RESEARCH ARTICLE OPEN ACCESS

COMPARATIVE APPROACH FOR THE DISCOVERY OF VIOLENCE IN DIGITAL MEDIA POSTS

KONDA SINDHUJA,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)
POREDDY NITHIN REDDY,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)
NAGARIGARI NITHIN GOUD,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)
BIRDHARAJU SAI DHEERAJ RAJU,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)
Dr.T.CHARAN SINGH,Assoc.Prof.CSIT,Sri Indu College Of Engineering & Technology(A)

Abstract

The automatic identification of harmful content is of major concern for so cialm ediaplatforms, policymakers, and society. Researchers have stud- ied 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 sur- vey with a focus on *harmful memes*.Based on a systematic canalysis of recent literature, pose 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,partlyduetothe lack of suitable datasets.We further find that ex- istingdatasetsmostlycapturemulti-classscenarios, which are not inclusive of the effective 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 sev- eral open-ended aspects such as delineating online harmonizing and critically examining related frameworks and assistive interventions, which we believe will motivate and drive future research.

1 Introduction

Socialmediahaveenabledindividualstofreelysharecontent online.Whilethiswasahugelypositivedevelopmentasit Hate speech[FortunaandNunes,2018],offensivelanguage [Zampieri *et al.*, 2019;Zampieri *et al.*, 2020], abusive lan- guage [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 businesses. Social media has defined various categories of harmful

content that they do not allow on their platforms [Halevy *etal.*, 2022;Nakov *et al.*, 2021b], and various categorizations havecomefromtheresearchcommunity[Banko*etal.*,2020;Pram anick *et al.*, 2021a].

Social media content is often multimodal, combining text, images,and/orvideos.Inrecentyears,*Internetmemes* have emerged as a popular type of content on social me-dia.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 circu- lated, imitated, or transformed via the Internet by many users"[Shifman,2013].Memestypicallyconsistofone or more images with some text on top [Shifman, 2013;Suryawanshi*etal.*,2020a].Themotivationandaimofmem es 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*etal.*,2021;Nakov*etal.*,2021b;Pramanick*etal.*, 2021a;Pramanick*etal.*,2021b].Thisincludesdetecting hatefulusersonTwitter[Ribeiro*etal.*,2018],understanding the virality patterns of memes [Ling *et al.*, 2021], detecting offensiveandnon-compliantcontent/logosinproductimages [Gandhi*etal.*,2020],spottinghatespeechinvideosandother 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.

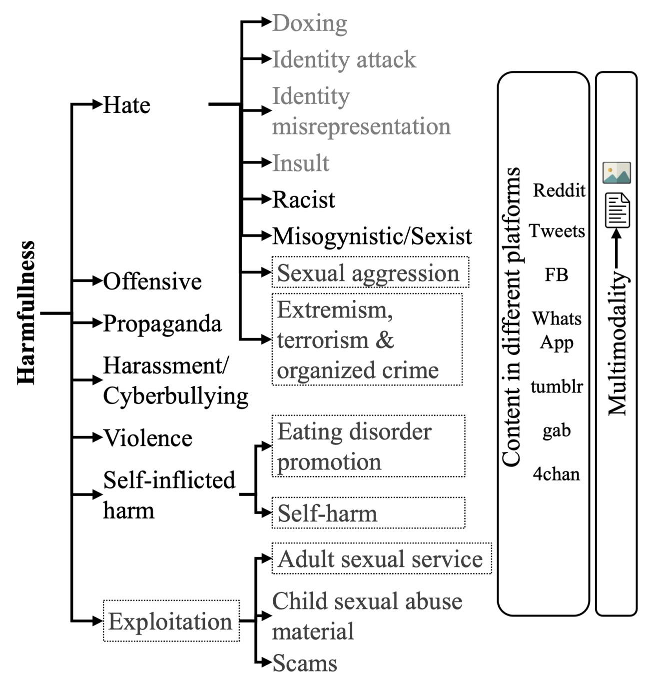

Figure 1: Typology of harmful memes. We show in *grey color* the categories for which we found no memes and no research publications; dotted boxes indicate that this type of memes exist, but we found no publications trying to detect it.

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;KotonyaandToni,2020],disinformation[Alam*etal.*,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 embeddedtextthathasthepotentialtocauseharmtoanindividual, an organization, a community, or society in general.

