny and sexism against women have grown a foothold within social media communities, reinvigorating age-old patriarchal establishments of baseless name-calling, objectifying their appearances, and stereotyping gender roles, which has been explored in the literature [Gasparini *et al.*, 2021]. This is especially fueled by the cryptic use of sexism disguised as humor via memes. Qualitative analysis involving the identification of the dominant themes present within sexist memes followed by their detailed interpretation was done via adjectival assessment and with a focus on themes like *technological privilege, others, dominance of patriarchy, gender stereotypes, and women as manipulators* in [Drakett *et al.*, 2018; Siddiqi *et al.*, 2018]. Further analysis by the same authors showed use of derogatory language in these memes, accompanied by the depiction of confident, strong, and poised women, essentially suggesting the threat perceived by sexist and chauvinistic people. When considered for a more extensive set of online memes, such imagery could also be present in non-sexist memes, which highlights the importance of the textual modality. This is further corroborated for sexist meme detection in [Fersini *et al.*, 2019], where textual cues with a late-fusion strategy yields an F1 score of 0.76, thus highlighting the efficacy of distinctly modeling textual cues for such scenarios.

II: Offensive Memes

Offensive content aims to upset or to embarrass people by being rude [Suryawanshi *et al.*, 2020a]. Several studies have focused on content and implicit offensive analogies within memes. Some leveraged unimodal [Giri *et al.*, 2021] and multimodal information [Suryawanshi *et al.*, 2020a], and investigating simple encoder and early fusion strategies for classifying offensive memes, while using techniques such as stacked LSTM/BiLSTM/CNN (Text) along with VGG-16 [Simonyan and Zisserman, 2015] to model multimodality, ultimately achieving an F1 score of 0.71 and accuracy of 0.50. To address contextualization, [Shang *et al.*, 2021b] used analogy-aware multimodality, by combining ResNet50 [He *et al.*, 2016] and GloVe-based LSTM, and attentive multimodal analogy alignment via supervised learning, while incorporating contextual discourse, yielding 0.72 and 0.69 accuracy for Reddit and Gab-based datasets, respectively. [Shang *et al.*, 2021a] extended this study via a graph neural network approach for multimodal entity extraction (KMEE) by leveraging commonsense knowledge for detecting offensive memes, which led to 1% accuracy enhancement in both scenarios. These results demonstrate the importance of contextual and commonsense knowledge for modeling offensive content.

III: Propaganda

Harmful propaganda memes are prominent in online fora that promote *xenophobia*, *racism*, *anti-semitism*, and *anti-feminism/anti-LGBTQ* [Askanius, 2021; Dafaure, 2020]. The memetic language involves similar style, symbolism, and iconography for contrasting inclinations [Greene, 2019] towards recruitment and promoting violent racial supremacy [DeCook, 2018; Askanius and Keller, 2021]. [Mittos *et al.*, 2020] investigated the *genetic testing* discourse, involved in establishing racial superiority and promoting far-right ideologies, by studying correlations using topic modeling, contextual semantics, toxic content analysis, and pHash to characterize the visual cues in memes. Recently, a novel multimodal, multi-label, fine-grained propaganda detection task was proposed [Dimitrov *et al.*, 2021a] as a shared task at SemEval-2021 [Dimitrov *et al.*, 2021b], which focused on detecting fine-grained propaganda techniques in text and in the entire meme, once again confirming the importance of multimodal cues.

IV: Harassment/Cyberbullying

The terms *harassment* and *cyberbullying*, are often used interchangeably. The difference between them is subtle: when the bullying is directed at the target based on protected attributes such as race, skin color, religion, sex, age, disability, nationality, etc., it is considered a harassment. In the past decade, there has been a lot of research effort as well as initiatives by policymakers and social media platforms to address online harassment and cyberbullying as they have been leading to suicides and psychological distress [Rosa *et al.*, 2019]. A recent study has highlighted the increase of harassment over time, most of which happens on social media [Vogels, 2021]. The automatic detection of harassment and cyberbullying has become an important focus of computational social science. [Rosa *et al.*, 2019] systematically reviewed the work on cyberbullying detection and listed the available datasets, the methodologies, and the state-of-the-art performance. They also provided an operational definition exemplifying cyberbullying while delineating annotation guidelines and agreement measures, along with ethical aspects. Besides focusing on the textual modality, [Hosseinmardi *et al.*, 2016] also investigated Instagram images and their associated comments for detecting cyberbullying and online harassment. They manually curated a dataset of 998 examples, including images and their associated comments. Interestingly, they noted that 48% of the posts with loaded language were not labelled as cyberbullying. [Singh *et al.*, 2017] also investigated cyberbullying detection using the same dataset and observed that the image and the text modalities complemented each other. Despite the continued use of multimodal content and memes for cyberbullying, we could not find any major efforts towards its automated detection. However, *name-calling*, which is a prominent tool for cyberbullying, has been explored in the context of propaganda detection [Dimitrov *et al.*, 2021b].

V: Violence

Violence is defined as “*the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either results in or have a high likelihood of resulting in injury, death, psychological harm, maldevelopment or deprivation*” [Krug *et al.*, 2002].

