Findings of the CONSTRAINT 2022 Shared Task on Detecting the Hero, the Villain, and the Victim in Memes

Shivam Sharma[1,3], **Tharun Suresh**[1], **Atharva Kulkarni**[1], **Himanshi Mathur**[1], **Preslav Nakov**[2], **Md. Shad Akhtar**[1], **Tanmoy Chakraborty**[1]
[1]Indraprastha Institute of Information Technology - Delhi, India
[2]Qatar Computing Research Institute, HBKU, Doha, Qatar
[3]Wipro AI Labs, India
{shivams, tharun20119, atharvak, himanshi18037, shad.akhtar, tanmoy}@iiitd.ac.in
pnakov@hbku.edu.qa

Abstract

We present the findings of the shared task at the CONSTRAINT 2022 workshop on "*Hero, Villain, and Victim: Dissecting Harmful Memes for Semantic Role Labeling of Entities*." The task aims to delve deeper into meme comprehension by deciphering the connotations behind the entities present in a meme. In more nuanced terms, the shared task focuses on determining the victimizing, glorifying, and vilifying intentions embedded in meme entities to explicate their connotations. To this end, we curate `HVVMemes`, a novel meme dataset of about 7,000 memes spanning the domains of COVID-19 and US Politics, each containing entities and their associated roles: *hero*, *villain*, *victim*, or *other*. The shared task attracted 105 registered participants, but eventually only nine of them made official submissions. The most successful systems used ensembles combining textual and multimodal models, with the best system achieving an F1-score of 58.67.

1 Introduction

The unwarranted spread of misinformation (Wu et al., 2019; Hardalov et al., 2022), propaganda (Da San Martino et al., 2020a,b), fake news (Lazer et al., 2018; Vosoughi et al., 2018), COVID-19 infodemic (Alam et al., 2021b; Nakov et al., 2022), hate speech (MacAvaney et al., 2019; Zampieri et al., 2019a), and other harmful content (Nakov et al., 2021) has plagued social media. Lately, *memes* have emerged as a powerful multimodal means to disseminate malicious content due to their ability to circumvent censorship norms (Mina, 2014) and to their fast-spreading nature. With an aptly crafted combination of images and text, a seemingly naïve meme can easily become a source of harmful information diffusion. As a result, exploring the noxious side of memes has become a pressing research topic; see also recent surveys on harmful memes (Sharma et al., 2022b) and on multimodal disinformation detection (Alam et al., 2021a).

While meme analysis has been studied in a variety of contexts, such as hate speech (Zhou et al., 2021; Kiela et al., 2020) harmfulness (Pramanick et al., 2021a,b), emotions (Sharma et al., 2020), misinformation (Zidani and Moran, 2021), sarcasm (Kumar and Garg, 2019), offensiveness (Suryawanshi et al., 2020), and propaganda (Dimitrov et al., 2021a,b), limited forays have been made on comprehending the role of the entities that make up a meme. This is our main focus here: on identifying the *hero*, the *villain*, and the *victim* entities present in a meme. Given a meme and a list of the entities it involves, the task is to identify which entity plays what role. Such categorization of the entities in the meme can help understand the entity-specific connotation and their nature, attitudes, decisions, and demeanour. For instance, when the meme creators intend to spread misinformation and hatred towards minority communities or to defame certain individuals, politicians, or organizations, they would depict the target entities as *villains*. Similarly, when the intent is to shed light on the deplorable state of certain entities or to glorify them, these entities would be portrayed as *victims* or as *heroes*, respectively.

Fig. 1 depicts apt examples for *hero*, *villain*, and *victim* categorization of the entities in a meme. The meme in Fig. 1a draws a comparison between Abraham Lincoln, John F. Kennedy, Barack Obama, and Donald Trump, where the former three are portrayed as *heroes*, while Donald Trump is shown in negative light, as a *villain*. Similarly, Fig. 1b mocks Jill Stein and the Green Party as *villains* for allegedly getting bribed by the rich. Fig. 1c on the other hand, frames the Republican Party as a *villain*, for their inconsiderate views on the poor, the minorities, and women, thus making them the *victims*. In conclusion, through depictions of heroism, villainy, and victimization, memes act as an appealing means to propagate certain views about the targeted entities.