Figure1showsourtypologyofharmfulmemes,whichwe defined based on an extensive literature survey; examples of differenttypesofharmfulmemesareshowninFigure2. Below, we discuss various aspects of the typology, as well as multimodality,multilinguality,culturalinfluences,andglobal propagationthroughrepackaging. Wefurtherhighlightkeyissuesincludingtheneedforfine-grainedanalysis,thecomplex abstractionofthememes,andthechallengesofthesubjectivity of the annotations and of multimodal learning.

2 HarmfulMemes

Figure1showsournewtypologyofharmfulcontentonsocial media, with focus on memes, which is inspired, but differs, fromwhatwasproposedinpreviouswork[Banko*etal.*,2020;Nakov *et al.*, 2021b;Pramanick *et al.*, 2021a].

Figure2:Examplesofdifferenttypesofharmfulmemes.

For example, [Banko *et al.*, 2020] categorized misinformation as ideological harm, which we excluded from our typologyasmisinformationisnotalwaysharmful. Similarly, whiletheintentofdisinformationisharmfulbydefinition,we do not specifically include it in our typology as most of our subcategories(*e.g.*,*hate*and*violence*)fallunderdisinformation[Alam*etal.*,2021].Similarly,someofoursub-categories (*e.g.*, *doxing* and *identity attack*) fall under malinformation. Figure1highlightsthecategorieswithgrey-coloredtextin a dotted box for which we could not find any studies, even though they are prominent in social media:for example, a queryinamajorsearchengineusingthekeywordsfromFigure 1 will return many memes expressing the respective type of harm [Sabat *et al.*, 2019].

TypesofHarmfulMemes

I:Hate

Studies on hate speech detection have focused primarily on textualcontent[FortunaandNunes,2018],andlessonthevisualmodality[WuandBhandary,2020],withlimitedresearch focus on memes [Kiela *et al.*, 2020; Zhou *et al.*, 2021].An enabling effort in this respect was the *Hateful Memes Challenge*[Kiela*etal.*,2020],whichaimedtoidentifythetargeted protectedcategories(*e.g.*,*race*and*sex*)andthetypeofattack (*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 *etal.*,2019],VL-BERT[Su*etal.*,2020],UNITER[Chen*et al.*,2020],VILLA[Gan*et al.*,2020], andensemblesthereof [Kiela*etal.*,2021].Usingthesamedataset,[Zhou*etal.*,2021] proposed a novel method by incorporating image captioning and data augmentation.The shared task on hateful memesat WOAH 2021 introduced new labels and tasks, which [Zia*et al.*, 2021] addressed using state-of-the-art pre-trained visualandtextualrepresentationsalongwithlogisticregression. Therehavealsobeenefforts todetectthespecific*protectedcategories*beingtargeted.Below,weelaborateontwosuchmajor protectedcategories:racistandmisogynistic/sexist,whichare most common in hateful memes on social media.

I. A:Racist: Raceisonesuchprotectedcategorythathas multi-dimensional aspects in which a systematic out-casting takesplacewithinsocial,economic,andculturalecosystems. Itisdefinedas,[2] "*Policies,behaviors,rules,etc. thatresultin a continued unfair advantage to some people and unfair or harmfultreatmentofothersbasedonrace*." Memeticracism mostly leverages the following:

a: Physical Appearance: Online racism through memes was found to be prominently based on physical appearance. Research studies used keyword-based scraping of memesfromplatformssuchasGab,Twitter,4chan,etc.,followedbyanin-depthqualitativediscussionofthecharacteristics of online discourse, and supported by thematic analysis. [Williams *et al.*, 2016] investigated the correlation between theofflineracialexperiencesandonlineperceptionofracism, where user feedback from white people and people of color wasobtainedforunderstandingthedifferencesintheperceptionofracism.Theirfindingssuggestedahigherlikelihoodof perceiving racism online, primarily by offline victims.

One of the classic scenarios of demeaning people of color andcamouflagingsystematicracism,alsoreferredtoas*color-blindnessracism*[Yoon,2016],againstAfrican-Americansis the usage of standard meme templates that primarily target blackNBAathletes,whilstjuxtapositioningagainstwhitemen from the NFL, thereby promoting white supremacy [Dickerson, 2016].This is also exemplified within racism by non-indigenous Australians against Aboriginals, which primarily leverages skin tone, stereotypes, and phenotypical characteristics. These memes use either slur/racist words like *abo* and *abbos*orcropfacialdepictionsofaboriginalstoconveywhite supremacy and vilification [Al-Natour, 2021].