There has been a lot of research in the past decades focusing on multimodal violence detection in surveillance videos [Yao and Hu, 2021], based on video and audio modalities [Acar *et al.*, 2013]. Another line of research investigated the *threat of violence* [Banko *et al.*, 2020] in comments on YouTube videos and Wikipedia [Wulczyn *et al.*, 2017]. In the existing literature, the automatic detection of violent memes has been studied in various contexts, *e.g.*, detecting hateful memes [Kiela *et al.*, 2020]. Yet, we could not find any work specifically focusing on violent meme detection.

VI: Self-Inflicted Harm

Self-inflicted harm includes different forms of harmful behavior, such as self-injury, eating disorders, suicide attempts, etc. [Seko and Lewis, 2018; Banko *et al.*, 2020; Sawhney *et al.*, 2022]. It can be both physical and psychological, and most people self-injure to cope with negative emotions, to punish themselves, or to solicit help from others [Seko and Lewis, 2018]. Self-injuring images are widely spread on Tumblr [Seko and Lewis, 2018], and exposure to them can yield a risk of self-harm and suicide for vulnerable users [Arendt *et al.*, 2019]. While social media platforms are constantly updating and improving their content moderation policies, a significant part of the self-harm content that is posted online remains undetected. At the same time, there are several positive narratives by self-injuries survivors, which require a proactive stance to be promoted [Seko and Lewis, 2018]. The majority of the studies on automated detection of such content are based on textual, visual, and network content analysis, *e.g.*, eating disorder [Wang *et al.*, 2017], and self-harm detection based on textual and visual content, and social context [Losada *et al.*, 2020; Parapar *et al.*, 2021; Wang *et al.*, 2017]. We could not find any literature on automatic detection and analysis of self-inflicted harmful memes.

2.2 Summary

Table 1 summarizes the state of the art on automatic detection of different types of harmful memes, exploring various tasks, datasets, and approaches. In the majority of these studies, the tasks are formulated in a binary setting. While the outcome of a binary setting is useful, multi-class and multi-label settings would be more desirable, *e.g.*, as addressed in [Dimitrov *et al.*, 2021a] for propaganda detection and protected category detection [Zia *et al.*, 2021]. The majority of the studies used large-scale state-of-the-art pre-trained neural networks for the visual content (*e.g.*, VGG and ResNet), for the textual content (*e.g.*, BERT), or for both (*e.g.*, Visual BERT and CLIP). Data augmentation and ensembles were further used in several studies. Table 1 shows variations of F1 such as micro, macro, and weighted; more details can be found at <https://github.com/firojalam/harmful-memes-detection-resources>. Overall, the results are comparatively better for detecting harmful and hateful memes than for the other tasks. For binary classification tasks such as troll identification, the results are only slightly better than random, which highlights the complexity of these tasks.

Types	Publication	Task	Dataset	Cl. T	Approach	AUC	Acc.	F1
Harm	[Pramanick <i>et al.</i> , 2021b]	Y/N	HarMeme	B	VisualBERT	0.81	0.80	
		VH/Ph/NH		M		0.74	0.54	
		Tar. Ident.		M		0.76	0.66	
Harm	[Pramanick <i>et al.</i> , 2021b]	Y/N	Harm-C	B	MOMENTA: CLIP, VGG-19, DistilBERT, CMAF	0.84	0.83	
		VH/Ph/NH		M		0.77	0.55	
		Tar. Ident.	M	0.78		0.70		
		Y/N	Harm-P	B		0.90	0.88	
VH/Ph/NH	M	0.87		0.67				
		Tar. Ident.	M	0.79	0.69			
Hate	[Zia <i>et al.</i> , 2021]	PC	FBHM	ML	CIMG, CTXT LASER, LaBSE	0.96		
		PC. AT.		ML		0.97		
Hate	[Chandra <i>et al.</i> , 2021]	Antisemitism	Gab	B	MFAS	0.91		
		Antisemitism	Twitter	M		0.71		
		Category	Gab			0.67		
		Category	Twitter			0.68		
Hate	[Kirk <i>et al.</i> , 2021]	Hateful	FBHM Pinterest	B	CLIP			0.56 0.57
Hate	[Lee <i>et al.</i> , 2021]	Hateful	FBHM MultiOFF	B	DisMultiHate	0.83	0.76	
Hate	[Gomez <i>et al.</i> , 2020]	Hatespach	MMHS150K	B	FCM, Inception-V3, LSTM	0.73	0.68	0.70
Hate	[Fersini <i>et al.</i> , 2019]	Sexist	The MEME	B	Late fusion			0.76
Hate	[Sabat <i>et al.</i> , 2019]	Hateful	Google	B	BERT, VGG-16, MLP			0.83
Off.	[Shang <i>et al.</i> , 2021b]	Offensive	Gab Reddit	B	Faster R-CNN, ResNet50, Glove-based LSTM, BERT, MLP	0.69	0.56	
						0.72	0.49	
Off.	[Shang <i>et al.</i> , 2021a]	Offensive	Reddit Gab	B	YOLO V4, ConceptNET, GNN	0.73	0.49	
						0.70	0.55	
Off.	[Giri <i>et al.</i> , 2021]	Offensive Off. Int.	Off. Int.	B	CNN, GloVe, LSTM CNN, FastText, LSTM	0.71		
				M		0.99		
Off.	[Suryawanshi <i>et al.</i> , 2020a]	Offensive	MultiOFF	B	Early fusion: Stacked LSTM, BiLSTM/CNN-Text, VGG16			0.50
Prop.	[Dimitrov <i>et al.</i> , 2021a]	Prop. Tech.	Facebook	ML	VisualBERT			0.48
Prop.	[Tian <i>et al.</i> , 2021]	Prop. Tech.: (T)	Facebook	ML	Ensemble: BERT, RoBERTa, XLNet, ALBERT, DistilBERT, DeBERTa, Char n-gram			0.59
Prop.	[Gupta <i>et al.</i> , 2021]	Prop. Tech.: (S)	Facebook	ML	RoBERTa			0.48
Prop.	[Feng <i>et al.</i> , 2021]	Prop. Tech.	Facebook	ML	RoBERTa, Embeddings			0.58
CB	[Hosseinmardi <i>et al.</i> , 2016]	CB Inci.	Instagram	B	SVD +(Unigram, 3-gram), kernelPCA+ meta data, lin. SVM			0.87
CB	[Suryawanshi <i>et al.</i> , 2020b]	Troll	TamilMemes	B	ResNet (Tr: TM) ResNet (Tr: TM + iNet) MobileNet (Tr: TM + iNet + F1k) ResNet (Tr: TM + iNet + F130k)	0.52		
				B		0.52		
				B		0.47		
				B		0.52		