[Barack Obama, John F. Kennedy, Abraham Lincoln] [Donald Trump]

(a)

[Green Party, Jill Stein]

(b)

[Women, Poor, Minorities] [Republican Party]

(c)

Figure 1: Examples of heroes, villains and victims, as portrayed within memes.

While some previous meme studies have sought to identify harmfulness and the entities (Sharma et al., 2022a) or the categories that are being targeted, e.g., a person, a group, an organization, or society (Pramanick et al., 2021a,b), none of them has scrutinized the entity's connotation. Our shared task aims to bridge this gap. We release HVVMemes, a meme dataset with about 7,000 memes on COVID-19 and US Politics, where each meme is annotated with a list of entities, each labeled with its role: *hero*, *villain*, *victim*, or *other*. The shared task attracted 105 teams, and nine of them made official submissions. Most teams fine-tuned pre-trained language and multimodal models or used ensembles, with the best system achieving an F1-score of 58.67. We discuss the submissions and their approaches in more detail in Section 5.

Despite the growing body of research on meme analysis, understanding the connotation underlying the individual entities in the meme remains a challenging endeavour. Their camouflaged semantics, satirical outlook, and cryptic nature make their analysis a daunting task (Sabat et al., 2019). Moreover, categorizing the entities as *heroes*, *villains*, or *victims* requires real-world and commonsense knowledge, which often are not present in popular pre-trained language models. Thus, it should not be surprising that, as the shared task's results show, off-the-shelf multimodal models, as well as various ensembles thereof, struggle with this task (Kiela et al., 2020). This highlights that the current state-of-the-art visual-linguistic models are unable to grasp the veiled information present in the memes. Thus, we hope that the dataset and task will foster further research in this interesting direction.

More details about the shared task is available at http://constraint-lcs2.github.io/

2 Related Work

Studies on Online Targeting. Previous work studied affective content in the context of harmful discourse in social media (Zainuddin et al., 2017, 2018; Gautam et al., 2020; Ousidhoum et al., 2019). Sarcastic content was detected by leveraging data sparseness (Zainuddin et al., 2019) towards studying aspect-based sentiment analysis. Shvets et al. (2021) established enhancements in target detection by examining generic concept extraction for hate speech detection. Targeted protected categories were characterised by harmful online engagements whilst addressing societal bias along with explainability (Sap et al., 2020; Mathew et al., 2021). For affective target characterisation, sequence modeling was explored in a hierarchical formulation of stacked BiGRUs (Ma et al., 2018) as well as in low-resource scenarios (Mitchell et al., 2013). Most approaches did not consider the variability in target referencing and the associated affective spectrum (Shvets et al., 2021). Finally, (Gomez-Zara et al., 2018) discussed hero/villain/victim analysis of news text; unlike their work, here we focus on multi-modality and memes.

Studies on Detecting Harmful Memes. The constant transitioning of harmful memes from unfiltered and largely anonymous communities and platforms such as 4chan, Reddit, and Gab to more mainstream social media has made the entire social media ecosystem both sensitive and vulnerable to extremism (Zannettou et al., 2018). Research on offense (Suryawanshi et al., 2020), hate speech (Kiela et al., 2020; Gomez et al., 2020), and online harm detection (Pramanick et al., 2021b) has found the availability of large datasets and the use of multi-modal frameworks crucial for these tasks.

Additional contextual cues involving commonsense knowledge (Shang et al., 2021), semantic entities, cues about the protected categories (Pramanick et al., 2021b; Karkkainen and Joo, 2021), along with other meta information, have also been explored for characterising various aspects of the online harm conveyed by memes. Most such tasks address *affect* detection at various levels of granularity, sometimes organised in a taxonomy. Still, none of these tasks has focused on explicitly modeling the complex narrative framework of the memetic discourse surrounding the specific entities referred to in the meme. With this in mind, here we attempt to alleviate a few associated challenges by exploring the feasibility of entity-specific visual-semantic role labelling for memes.