b: Ethnicity: Ethnoculturalaspectsareprominent online, as users from various cultural backgrounds share a commonplatformforexchangingideas. [Fairchild,2020]presented a generic thematic analysis of nine codesets focusing on*raceandethnicity,slursandlanguage,stereotypes,typology,politics,andculture*,followedbyacontextualanalysisof the racist discourse and associated tags.[Tuters and Hagen,2020]presentedaqualitativeperspectiveoftheprevalenceof *tripleparenthesis*memespromotinghostilityagainstJewson 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 normalizationintomainstreammediausingtheHawkesprocessand change-point analysis. They highlighted the use of derogateTory slang words, nationalism, conspiracy theories grounded inbiblicalliterature,andhatredtowardsJews,encodedusing visual-linguistic instruments [Fairchild, 2020]. *Floating signifiers*(*e.g.*,the*PepetheFrog*meme)alongwiththeadversarial language games [Tuters and Hagen, 2020], lend themselves asversatileand highlyaccessibleplatforms formalevolence. Asmentionedearlier,socialmediaplatformsareinstrumental in propagating various types of harmful memes. [Zannettou*etal.*,2020]studied4chan's/pol/asthemajorunidirectional spreader of the *Happy Merchant* meme, among many other platforms.

communities play in facilitating the spread of highly racist contentagainstthegenericonesthatenablemoderatelyracist content.From computation studies' viewpoint, [Chandra *etal.*, 2021] emphasized optimal encoding of different modal- ities using models like ResNET152 [He *et al.*, 2016] and RoBERTa [Liu *et al.*, 2019], along with Multimodal Fusion ArchitectureSearch(MFAS),yielding0.71and0.90F1score forTwitterandGabdatasets,respectively,suggestinggreater propensity for multimodality in the latter.

I.B:Misogynistic/Sexist: Misogynyandsexismagainst 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*etal.*,2021].Thisisespeciallyfu- eled 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

showeduseofderogatorylanguageinthesememes,accompa - niedbythedepictionofconfident,strong,andpoisedwomen, essentiallysuggestingthethreatperceivedbysexistandchauvinistic people.when considered for a more extensive setof online memes, such imagery could also be present in non-sexist memes, which highlightstheimportance ofthe textual modality.This is further corroborated for sexist meme detection in [Fersini *et al.*, 2019], where textual cues with a late-fusionstrategyyieldsanF1scoreof0.76,thushighlight- ing the efficacy of distinctly modeling textual cues for such scenarios.

II:OffensiveMemes

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 in- vestigating simple encoder and early fusion strategies for classifying offensive memes, while using techniques such as stackedLSTM/BiLSTM/CNN(Text)alongwithVGG-16[Si-monyan and Zisserman, 2015] to model multimodality, ultimatelyachievinganF1scoreof0.71andaccuracyof0.50.To addresscontextualization,[Shang*etal.*,2021b]usedanalogy-awaremultimodality,bycombiningResNet50[He*etal.*,2016] and GloVe-based LSTM, and attentive multimodal analogy alignment via supervised learning, while incorporating contextualdiscourse,yielding0.72and0.69accuracyforReddit- and Gab-based datasets, respectively.[Shang *et al.*, 2021a] extended this study via a graph neural network approach for multimodalentityextraction(KMEE)-

onmultimodalviolencedetectioninsurveillancevideos[YaoandHu,2021],[Acar*etal.*,2013]. Anotherlineofresearchinvestigatedthe*threatof violence*[Banko*etal.*,2020]incommentsonYouTubevideos andWikipedia[Wulczyn*etal.*, 2017].Intheexistingliterature, theautomaticdetectionofviolentmemeshasbeenstudiedin various contexts, *e.g.*, detecing hateful memes [Kiela *et al.*,2020]. Yet,wecouldnotfindanyworkspecificallyfocusing on violent meme detection.

VI:Self-InflictedHarm

Self-inflicted harm includes different forms of harmful be- havior, such as self-injury, eating disorders, suicide at- tempts, etc.[Seko and Lewis, 2018;Banko *et al.*, 2020;Sawhney *et al.*, 2022].It can be both physical and psycho- logical, and most people self-injure to cope with negative emotions, to punish themselves, or to solicit help from oth- ers [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 platformsare constantly updating and improving their content modera- tion policies, a significant part of the self-harm content thatis posted online remains undetected. At the same time, there areseveralpositivenarrativesbyself-injuriessurvivors,which require a proactive stance to be promoted [Seko and Lewis,2018].The majority of the studies on automated detectionof 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;Wangetal.,2017]. Wecouldnotfindanyliteratureonauto-maticdetectionandanalysisofself-inflictedharmfulmemes.