Table 1: Summary of the experimental results for the automatic detection of harmful memes. Y/N: positive and negative class labels; VH: Very harmful, PH: Partially-harmful, NH: Non-harmful; Tar. Ident.: Target Identification; PC: Protected category identification; PC. AT. : Protected category attack type; Off. Int.: Offense intensity prediction; Off: Offensive; Prop.: Propaganda; Prop. Tech.: Propaganda techniques, Prop. Tech.: (T): Text, Prop. Tech.: (S): text span; CB Inci.: Cyberbullying Incidents; CMAF: Cross-modal attention fusion. Cl.T: Classification task; B: Binary, M: Multi-class, ML: Multi-class and Multilabel; TM: TamilMemes, iNet: ImageNet, Fl: Flickr. **More detail can be found at** <http://github.com/firojalam/harmful-memes-detection-resources>

3 Repackaging Memes for Harmful Agendas

Repackaging via remixing or mimicking the meme is a common practice facilitating their adoption across languages and cultures [Shifman, 2013], which often implies harm. For example, popular memes are often repackaged with misogynistic intent. Common ideas that mock specific female identities include *the terrible wife* or *the crazy girlfriend*. For example, the *Distracted Boyfriend* meme³ has been repackaged many times with varying intent, including harm and humor.

Another example is the *Proud Boys* meme,⁴ which has peculiar characteristics. Its proponents work in gangs, indulge in violence and alcohol, follow a uniform code for appearances and collectively accepted logos to depict their identity.

³<https://knowyourmeme.com/memes/distracted-boyfriend>

⁴<https://www.populismstudies.org/wp-content/uploads/2021/03/ECPS-Organisation-Profile-Series-1.pdf>

The use of *Pepe the Frog* reinstates their deeply rooted affiliation to far-right ideologies. Their version of Pepe is a variation that depicts him donning the Proud Boys uniform (black Fred Perry polo with gold trim), whilst displaying the OK hand gesture.

4 Cultural Influence and Multilinguality

[Shifman, 2013] introduced the term *user-generated globalization*, which refers to translation, customization, and distribution of memes across the globe by ordinary online users. In particular, they studied a joke related to computers and romantic relations and its translated version in the top nine non-English languages and found that the joke adapted very well for most of these languages, except for Arabic, which might be due to culture-specific inappropriateness. They further found limited localization of the joke in Chinese, German, and Portuguese.

A recent study [McSwiney *et al.*, 2021] found that most memes either predominantly belonged to anglophone organizations or were derived from anglophone references like the “One Does Not Simply Walk Into Mordor” meme, which appeared in Germany’s Ein Prozent. Most European organizations leverage different genres of images like *share-posts* and *templates* specifically designed for online circulation and orthogonal to the irreverent and participatory nature of memes. In addition to the localized cultural adaptation and customization, memes can use multiple languages. Such examples can be found in the TamilMemes dataset [Suryawanshi *et al.*, 2020b]. Modelling such memes is complex, as is evident from the results reported in [Hegde *et al.*, 2021].