Other Related Shared Tasks. Several shared tasks have targeted the broad field of harmful social media content. Some tasks investigated the characterisation of *offensive language, hate speech, profanity*, and associated fine-grained attributes such as *implicit* and *explicit* implications in binary, multi-class, multi-label, and hierarchical settings (Struß et al., 2019; Zampieri et al., 2019b, 2020). Their coverage has been fairly comprehensive in terms of the languages covered including *Arabic, Danish, Greek, English, Turkish*, and *Dravidian Languages* like *Tamil, Malayalam, Kannada* as well as *German and English/Indo-Aryan code-mixing* (Zampieri et al., 2019b; Mubarak et al., 2020; Zampieri et al., 2020; Chakravarthi et al., 2021; Modha et al., 2021). They also address harmful content dissemination, targeting various protected categories such as *religious affiliation, national origin, sex*, etc. (Zhang et al., 2019). Other efforts have targeted misinformation, propaganda, and persuasiveness detection (Aly et al., 2021; Shaar et al., 2021; Da San Martino et al., 2020a), where the goal is to detect verifiable claims, their veracity, span, and check-worthiness. Persuasive technique detection has also been explored for images besides text-based content, e.g., Dimitrov et al. (2021b) introduced the task of propaganda in *memes*.

Some tasks have attempted to address affect concerning various targets. Xu et al. (2016) focused on stance prediction for given targets, i.e., whether the comment is in favour or against the target, both in supervised and in unsupervised scenarios. Molla and Joshi (2019) modeled sarcastic targeting of specific entities. Rosenthal et al. (2017) focused on sentiment analysis in Twitter.

Domain	Splits	# Memes	# Referenced Entities				
			Hero	*Villain*	*Victim*	*Other*	**Total**
COVID-19	Train	2,700	163	576	317	2,438	3,494
	Val	300	19	65	40	268	392
	Test	381	18	106	50	359	533
	Total	3,381	200	747	407	3,065	4,419
Politics	Train	2,852	230	1,308	441	2,617	4,596
	Val	350	27	166	58	317	568
	Test	350	31	167	45	308	551
	Total	3,552	288	1,641	544	3,242	5,715

Table 1: Statistics about our HVVMemes dataset.

In contrast, here we focus not only on the polarity of the target entity, but also on understanding complex connotations such as *glorification*, *vilification*, and *victimisation* in memes. This is both challenging and important, as memetic discourse has taken over a sizable portion of online engagement and as it requires specialised moderation given its multimodal nature.