Summary

Table 1 summarizes the state of the art on automatic detec- tion of different types of harmful memes, exploring various tasks,datasets,andapproaches. Inthemajorityofthesestud- ies, 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 pro- tected category detection [Zia *et al.*, 2021]. The majority of the studies used large-scale state-of-the-art pre-trained neu- ral networks for the visual content (*e.g.*, VGG and ResNet), for the textual content (*e.g.*, BERT), or for both (*e.g.*, Vi- sual BERT and CLIP). Data augmentation and ensembles were further used in several studies.Table 1 shows varia- tionsofF1suchasmicro,macro,andweighted;moredetails canbefoundathttps://github.com/firojalam/harmful-memes-detection-resources.Overall, the results are comparatively better for detecting harmful and hateful memes than for the othertasks.Forbinaryclassificationtaskssuchastrollidenti-fication,theresultsareonlyslightlybetterthanrandom,which highlights the complexity of these tasks.

III:Propaganda

Harmful propaganda memes are prominent in online forathat 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] to- wards recruitment and promoting violent racial supremacy [DeCook, 2018;Askanius and Keller, 2021].[Mittos *et al.*,2020] investigated the *genetic testing* discourse, involved in establishingracialsuperiorityandpromotingfar-rightideolo-gies,bystudyingcorrelationsusingtopicmodeling,contextual semantics, toxic content analysis, and pHash to characterize thevisualcuesinmemes.Recently,anovelmultimodal,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-grainedpropagandatechniquesintextandintheentirememe, once again confirming the importance of multimodal cues.

IV: Harassment/Cyberbullying

Theterms*harassment*and*cyberbullying*,areoftenusedinter-changeably. Thedifferencebetweenthemissubtle: whenthe bullyingisdirectedatthetargetbasedonprotectedattributes such as race, skin color, religion, sex, age, disability, nation- ality, etc., it is considered a harassment.In the past decade, therehasbeenalotofresearcheffortaswellasinitiativesby 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,mostofwhichhappensonsocialmedia[Vogels,2021]. Theautomaticdetectionofharassmentandcyberbullyinghas become an important focus of computational social science. [Rosa *et al.*, 2019] systematically reviewed the work on cy- berbullying detection and listed the available datasets, the methodologies, and the state-of-the-art performance.They alsoprovidedanoperationaldefinitionexemplifyingcyberbul-lying while delineating annotation guidelines and agreement measures,alongwithethicalaspects. Besidesfocusingonthe textual modality, [Hosseinmardi *et al.*, 2016] also investigated Instagramimagesandtheirassociatedcommentsfordetecting cyberbullyingandonlineharassment. Theymanuallycurated a dataset of 998 examples, including images and their asso- ciated comments.Interestingly, they noted that 48% of the postswithloadedlanguagewerenotlabelledascyberbullying. [Singh*etal.*,2017]alsoinvestigatedcyberbullyingdetection usingthesamedatasetandobservedthattheimageandthetext modalities complemented each other. Despite the continued use of multimodal content and memes for cyberbullying, we couldnotfindanymajoreffortstowardsitsautomateddetec-tion.However, *name-calling*, which is a prominent tool for cyberbullying,hasbeenexploredinthecontextofpropaganda detection [Dimitrov *et al.*, 2021b].

V:Violence

Violenceisdefinedas*"theintentionaluseofphysicalforceor power,threatenedoractual,againstoneself,anotherperson, oragainstagrouporcommunity,thateitherresultsinorhave a*