5 Major Challenges

- **Complex abstraction:** Memes can efficiently abstract away complex ideas using creative and powerful customization of visual and linguistic nuances. At the same time, memes with overlapping snippets, patterned text and irony, sarcasm or implicit anti-semitism are non-trivial [Chandra *et al.*, 2021]. For instance, the subtle use of triple parentheses in memes can insinuate a targeted entity whilst underlining an anti-semitic narrative [Tuters and Hagen, 2020]. Moreover, *sexist memes* can promote casual sexism, disguised as humor, irony, sarcasm, and mockery [Siddiqi *et al.*, 2018]. This multi-layering of influential notions via multimodality poses major challenges for automatic meme analysis and requires sophisticated multimodal fusion to understand novel digital vernaculars.
- **Subjectivity in the annotation:** Subjective perceptions play a significant role for memes as a consequence of the complex interplay between the visual and the linguistic content, complemented by the lack of context [Crane and French, 2021]. Moreover, harmful memes, which are prominently used for propaganda warfare violate one’s logic and rational thought. This reverberates as conflicting opinions during data collection and annotation. As noted in [Suryawanshi *et al.*, 2020a], uninitiated annotators were observed to incorrectly mark memes as offensive simply if their sentiments were hurt. This was also concluded from a user study in [Gasparini *et al.*, 2021], wherein out of 59 ambiguous misogynistic memes, only 23% were correctly identified by crowd-sourced workers, while domain experts achieved 77% expert agreement.
- **Inadequate solutions:** Understanding the visual content in memes requires sophisticated solutions, as conventional approaches rely too much on hand-crafted features like low-level grey-scaling, colored, photographic, and semantic features, along with ineffective modelling [Fersini *et al.*, 2019]. This is amplified by the predominantly non-discriminatory nature of visual descriptors in memes, emphasizing textual and discourse-intensive modelling [Shang *et al.*, 2021a; Shang *et al.*, 2021b]. Visual clustering techniques such as pHash used for memes depicting standardized imagery like popular alt-right figures (*e.g.*, *Lauren Southern*, *Richard Spencer*), as well as alt-right memes such as *Pepe the Frog*, and anti-semitic ones such as the *Happy Merchant*, are insufficient to model the visual role-play, indicating the need for sophisticated visual analysis [McSwiney *et al.*, 2021; Zannettou *et al.*, 2020].
- **Insufficient dataset size:** Meme analysis requires a rich set of features and meta-data, which in turn needs a dataset size large enough to enable generalization at scale [AI-Natour, 2021]. Similarly, a keyword-based platform-dependent collection of memes could yield a biased representation of the sample space, and hence could over-represent typical memetic characteristics [Fairchild, 2020].
- **Rapid evolution:** Harmful memes evolve quickly, fueled by new events or by malicious adversaries looking for new ways to bypass existing online detection systems. While humans can generally use prior knowledge to understand new harmful concepts and tasks by looking at a few examples, AI systems struggle to generalize well from a few examples [Wang *et al.*, 2020]. Few-shot learning (FSL) is a new machine learning paradigm that has recently shown breakthrough results in NLP [Brown *et al.*, 2020] and vision tasks [Fan *et al.*, 2021]. It is crucial to advance FSL in the multimodal domain to adapt rapidly and to recognize new evolving types of harmful memes [Tsimpoukelli *et al.*, 2021; Tejankar *et al.*, 2021]. Unlike traditional AI that mainly relies on pattern-matching with labelled data, FSL-based AI systems can evolve to new harmful memes and policies using a handful of examples and can take action immediately instead of waiting for months for the labelled data to be collected.
- **Contextualization:** Understanding many memes requires complex and multimodal reasoning that is based upon a certain contextual background, which may span over diverse levels of abstraction, such as *common sense* [Shang *et al.*, 2021a], *factual* [Zhu, 2020] and *situational* [Sabat *et al.*, 2019]. This contextual information may be conveyed both independently and jointly via textual and visual cues. Analysing this information can be crucial, but it is often not explicitly available for the target meme.
- **Platform restrictions:** The non-standardization of *user accountability and transparency* across constantly evolving social networking services have posed challenges for the systematic study of online harm detection. For example, the freedom of being anonymous has obscured racial integrity and accountability, effectively complicating harmful discourse analysis [Dickerson, 2016]. Moreover, the complex designs and governance policies of platforms such as WhatsApp meant that they focused on their *secure* but unabated use for disseminating systematic racism [Matamoros-Fernández, 2020]. As observed by [Zannettou *et al.*, 2020], the investigation of an actively evolving community like Gab, using a Hawkes process, might err the observations [Zannettou *et al.*, 2020].
- **Identifying real instigators of harm:** Poe’s law emphasizes the understanding of the actual intent while distinguishing between online satire and extremism [Greene, 2019]. Similar ambiguity could also be observed while distinguishing between the real faces of white supremacy and its participatory audience [Greene, 2019]. Interestingly, memes like *triple parenthesis* can render the targets obscure [Tuters and Hagen, 2020]. Even the regulatory bodies find it challenging to clearly distinguish between anti-democracy extremists and anti-democratic alt-right factions [Askanius, 2021]. Consequently, one must also be careful while associating the alt-right with culture. It is instead a historical phenomenon that leverages culture as a tool for its propagation [Dafaure, 2020].

6 Future Forecasting

• **Characterizing vehicles of harm:** Satire is not only used as a progressive tool to resist bigotry, but it is also weaponized by malicious actors towards hijacking the online discussion [Greene, 2019]. It is thus important to decode the discourse and to understand the communication that memes are part of [DeCook, 2018]. Exploring the points of the confluence of youth with far-right memes will help highlight where and how messages of extreme violence circulate and transit back and forth between malicious actors and receptive users [Askanius and Keller, 2021]. It could also be insightful to examine how symptomatic the discourse rhetoric of the anecdotal reference is, within the backdrop of rooted antisemitic perspectives, like the nebulous *Othering*.