3 Dataset Curation

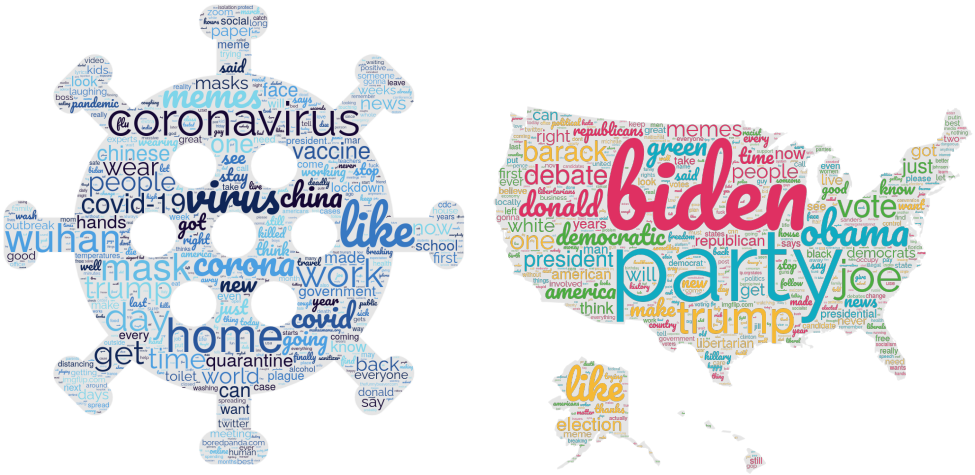
Towards curating a dataset that would enable the identification of *hero*, *villain*, and *victim* as roles in memes, we leveraged and reannotated the HarMeme dataset released in (Pramanick et al., 2021b), and we call this new dataset HVVMemes. HarMeme includes 3,544 memes about COVID-19 and 3,552 memes about US Politics, which are annotated for *harmfulness* as well as for *target type*, in case the meme is harmful, with four categories for the latter: *individual*, *organisation*, *community*, and *society*. Table 1 gives some statistics about HVVMemes (note that for COVID-19, we filtered out some of the memes in HarMeme, keeping 3,381 of the original 3,554 memes). As a general trend for both domains, we observe a neutral reference for most of the entities mentioned in the memes (3,065 for COVID-19, and 3,242 for US Politics); for such cases, we assign a fourth category: *other*. We further see that *villain* is the second most frequent role (747 memes for COVID-19, and 1,641 for US Politics), followed by *victim* (407 memes for COVID-19, and 544 for US Politics), and then *hero* (200 memes for COVID-19, and 288 for US Politics). We believe that this is a realistic representation of social media engagement involving memes, which are mostly humorous with neutral connotations, and less frequently harmful by indulging in vilification. Victimisation can also be interpreted as a countering resistance to incessant vilification. Finally, glorification is generally the weakest voice in memetic discourse.

S. No.	Annotation Guidelines
1	Meme author's perspective needs to be considered as the frame of reference, while assigning roles.
2	Towards complete assimilation, both visual and textual cues should be factored in.
3	Relevant background context should be acquired before assigning roles.
4	Ambiguous memes can be categorised as *other*.
5	A 3-point Likert scale based mental frame of reference, implying *negative*, *neutral* and *positive* sentiments involved, should steer the connotation adjudication.
6	All reasonably *intelligible* (without ambiguity) entities that are referred to in the meme must be considered as valid targets.
7	Entities with multiple interpretations should be categorised as *other*.
8	The role of the original speaker of a quote, as expressed within a meme, must not be presumed.

Table 2: Key considerations in our annotation guidelines.

Entity	Resolution Remark
`Corona`	resolved to `Corona Beer` (whenever valid).
`Govt.`	resolved to `Government`.
`Putin`	resolved to `Vladimir Putin`.
`CDC`	standardised as `Centre of Disease Control (CDC)`.

Table 3: Examples of *resolution remarks* that we provided to the annotators towards entity identification.



(a) COVID-19 (b) US Politics

Figure 2: Word clouds for (a) COVID-19 and (b) US Politics domains in `HVVMemes`.

3.1 Annotation Setup

Since entity role labelling is complex and subjective, we formulated clear annotation guidelines, which are summarized in Table 2. Each meme was annotated by three annotators, and the disagreements were resolved with the help of a consolidator. We asked the annotators (*i*) to identify the entities, and (*ii*) to assign roles to these entities.

3.1.1 Identifying the Entities

This step requires the annotators to elicit all entities that the meme refers to. This includes *persons*, *norp* (nationalities, religious, or political groups), *facilities, organizations, geopolitical entities, locations*, *products*, and other, as defined by spaCy's label scheme for named entity recognition.[1]

To assist the annotators, we provided them an exhaustive list of all automatically identified entities along with *resolution remarks* whenever needed as shown in Table 3. Note that the annotators were not restricted to select entities from our provided list, which can be error-prone as automatic named entity recognition is not perfect; in fact, they were encouraged to add additional entities as needed, e.g., such shown in the image, but not mentioned in the textual part of the meme.

Fig. 2a shows a word cloud visualization of the entities referenced in COVID-19 memes: we can see social, global, political, and economic entities such as *coronavirus, China, home, Wuhan, mask, work*, etc. Similarly, in Fig. 2b shows a word cloud for US Politics memes, where we see entities like *Biden, party, Donald, Democratic, Obama*, etc.