Types	Publication	Task	Dataset	Cl.T	Approach	AUCAcc.F1
Harm[	Pramanick*etal.*,2021b]	Y/N	HarMeme	B	VisualBERT	0.810.80
		VH/Ph/NH		M		0.740.54
		Tar.Ident.		M		0.760.66
Harm[	Pramanick*etal.*,2021b]	Y/N	Harm-C	B	MOMENTA:	0.840.83
		VH/Ph/NH		M	CLIP,VGG-19,	0.770.55
		Tar.Ident.		M	DistilBERT,	0.780.70
		Y/N	Harm-P	B	CMAF	0.900.88
		VH/Ph/NH		M		0.870.67
		Tar.Ident.		M		0.790.69
Hate[Zia*etal.*, 2021]		PC	FBHM	ML	CIMG,CTXT	0.96
		PC.AT.		ML	LASER,LaBSE	0.97
Hate[Chandra*etal.*,2021]		Antisemitism	Gab	B	MFAS	0.91
		Antisemitism	Twitter			0.71
		Category	Gab	M		0.67
			Twitter			0.68
Hate[Kirk*etal.*, 2021]		Hateful	FBHM	B	CLIP	0.56
			Pinterest			0.57
Hate	[Lee*etal.*, 2021]	Hateful	FBHM	B	DisMultiHate	0.830.76
			MultiOFF			0.65
Hate	[Gomez*etal.*,2020]	Hatespech	MMHS150K	B	FCM,Inception-V3,LSTM	0.730.680.70
Hate	[Fersini*etal.*,2019]	Sexist	The MEME	B	Latefusion	0.76
Hate	[Sabat*etal.*,2019]	Hateful	Google	B	BERT,VGG-16,MLP	0.83
Off.[Shang*etal.*, 2021b]		Offensive	Gab	B	FasterR-CNN,ResNet50,	0.690.56
			Reddit		Glove-basedLSTM,BERT,MLP	0.720.49
Off.[Shang*etal.*, 2021a]		Offensive	Reddit	B	YOLOV4,ConceptNET, GNN	0.730.49
			Gab			0.700.55
Off.	[Giri*etal.*,2021]	Offensive	Off. Int.	B	CNN,GloVe,LSTM	0.71
		Off. Int.		M	CNN,FastText,LSTM	0.99
Off.[Suryawanshi*etal.*,2020a]		Offensive	MultiOFF	B	Earlyfusion:StackedLSTM, BiLSTM/CNN-Text,VGG16	0.50
Prop.[Dimitrov*etal.*,2021a]		Prop.Tech.	Facebook	ML	VisualBERT	0.48
Prop.[Tian*etal.*,2021]		Prop.Tech.:(T)	Facebook	ML	Ensemble:BERT,RoBERTa, XLNet,ALBERT,DistilBERT, DeBERTa, Charn-gram	0.59
Prop.[Gupta*etal.*,2021]		Prop.Tech.:(S)	Facebook	ML	RoBERTa	0.48
Prop.[Feng*etal.*, 2021]		Prop.Tech.	Facebook	ML	RoBERTa,Embeddings	0.58
CB	[Hosseinmardi*etal.*,2016]	CBInci.	Instagram	B	SVD+(Unigram,3-gram), kernelPCA+metadata,lin.SVM	0.87
CB	[Suryawanshi*etal.*,2020b]	Troll	TamilMemes	B	ResNet(Tr:TM)	0.52
					ResNet(Tr:TM+iNet)	0.52
					MobileNet(Tr.:TM+iNet+Fl1k)	0.47
				B	ResNet(Tr.:TM+iNet+Fl30k)	0.52

Table1:Summaryoftheexperimentalresultsfortheautomaticdetectionofharmfulmemes.Y/N:positiveandnegativeclasslabels;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-classandMultilabel;TM:TamilMemes,iNet: ImageNet,Fl: Flickr. **Moredetailcanbefoundat** http://github.com/firojalam/harmful-memes-detection-resources

3 RepackagingMemesforHarmfulAgendas

Repackagingviaremixingormimickingthememeisacom-monpracticefacilitatingtheiradoptionacrosslanguagesand cultures [Shifman, 2013], which often implies harm. For ex-ample,popularmemesareoftenrepackagedwithmisogynistic intent.Common ideas that mock specific female identities include*theterriblewife*or*thecrazygirlfriend*. Forexample, the *Distracted Boyfriend* meme[3] has been repackaged many times with varying intent, including harm and humor.

Another example is the *Proud Boys* meme,[4] which has peculiarcharacteristics. Itsproponentsworkingangs,indulge in violence and alcohol, follow a uniform code for appearances and collectively accepted logos to depict their identity.

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 (blackFredPerrypolowithgoldtrim),whilstdisplayingthe OK hand gesture.

4 Cultural Influence and Multilinguality

[Shifman, 2013] introduced the term *user-generated global-ization*,whichreferstotranslation,customization,anddistri-bution 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 mightbeduetoculture-specificinappropriateness. Theyfur-therfoundlimitedlocalizationofthejokeinChinese,German, and Portuguese.