• **Cross-cultural studies:** In order to be sensitive to racially hateful memes, systems need to factor in the prejudices and the stereotypes about minorities. One hypothesis is that the relationship between offline micro-aggression and online perception of racism will become more prominent in settings where Caucasians are not the majority. This presents the scope of investigating cross-cultural and cross-contextual implications of racism experienced and perceived online [Williams *et al.*, 2016].

• **Empirical in addition to theoretical:** There are few compelling questions arising from the existing understanding of harmful memes regarding the cause of their potency to instigate harm, cross-platform transitioning, and outcomes, few of them being as follows: To what extent are the “hate jokes” part of the slow yet steady process of normalizing online extremism in mainstream media? What are the consequences of transitioning from their original space to the mainstream? What is the reaction of the general public when exposed to such content [Askanius, 2021]? Clearly, the assessment of the prevalence of different visual forms like memes, photography, and artwork in online communications, along with the cryptic use of visual-linguistic semiotics requires more empirical analysis [McSwiney *et al.*, 2021].

• **Rich metadata:** The use of enriching features such as the tags associated with the social media posts, incorporating video data along with contextual information such as user profiles [Chandra *et al.*, 2021], and using intermediate representations to capture higher levels of abstractions that leverage both the image and the text modalities can help model complex tasks. Moreover, the contextual knowledge supplementing such abstract information becomes indispensable for automated meme analysis [Shang *et al.*, 2021a].

• **Multi-class and multi-label classification:** As highlighted in Table 1, the existing classification setups are primarily binary. However, a more fine-grained multi-class and multi-label setup can enhance the decision-making process, as required in many scenarios. For example, a meme labelled as hateful [Kiela *et al.*, 2020], but which has the characteristics of violence and misogyny, loses its specificity. Attempts in this direction include fine-grained analysis of hateful [Zia *et al.*, 2021] and propagandistic [Dimitrov *et al.*, 2021a] memes, detecting the victims targeted by harmful memes [Pramanick *et al.*, 2021a; Sharma *et al.*, 2022a], and understanding who is the hero, the villain, and the victim [Sharma *et al.*, 2022b].

• **Memetic moderation:** Counter-narratives can help address the selective targeting via harmful memes [Williams, 2020]. The utility of the post-modern transgression and humor must not be left to the alt-right extremists just because they were successful in weaponizing them, as essentially it reinstates their belief that the “left can’t meme” [Dafaure, 2020]. Creating counter memes can help raise awareness about racial issues [Yoon, 2016]. Reclaiming the digital space and indulging in subversive reactions by leveraging the participatory humor using *digilanties* (online vigilantes) can help mitigate the collective menace impended by the systematic and subtle oppression of women [Drakett *et al.*, 2018].

7 Conclusion

We presented a survey of the current intelligent technologies for detecting and understanding harmful memes. Based on a systematic analysis of recent literature, we first proposed a new typology of harmful memes, and then we highlighted and summarized the relevant state of the art. We then discussed the lessons learned and the major challenges that need to be overcome. Finally, we suggested several research directions, which we forecast will emerge in the near future.

Acknowledgments

The work was partially supported by a Wipro research grant, Ramanujan Fellowship, the Infosys Centre for AI, IIIT Delhi, and ihub-Anubhuti-iiitd Foundation, set up under the NM-ICPS scheme of the Department of Science and Technology, India. It is also part of the Tanbih mega-project, which is developed at the Qatar Computing Research Institute, HBKU, and aims to limit the impact of “fake news,” propaganda, and media bias by making users aware of what they are reading.

References

- [Acar *et al.*, 2013] Esra Acar, Frank Hopfgartner, and Sahin Albayrak. Violence detection in Hollywood movies by the fusion of visual and mid-level audio cues. In *MM*, pages 717–720, 2013.
- [Afridi *et al.*, 2021] Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young Koo Lee. A multimodal memes classification: A survey and open research issues. In *SCA*, pages 1451–1466, 2021.
- [Al-Natour, 2021] Ryan Al-Natour. The digital racist fellowship behind the anti-aboriginal internet memes. *J. of Soc.*, 57(4):780–805, 2021.
- [Alam *et al.*, 2021] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. A survey on multimodal disinformation detection. *arXiv:2103.12541*, 2021.
- [Arendt *et al.*, 2019] Florian Arendt, Sebastian Scherr, and Daniel Romer. Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults. *New Media & Soc.*, 21(11-12):2422–2442, 2019.
- [Askanius and Keller, 2021] Tina Askanius and Nadine Keller. Murder fantasies in memes: fascist aesthetics of death threats and the banalization of white supremacist violence. *Info., Comm. & Soc.*, 24(16):2522–2539, 2021.