[1] `spacy.io/models/en#en_core_web_sm`

To assess the general agreement between the annotators, we considered an agreement towards entity identification if at least two annotators agreed on an entity in the meme. The number of memes with agreed entities was normalised by the total number of memes with at least one valid entity assignment by the annotators. This was done independently of the implied role category, as the emphasis in this first step is on entity identification. The highest agreement towards this was 0.98, which suggests the reliability associated with the annotator's collective understanding of the task. We followed a similar approach for the overall role-wise inter-annotator agreement; see below.

3.1.2 Role Assignment

The annotation was done in three stages: (*i*) dry-run, (*ii*) complete annotation, and (*iii*) consolidation. As part of the dry-run, the annotators and the consolidator annotated a random subset of 250 memes, assigning the entities the roles of *hero*, *villain*, *victim*, and *other*. Then, we gave them feedback and we trained them carefully by issuing detailed guidelines that included the formal definitions of the role categories and the instructions exemplifying the edge scenarios identified as part of the dry-run disagreements. In the second stage, the annotators performed a complete annotation. This was followed by a third consolidation stage with the help of a consolidator.

Due to the varying annotation responses and co-referencing for each role, conventional annotation agreement measures are not suitable for our setup. We consider an agreement when at least two annotators agree on one of the candidate entities for a particular role, which we formalize as the following role-wise agreement score a:

$$a = \frac{v_{agr}}{v_{tot}} \tag{1}$$

We define v_{agr}, which refers to the total number of valid agreements, and v_{tot}, which is the total number of valid responses, as follows:

$$v_{agr} = \sum_{i=1}^{N} I_i;\ v_{tot} = \sum_{i=1}^{N} Z_i \tag{2}$$

where I_i is a valid agreement (1, iff two or more annotators agree on an entity in example i), Z_i is a valid response (1, iff at least one annotator provides a valid entity as a response in example i), and N is the total number of examples in the dataset.

Roles	**Covid-19** (a)		**US Politics** (a)		**Stage-3**
	Stage-2	**Stage-3**	**Stage-2**	**Stage-3**	**Avg.** (a)
Hero	0.30	0.54	0.36	0.51	0.53
Villain	0.31	0.55	0.55	0.73	0.64
Victim	0.21	0.55	0.24	0.43	0.49
Other	0.58	0.68	0.76	0.88	0.78
Avg.	0.35	**0.58**	0.48	**0.64**	**0.61**

Table 4: Inter-annotator agreement (IAA) summary for *completed* (Stage-2) and *consolidated* (Stage-3) stages of the annotation process. Note that the average IAA for the dry-run (Stage-1), for COVID-19 and US Politics combined, was 0.50 (hero), 0.35 (villain), 0.14 (victim), and 0.55 (other).

In the *first* dry-run stage of the annotation process, the annotators worked on 250 memes, and then we examined their agreement, which was 0.50, 0.35, 0.14, and 0.55, for the roles of *hero*, *villain*, *victim*, and *other*, respectively, for COVID-19 and US Politics combined. The inter-annotator agreement for stages 2 and 3 is shown in Table 4. We can see that the average agreement scores after the *completion* stage (stage-2) are 0.35 and 0.48 for COVID-19 and US Politics, respectively. After the consolidation stage (stage-3), these numbers increased to 0.58 and 0.64, respectively.

3.2 Role-wise Analysis of `HVVMemes`

The distribution of the referencing entities within our `HVVMemes` dataset is somewhat skewed towards specific entities as well as towards specific predominant roles for these specific entities. The entities fairly emulate the prevalent trends and discourse topics that social media engagement around the period of the dataset collection reflected, which was at the onset of the COVID-19 pandemic and the surrounding political outlook within the United States of America. We observed that entities like *Donald Trump* and *China* were referenced almost equally in *COVID-19* memes as a *villain* and *other*, while other entities are invariably referenced as *other* using humor, sarcasm, limerick, etc. For the domain of *US Politics*, on one hand, entities like *Donald Trump*, the *Democratic Party*, the *Republican Party*, and the *Democrats* are observed to have similar trend of pre-dominantly being referenced as a *villain* and *other*, and on the other hand, as a general trend, most of the memes have at least one vilified reference.