[3]https://knowyourmeme.com/memes/distracted-boyfriend

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

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"OneDoesNotSimplyWalkIntoMordor"meme,which appeared in Germany's Ein Prozent. Most European organizations leverage different genres of images like *share-posts* and*templates*specificallydesignedforonlinecirculationand orthogonaltotheirreverentandparticipatorynatureofmemes. Inadditiontothelocalizedculturaladaptationandcustomization,memes can use multiple languages.Such examples can be found in the TamilMemes dataset [Suryawanshi *etal.*,2020b]. Modellingsuchmemesiscomplex,asisevident from the results reported in [Hegde *et al.*, 2021].

5 MajorChallenges

• **Complexabstraction:** Memescanefficientlyabstractaway complex ideas using creative and powerful customization of visual and linguistic nuances. At the same time, memes with overlappingsnippets,patternedtextandirony,sarcasmorimplicitanti-semitismarenon-trivial[Chandra*etal.*,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,andmockery[Siddiqi*etal.*,2018]. Thismulti-layering ofinfluentialnotionsviamultimodalityposesmajorchallenges forautomaticmemeanalysisandrequiressophisticatedmultimodal fusion to understand novel digital vernaculars.

• **Subjectivityintheannotation:**Subjectiveperceptionsplay asignificantroleformemesasaconsequenceofthecomplex 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. Thisreverberatesasconflictingopinionsduringdatacollection andannotation.Asnotedin[Suryawanshi*etal.*,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], whereinoutof59ambiguousmisogynisticmemes,only23% were correctly identified by crowd-sourced workers, while domain experts achieved 77% expert agreement.

• **Inadequate solutions:**Understanding the visual contentin 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 natureofvisualdescriptorsinmemes,emphasizingtex- tual and discourse-intensive modelling [Shang *et al.*, 2021a;Shang*etal.*,2021b].Visualclusteringtechniquessuch 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 sizelargeenoughtoenablegeneralizationatscale[Al-Natour,2021].Similarly, a keyword-based platform-dependent col- lection of memes could yield a biased representation of the samplespace,andhencecouldover-representtypicalmemetic 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 understandnew harmful concepts and tasks by looking at a few ex- amples, AI systems struggle to generalize well from a few examples[Wang*etal.*,2020].Few-shotlearning(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 evolvingtypesofharmfulmemes[Tsimpoukelli*etal.*,2021;Tejankar*etal.*,2021]. UnliketraditionalAIthatmainlyrelies onpattern-matchingwithlabelleddata,FSL-basedAIsystems canevolvetonewharmfulmemesandpoliciesusingahand- ful of examples and can take action immediately instead of waiting for months for the labelled data to be collected.

• **Contextualization:** Understanding many memes requires complexandmultimodalreasoningthatisbaseduponacertain 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 contextualinformationmaybeconveyedbothindependently 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 socialnetworkingserviceshaveposedchallengesforthesystematic study of online harm detection.For example, the freedom of being anonymous has obscured racial integrity andaccountability,effectivelycomplicatingharmfuldiscourse analysis [Dickerson, 2016]. Moreover, the complex designs andgovernancepoliciesofplatformssuchasWhatsAppmeant thattheyfocusedontheir*secure*butunabatedusefordisseminatingsystematicracism[Matamoros-Ferna´ndez,2020]. As observedby[Zannettou*etal.*,2020],theinvestigationofanactivelyevolvingcommunitylikeGab,usingaHawkesprocess, might err the observations [Zannettou *et al.*, 2020].

• **Identifyingrealinstigatorsofharm:**Poe'slawemphasizes the understandingoftheactualintentwhiledistinguishingbe- tween 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 challengingtoclearlydistinguishbetweenanti-democracyextremistsand anti-democratic alt-right factions [Askanius, 2021].Consequently,onemustalsobecarefulwhileassociatingthealt-right withculture. Itisinsteadahistoricalphenomenonthatlever- ages culture as a tool for its propagation [Dafaure, 2020].

6 FutureForecasting

• **Characterizing vehicles of harm:** Satire is not only used asaprogressivetooltoresistbigotry,butitisalsoweaponized by malicious actors towards hijacking the online discussion [Greene, 2019]. It is thus important to decode the discourse andtounderstandthecommunicationthatmemesarepartof [DeCook, 2018].Exploring the points of the confluence of youthwithfar-rightmemeswillhelphighlightwhereandhow messages of extreme violence circulate and transit back and forthbetweenmaliciousactorsandreceptiveusers[Askaniusand Keller,2021]. Itcouldalsobeinsightfultoexaminehow symptomaticthediscourserhetoricoftheanecdotalreference is,withinthebackdropofrootedantisemiticperspectives,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 relationshipbetweenofflinemicro-aggressionandonlineper- ception of racism will become more prominent in settings whereCaucasiansarenotthemajority.Thispresentsthescope of investigating cross-cultural and cross-contextual implica- tionsofracismexperiencedandperceivedonline[Williams*etal.*, 2016].