- [Askanius, 2021] Tina Askanius. On frogs, monkeys, and execution memes: Exploring the humor-hate nexus at the intersection of neo-Nazi and alt-right movements in Sweden. *Tel. & New Media*, 22(2):147–165, 2021.
- [Banko *et al.*, 2020] Michele Banko, Brendon MacKeen, and Laurie Ray. A unified taxonomy of harmful content. In *WOAH*, pages 125–137, 2020.
- [Bondielli and Marcelloni, 2019] Alessandro Bondielli and Francesco Marcelloni. A survey on fake news and rumour detection techniques. *Info. Sci.*, 497:38–55, 2019.
- [Brooke, 2019] Sian Brooke. “Condescending, Rude, Assholes”: Framing gender and hostility on Stack Overflow. In *WALO*, pages 172–180, 2019.
- [Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.
- [Chandra *et al.*, 2021] Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. “Subverting the Jewtocracy”: online antisemitism detection using multimodal deep learning. In *WebSci*, page 148–157, 2021.
- [Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, pages 104–120, 2020.
- [Crane and French, 2021] Tim Crane and Craig French. The problem of perception. In *The Stanford Encyclopedia of Philosophy*. Stanford University, 2021.
- [Da San Martino *et al.*, 2019] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news article. In *EMNLP-IJCNLP*, pages 5636–5646, 2019.
- [Da San Martino *et al.*, 2020] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. In *IJCAI*, pages 4826–4832, 2020.
- [Dafaure, 2020] Maxime Dafaure. The “great meme war”: The alt-right and its multifarious enemies. *Angles*, (10), 2020.
- [DeCook, 2018] Julia R. DeCook. Memes and symbolic violence: #proudbots and the use of memes for propaganda and the construction of collective identity. *LMT*, 43(4):485–504, 2018.
- [Dickerson, 2016] Nikolas Dickerson. Constructing the digitalized sporting body: Black and white masculinity in NBA/NHL internet memes. *Comm. & Sport*, 4(3):303–330, 2016.
- [Dimitrov *et al.*, 2021a] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. Detecting propaganda techniques in memes. In *ACL-IJCNLP*, pages 6603–6617, 2021.
- [Dimitrov *et al.*, 2021b] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *SemEval*, pages 70–98, 2021.
- [Drakett *et al.*, 2018] Jessica Drakett, Bridgette Rickett, Katy Day, and Kate Milnes. Old jokes, new media – online sexism and constructions of gender in internet memes. *Fem. & Psy.*, 28(1):109–127, 2018.
- [Fairchild, 2020] Tabitha Fairchild. It’s funny because it’s true: The transmission of explicit and implicit racism in internet memes. Virginia Commonwealth University, 2020.
- [Fan *et al.*, 2021] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In *CVPR*, pages 4527–4536, 2021.
- [Feng *et al.*, 2021] Zhida Feng, Jiji Tang, Jiayang Liu, Weichong Yin, Shikun Feng, Yu Sun, and Li Chen. Alpha at SemEval-2021 task 6: Transformer based propaganda classification. In *SemEval*, pages 99–104, 2021.
- [Fersini *et al.*, 2019] Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. Detecting sexist meme on the web: A study on textual and visual cues. In *ACIHW*, pages 226–231, 2019.
- [Fortuna and Nunes, 2018] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *CSUR*, 51(4):1–30, 2018.
- [Gan *et al.*, 2020] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *NeurIPS*, 33:6616–6628, 2020.
- [Gandhi *et al.*, 2020] Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. Scalable detection of offensive and non-compliant content / logo in product images. *WACV*, pages 2236–2245, 2020.
- [Gasparini *et al.*, 2021] Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *arXiv:2106.08409*, 2021.
- [Giri *et al.*, 2021] Roushan Kumar Giri, Subhash Chandra Gupta, and Umesh Kumar Gupta. An approach to detect offence in memes using natural language processing (NLP) and deep learning. In *ICCCI*, pages 1–5, 2021.
- [Gomez *et al.*, 2020] Raul Gomez, Jaume Gibert, Lluís Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multi-modal publications. In *WACV*, pages 1470–1478, 2020.
- [Greene, 2019] Viveca S. Greene. “Deplorable” satire: Alt-right memes, white genocide tweets, and redpilling normies. *Stud. in Am. Humor*, 5(1):31–69, 2019.
- [Gupta *et al.*, 2021] Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. Volta at SemEval-2021 task 6: Towards detecting persuasive texts and images using textual and multimodal ensemble. In *SemEval*, pages 1075–1081, 2021.
- [Haidar *et al.*, 2016] Batoul Haidar, Maroun Chamoun, and Fadi Yamout. Cyberbullying detection: A survey on multilingual techniques. In *EMS*, pages 165–171, 2016.
- [Halevy *et al.*, 2022] Alon Halevy, Cristian Canton-Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. Preserving integrity in online social networks. *Commun. ACM*, 65(2):92–98, jan 2022.
- [Hardalov *et al.*, 2022] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. A survey on stance detection for mis- and disinformation identification. In *NAACL (Findings)*, 2022.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