Rank	System	Precision	Recall	F1
1	shiroe	55.76	62.73	58.67
2	jayeshbankoti	53.58	59.45	56.01
3	c1pher	53.91	57.25	55.24
4	zhouziming	54.19	55.36	54.71
5	smontariol	57.96	44.97	48.48
6	zjl123001	47.98	44.97	46.18
7	amanpriyanshu	30.98	34.35	31.94
8	IIITDWD	25.57	23.79	23.86
9	rabindra.nath	25.30	25.30	23.72

Table 5: Leaderboard summary for the shared task.

4 Shared Task Details

The CONSTRAINT 22 Shared Task on Detecting the Hero, Villain, and the Victim in Memes asked to predict which entities are glorified, vilified, and victimised in a given meme. We gave the participants the above-described labeled training and validation datasets, where for each meme, we had the list of corresponding entities and their labeled role. The task was, given a meme and a list of entities, to predict the role of each of these entities in the meme. We provided the data split by topic (COVID-19 and US Politics), as discussed in Section 3. For the test set, we combined and shuffled the memes from the two topics, and we provided the memes with a list of corresponding entities, but no labels.

The task was organized on CodaLab, an open-source platform widely used to host machine learning and data science competitions. Our competition link[2] provided all the necessary resources for the participants including archived news, notifications, and forum posts communicated during the running of the competition. We allowed the participants a maximum of 25 submissions, and the best submission was considered for the leaderboard.

The official evaluation measure was macro-F1 score, as we have an imbalanced multi-class problem. We further report precision and recall.

5 Participation and Results

The total of 105 teams registered for the competition, and nine of them made submissions to the leaderboard, making a total of 71 attempts to improve their scores. The teams tried a variety of approaches, and below we discuss the approaches by the six teams who also submitted a system description paper with information about their runs.

- **shiroe/jayeshbankoti** (Kun et al., 2022) achieved the best results overall. One of the distinctive approaches that the authors followed was to make use of Celebrity face detection from the input meme images using Giphy's Github.[3] In addition, a sub-image detector using YoloV5[4] leveraged the bounding boxes for memes with multiple images. This was input into an ensemble model of DeBERTa (He et al., 2021) + RoBERTa (Liu et al., 2019) + ViLT (Kim et al., 2021) + EfficientNetB7 (Tan and Le, 2019) with averaging of the predictions in the final layer. Though they incorporated a celebrity detector, the lack of other external knowledge limited their system performance. Their source code is available at https://bitbucket.org/logicallydevs/constraint_2022/src/master/

- **c1pher** (Singh et al., 2022) were ranked third. It is remarkable that they achieved this result using just the text input. They formulated the problem as a Multiple Choice Question Answering Task (MCQA), and they used an ensemble of three modules: twitter-xlm-roberta + COVID-BERT (Müller et al., 2020) + BERT-tweet (Nguyen et al., 2020). They further added a sentiment module trained using RoBERTa, with the final classification layer comprising Support Vector Machine (SVM). A major drawback of this approach is that they ignored the image as an input altogether.

- **zhouziming/zjl123001** (Zhou et al., 2022) leveraged the Visual Commonsense Reasoning (VCR) framework in a multimodal model. They built an ensemble of VisualBERT (Li et al., 2019) + UNITER (Chen et al., 2020) + OSCAR (Li et al., 2020) + ERNIE-Vil (Yu et al., 2021), combined using an SVM. To handle the disproportionately large number of *Other* examples, they introduced loss-reweighting. The lack of sufficient external knowledge and position information about the OCR text with the image restricted their system performance. Their source code is available at https://github.com/zjl123001/DD-TIG-Constraint