• **Empirical *in addition to* theoretical:** There are few com- pelling questions arising from the existing understanding of harmful memes regarding the cause of their potency to insti- gate harm, cross-platform transitioning, and outcomes, fewofthembeingasfollows: Towhatextentarethe"hatejokes" part of the slow yet steady process of normalizing online ex- tremism 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 suchcontent[Askanius,2021]?Clearly,theassessmentofthe prevalenceof differentvisualformslikememes,photography, and artwork in online communications, along with the cryp- tic 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*etal.*,2021],andusingintermediaterepre- sentations to capture higher levels of abstractions that lever- age both the image and the text modalities can help model complex tasks. Moreover, the contextual knowledge supple- mentingsuchabstractinformationbecomesindispensablefor automated meme analysis [Shang *et al.*, 2021a].

• **Multi-classandmulti-labelclassification:**Ashighlighted in Table 1, the existing classification setups are primarily bi- nary.However,amorefine-grainedmulti-classandmulti-label setup can enhance the decision-making process, as required in many scenarios.For example, a meme labelled as hate-ful [Kiela *et al.*, 2020], but which has the characteristics of violenceandmisogyny,losesitsspecificity. Attemptsinthis direction include fine-grained analysis of hateful [Zia *et al.*,2021]andpropagandistic[Dimitrov*etal.*,2021a]memes,de- tectingthevictimstargetedbyharmfulmemes[Pramanick*etal.*, 2021a;Sharma *et al.*, 2022a], and understanding who is the hero, the villain, and the victim [Sharma *et al.*, 2022b].

• **Memeticmoderation:**Counter-narrativescanhelpaddress the selective targeting via harmful memes [Williams, 2020]. Theutilityofthepost-modern transgressionandhumormust 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]. Cre- ating counter memes can help raise awareness about racial issues [Yoon, 2016].Reclaiming the digital space and in- dulginginsubversivereactionsbyleveragingtheparticipatory humor using *digilanties* (online vigilantes) can help mitigate thecollectivemenaceimpendedbythesystematicandsubtle oppression of women [Drakett *et al.*, 2018].

7 Conclusion

Wepresentedasurveyofthecurrentintelligenttechnologies fordetectingandunderstandingharmfulmemes.Basedon a systematic analysis of recent literature, we first proposed a newtypologyofharmfulmemes,andthenwehighlightedand summarized the relevant state of the art.We then discussed the lessons learned and the major challenges that need to be overcome. Finally,wesuggestedseveralresearchdirections, which we forecast will emerge in the near future.

Acknowledgments

The work was partially supported by aWiproresearchgrant, RamanujanFellowship,theInfosysCentreforAI,IIITDelhi, and ihub-Anubhuti-iiitd Foundation, set up under the NM-ICPSschemeoftheDepartmentofScienceandTechnology, India.It is also part of the Tanbih mega-project, which is developedattheQatarComputingResearchInstitute,HBKU, andaimstolimittheimpactof"fakenews,"propaganda,and media bias by making users aware of what they are reading.

References

[Acar*etal.*,2013]Esra Acar, Frank Hopfgartner, and Sahin Al- bayrak.ViolencedetectioninHollywoodmoviesbythefusionof visual and mid-level audio cues.In *MM*, pages 717–720, 2013.

[Afridi*etal.*,2021]TariqHabibAfridi,AftabAlam,Muham- mad Numan Khan, Jawad Khan, and Young Koo Lee.A multi- modal memes classification: A survey and open research issues. In *SCA*, pages 1451–1466, 2021.

[Al-Natour, 2021]Ryan Al-Natour.The digital racist fellowship behindtheanti-aboriginalinternetmemes.*J.ofSoc.*,57(4):780– 805, 2021.

[Alam*etal.*,2021] FirojAlam,StefanoCresci,Tanmoy Chakraborty,FabrizioSilvestri,DimiterDimitrov,Gio- vanniDaSanMartino,ShadenShaar,HamedFirooz,andPreslav 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: Evi- dence from a two-wave panel study among young adults.*New Media & Soc.*, 21(11-12):2422–2442, 2019.