- [Hegde *et al.*, 2021] Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Sathiyaraj Thangasamy, B Bharathi, and Bharathi Raja Chakravarthi. Do images really do the talking? Analysing the significance of images in Tamil troll meme classification. *arXiv:2108.03886*, 2021.
- [Hosseinmardi *et al.*, 2016] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Prediction of cyberbullying incidents in a media-based social network. In *ASONAM*, pages 186–192, 2016.
- [Husain and Uzuner, 2021] Fatemah Husain and Ozlem Uzuner. A survey of offensive language detection for the Arabic language. *TALLIP*, 20(1):1–44, 2021.
- [Islam *et al.*, 2020] Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *SNAM*, 10(1):1–20, 2020.
- [Joksimovic *et al.*, 2019] Srecko Joksimovic, Ryan S. Baker, Jaelyn Ocumpaugh, Juan Miguel L. Andres, Ivan Tot, Elle Yuan Wang, and Shane Dawson. Automated identification of verbally abusive behaviors in online discussions. In *WALO*, pages 36–45, 2019.
- [Kiela *et al.*, 2020] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, volume 33, pages 2611–2624, 2020.
- [Kiela *et al.*, 2021] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, et al. The hateful memes challenge: competition report. In *NeurIPS*, pages 344–360, 2021.
- [Kirk *et al.*, 2021] Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. In *WOAH*, pages 26–35, 2021.
- [Kotonya and Toni, 2020] Neema Kotonya and Francesca Toni. Explainable automated fact-checking: A survey. In *COLING*, pages 5430–5443, 2020.
- [Krug *et al.*, 2002] Etienne G Krug, James A Mercy, Linda L Dahlberg, and Anthony B Zwi. The world report on violence and health. *The Lancet*, 360(9339):1083–1088, 2002.
- [Kumar *et al.*, 2018] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking aggression identification in social media. In *TRAC*, pages 1–11, 2018.
- [Lee *et al.*, 2021] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. Disentangling hate in online memes. In *ICMEW*, pages 5138–5147, 2021.
- [Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv:1908.03557*, 2019.
- [Ling *et al.*, 2021] Chen Ling, Ihab AbuHilal, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Dissecting the meme magic: Understanding indicators of virality in image memes. *CSCW*, 5:1–24, 2021.
- [Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019.
- [Losada *et al.*, 2020] David E Losada, Fabio Crestani, and Javier Parapar. Overview of eRisk 2020: early risk prediction on the internet (extended overview). In *CLEF*, 2020.
- [Matamoros-Fernández, 2020] Ariadna Matamoros-Fernández. ‘El Negro de WhatsApp’ meme, digital blackface, and racism on social media. *First Monday*, 25(12), 2020.
- [McSwiney *et al.*, 2021] Jordan McSwiney, Michael Vaughan, Annett Heft, and Matthias Hoffmann. Sharing the hate? Memes and transnationality in the far right’s digital visual culture. *Info., Comm. & Soc.*, 24(16):2502–2521, 2021.
- [Mittos *et al.*, 2020] Alexandros Mittos, Savvas Zannettou, Jeremy Blackburn, and Emiliano De Cristofaro. “And we will fight for our race!” A measurement study of genetic testing conversations on Reddit and 4chan. In *ICWSM*, pages 452–463, 2020.
- [Mubarak *et al.*, 2017] Hamdy Mubarak, Kareem Darwish, and Walid Magdy. Abusive language detection on Arabic social media. In *WALO*, pages 52–56, 2017.
- [Nakov *et al.*, 2021a] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. Automated fact-checking for assisting human fact-checkers. In *IJCAI*, pages 4551–4558, 2021.
- [Nakov *et al.*, 2021b] Preslav Nakov, Vibha Nayak, Kyle Dent, Ameya Bhatawdekar, Sheikh Muhammad Sarwar, Momchil Hardalov, Yoan Dinkov, Dimitrina Zlatkova, Guillaume Bouchard, and Isabelle Augenstein. Detecting abusive language on online platforms: A critical analysis. *arXiv:2103.00153*, 2021.
- [Nakov *et al.*, 2021c] Preslav Nakov, Husrev Taha Sencar, Jisun An, and Haewoon Kwak. A survey on predicting the factuality and the bias of news media. *arXiv:2103.12506*, 2021.
- [Parapar *et al.*, 2021] Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. Overview of eRisk 2021: Early risk prediction on the internet. In *CLEF*, pages 324–344, 2021.
- [Pramanick *et al.*, 2021a] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Detecting harmful memes and their targets. In *ACL-IJCNLP (Findings)*, pages 2783–2796, 2021.
- [Pramanick *et al.*, 2021b] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *EMNLP (Findings)*, pages 4439–4455, 2021.
- [Ribeiro *et al.*, 2018] Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on Twitter. In *ICWSM*, pages 676–679, 2018.
- [Rosa *et al.*, 2019] Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345, 2019.
- [Sabat *et al.*, 2019] Benet Oriol Sabat, Cristian Canton-Ferrer, and Xavier Giró-i-Nieto. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv:1910.02334*, 2019.
- [Sawhney *et al.*, 2022] Ramit Sawhney, Shivam Agarwal, Atula Tejaswi Neerkaje, Nikolaos Aletras, Preslav Nakov, and Lucie Flek. Towards suicide ideation detection through online conversational context. In *SIGIR*, 2022.

- [Schmidt and Wiegand, 2017] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *SocialNLP*, pages 1–10, 2017.
- [Seko and Lewis, 2018] Yukari Seko and Stephen P Lewis. The self-harmed, visualized, and reblogged: Remaking of self-injury narratives on tumblr. *New media & society*, 20(1):180–198, 2018.
- [Shang *et al.*, 2021a] Lanyu Shang, Christina Youn, Yuheng Zha, Yang Zhang, and Dong Wang. KnowMeme: A knowledge-enriched graph neural network solution to offensive meme detection. In *eScience*, pages 186–195, 2021.
- [Shang *et al.*, 2021b] Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. AOMD: An analogy-aware approach to offensive meme detection on social media. *Inf. Process. Manage.*, 58(5), 2021.
- [Sharma *et al.*, 2022a] Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. DISARM: Detecting the victims targeted by harmful memes. In *NAACL-HLT (Findings)*, 2022.
- [Sharma *et al.*, 2022b] Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes. In *CONSTRAINT*, pages 1–11, 2022.
- [Shifman, 2013] Limor Shifman. *Memes in digital culture*. MIT press, 2013.
- [Siddiqi *et al.*, 2018] Nasrina Siddiqi, Anjuman Bains, Sheema Aleem, and Sheema Aleem. Analysing threads of sexism in new age humour: A content analysis of internet memes. *Indian J. Soc. Res.*, 59:356, 2018.
- [Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [Singh *et al.*, 2017] Vivek K Singh, Souvick Ghosh, and Christin Jose. Toward multimodal cyberbullying detection. In *CHI*, pages 2090–2099, 2017.
- [Su *et al.*, 2020] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020.
- [Suryawanshi *et al.*, 2020a] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *TRAC*, pages 32–41, 2020.
- [Suryawanshi *et al.*, 2020b] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. A dataset for troll classification of TamilMemes. In *WILDRE*, pages 7–13, 2020.
- [Tejankar *et al.*, 2021] Ajinkya Tejankar, Bichen Wu, Saining Xie, Madian Khabza, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv:2112.13884*, 2021.
- [Thorne and Vlachos, 2018] James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. In *COLING*, pages 3346–3359, 2018.
- [Tian *et al.*, 2021] Junfeng Tian, Min Gui, Chenliang Li, Ming Yan, and Wenming Xiao. MinD at SemEval-2021 Task 6: propaganda detection using transfer learning and multimodal fusion. In *SemEval-2021*, pages 1082–1087, 2021.
- [Tsimpoukelli *et al.*, 2021] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *NeurIPS*, 34:200–212, 2021.
- [Tuters and Hagen, 2020] Marc Tuters and Sal Hagen. (((They))) rule: memetic antagonism and nebulous othering on 4chan. *New Media & Society*, 22(12):2218–2237, 2020.
- [Van Hee *et al.*, 2015] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. Detection and fine-grained classification of cyberbullying events. In *RANLP*, pages 672–680, 2015.
- [Vogels, 2021] Emily A Vogels. The state of online harassment. *Pew Research Center*, 13, 2021.
- [Wang *et al.*, 2017] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. Detecting and characterizing eating-disorder communities on social media. In *WSDM*, page 91–100, 2017.
- [Wang *et al.*, 2020] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *CSUR*, 53(3):1–34, 2020.
- [Wang *et al.*, 2017] Yilin Wang *et al.* Understanding and discovering deliberate self-harm content in social media. In *WWW*, page 93–102, 2017.
- [Williams *et al.*, 2016] Amanda Williams, Clio Oliver, Katherine Aumer, and Chanel Meyers. Racial microaggressions and perceptions of internet memes. *CHB*, 63:424–432, 2016.
- [Williams, 2020] Apryl Williams. Black memes matter: #living-whileblack with Becky and Karen. *Social Media + Society*, 6(4):2056305120981047, 2020.
- [Wu and Bhandary, 2020] Ching Seh Wu and Unnathi Bhandary. Detection of hate speech in videos using machine learning. In *CSCI*, pages 585–590, 2020.
- [Wulczyn *et al.*, 2017] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *WWW*, pages 1391–1399, 2017.
- [Yao and Hu, 2021] Huiling Yao and Xing Hu. A survey of video violence detection. *Cyber-Physical Systems*, pages 1–24, 2021.
- [Yoon, 2016] InJeong Yoon. Why is it not just a joke? Analysis of internet memes associated with racism and hidden ideology of colorblindness. *JCRAE*, 33, 2016.
- [Zampieri *et al.*, 2019] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In *NAACL-HLT*, pages 1415–1420, 2019.
- [Zampieri *et al.*, 2020] Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *SemEval*, pages 1425–1447, 2020.
- [Zannettou *et al.*, 2020] Savvas Zannettou, Joel Finkelstein, Barry Bradlyn, and Jeremy Blackburn. A quantitative approach to understanding online antisemitism. In *ICWSM*, pages 786–797, 2020.
- [Zhou *et al.*, 2021] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. Multimodal learning for hateful memes detection. In *ICMEW*, pages 1–6, 2021.
- [Zhu, 2020] Ron Zhu. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv:2012.08290*, 2020.
- [Zia *et al.*, 2021] Haris Bin Zia, Ignacio Castro, and Gareth Tyson. Racist or sexist meme? Classifying memes beyond hateful. In *WOAH*, pages 215–219, 2021.