[AskaniusandKeller,2021]TinaAskaniusandNadineKeller.Mur- derfantasiesinmemes: fascistaestheticsofdeaththreatsandthe banalizationofwhitesupremacistviolence.*Info.,Comm.&Soc.*, 24(16):2522–2539, 2021.

[Askanius,2021]TinaAskanius. Onfrogs,monkeys,andexecution memes:Exploring the humor-hate nexus at the intersection of neo-Naziandalt-rightmovementsinSweden.*Tel.&NewMedia*, 22(2):147–165, 2021.

[Banko*etal.*,2020] MicheleBanko,BrendonMacKeen,andLaurie Ray.A unified taxonomy of harmful content.In *WOAH*, pages 125–137, 2020.

[BondielliandMarcelloni,2019] AlessandroBondielliand Francesco Marcelloni.A survey on fake news and rumour detection techniques.*Info. Sci.*, 497:38–55, 2019.

[Brooke, 2019] Sian Brooke.“Condescending, Rude, Assholes”: FraminggenderandhostilityonStackOverflow.In*WALO*,pages 172–180, 2019.

[Brown*etal.*,2020]Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan,PranavShyam,GirishSastry,AmandaAskell,etal. Languagemodelsarefew-shotlearners.*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*etal.*,2020]Yen-ChunChen,LinjieLi,LichengYu,Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER:Universalimage-textrepresentationlearning.In*ECCV*, pages 104–120, 2020.

[CraneandFrench,2021]Tim Crane and Craig French. The problem of perception.In *The Stanford Encyclopedia of Philosophy*. Stanford University, 2021.

[DaSanMartino*etal.*,2019]GiovanniDaSanMartino,Seunghak Yu, Alberto Barro´n-Ceden˜o, Rostislav Petrov, and Preslav Nakov. Fine-grainedanalysisofpropagandainnewsarticle.In*EMNLP-IJCNLP*, pages 5636–5646, 2019.

[Da San Martino *et al.*, 2020]Giovanni Da San Martino, Stefano Cresci, Alberto Barro´n-Ceden˜o, Seunghak Yu, Roberto Di Pietro, andPreslavNakov.Asurveyoncomputationalpropagandadetection.In *IJCAI*, pages 4826–4832, 2020.

[Dafaure,2020]MaximeDafaure. The“greatmemewar:”Thealt- right and its multifarious enemies.*Angles*, (10), 2020.

[DeCook, 2018]Julia R. DeCook. Memes and symbolic violence: #proudboys and the use of memes for propaganda and the construction of collective identity.*LMT*, 43(4):485–504, 2018.

[Dickerson,2016]NikolasDickerson. Constructingthedigitalized sportingbody:BlackandwhitemasculinityinNBA/NHLinternet memes.*Comm. & Sport*, 4(3):303–330, 2016.

[Dimitrov*etal.*,2021a]Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino.Detecting propaganda techniquesinmemes.In*ACL-IJCNLP*,pages6603–6617,2021.

[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]TabithaFairchild. It’sfunnybecauseit’strue:The transmission of explicit and implicit racism in internet memes. Virginia Commonwealth University, 2020.

[Fan*etal.*,2021]ZhiboFan,YuchenMa,ZemingLi,andJianSun. Generalized few-shot object detection without forgetting.In *CVPR*, pages 4527–4536, 2021.

[Feng*etal.*,2021]Zhida Feng, Jiji Tang, Jiaxiang Liu, Weichong Yin,ShikunFeng,YuSun,andLiChen.AlphaatSemEval-2021 task6:Transformerbasedpropagandaclassification.In*SemEval*, pages 99–104, 2021.

[Fersini*etal.*,2019]Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs.Detecting sexist meme on the web: A study on textual and visual cues.In *ACIIW*, pages 226–231, 2019.

[FortunaandNunes,2018]PaulaFortunaandSe´rgioNunes. Asurveyonautomaticdetectionofhatespeechintext.*CSUR*,51(4):1– 30, 2018.

[Gan*etal.*,2020]Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, YuCheng,andJingjingLiu.Large-scaleadversarialtrainingfor vision-and-languagerepresentationlearning. *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. Scalabledetectionofoffensiveandnon-compliantcontent/logo in product images.*WACV*, pages 2236–2245, 2020.

[Gasparini*etal.*,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.