[2] https://codalab.lisn.upsaclay.fr/competitions/906

[3] http://github.com/Giphy/celeb-detection-oss

[4] https://github.com/ultralytics/yolov5

System	BERT	R-BERT	D-BERT	*CLIP*	EB7	OFA	ViLT	ViT	VB	U	O	E-V	SVM	XGB	BF	VADER	W-P
shiroe		✓	✓		✓		✓										
c1pher	✓	✓											✓				
zhouziming									✓	✓	✓	✓	✓				
smontariol				✓		✓			✓					✓			
IIITDWD																✓	✓
rabindra.nath	✓							✓					✓		✓		

Table 6: Models used by the participants as part of their system submissions. **R-BERT:** RoBERTa, **D-BERT:** DeBERTa, **EB7:** EfficientNetB7, **OFA:** Once-for-All, **ViLT:** Visual and Language Transformer, **ViT:** Visual Transformer, **VB:** Visual BERT, **U:** UNITER, **O:** OSCAR, **E-V:** ERNIE-Vil, **SVM:** Support Vector Machines, **XGB:** XGBoost, **BF:** Block Fusion and **W-P:** Wu-Palmer.

- **smontariol** (Montariol et al., 2022) experimented with sampling to handle data imbalance, trying six strategies. On top of that, they used an ensemble of CLIP (Radford et al., 2021) + VisualBERT + OFA (Cai et al., 2020) with XGBoost as the final layer for classification. The potential limitations of this approach include OCR errors and issues with image–text correspondence. Their source code is available at https://github.com/smontariol/mmsrl_constraint

- **IIITDWD** (Fharook, 2022) combined sentiment- and lexicon-based approaches to associate sentiment polarity and roles with each entity. For sentiment classification, they used VADER[5]. Moreover, to associate commonly used words for *hero*, *villain*, and *victim*, they developed a corpus and used Wu-Palmer similarity.[6] The way was done and its impact are described in insufficient detail. Their source code is available at https://github.com/fharookshaik/shared-task_constraint-2022

- **rabindra.nath** (Nandi et al., 2022) proposed an approach using BLOCK fusion (Benyounes et al., 2019) for combining the image with text embeddings. They used a combination of ViT (Bobicev and Sokolova, 2017) and BERT (Devlin et al., 2019) for the image and for the text, respectively, followed by SVM as the final layer for classification. The empirical approach limits their system performance despite adding several data augmentation techniques. Their source code is available at https://github.com/robi56/harmful_memes_block_fusion

[5] https://pypi.org/project/vaderSentiment/

[6] https://arxiv.org/ftp/arxiv/papers/1310/1310.8059.pdf

The evaluation results for the above systems are shown in Table 5. We can see that the macro-F1 scores range between 58.67 and 23.72, with a mean of 44.31 and a median of 48.48.

Table 6 further gives a summary of the most important components of the participating systems. We can see that one commonly used architecture is BERT and its variants, including multi-modal variants, whereas SVM is the preferred way to combine the components of ensemble systems.

6 Conclusion

Understanding and interpreting the connotations behind the entities in a meme is a difficult problem, which we pioneered in this shared task. Given a meme and a list of entities, the task asks to detect the role of each entity as a *hero*, a *villain*, a *victim*, or *other*. We curated HVVMemes, a large-scale meme dataset of 7,000 memes spanning the domains of COVID-19 and US Politics, annotated with the entities they refer to as well as with their role. The shared task attracted 105 registered participants, out of which nine made official submissions, and six submitted papers describing their systems. We hope that our dataset and task setup will enable further research towards understanding how entities are portrayed in memes.

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, thus promoting media literacy and critical thinking.

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021a. A survey on multimodal disinformation detection. *CoRR*, abs/2103.12541.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021b. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of EMNLP*, pages 611–649, Punta Cana, Dominican Republic.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Punta Cana, Dominican Republic.

Hedi Ben-younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. 2019. BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI '19, Honolulu, Hawaii, USA.

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '17, pages 97–102, Varna, Bulgaria.

Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. 2020. Once-for-All: Train One Network and Specialize it for Efficient Deployment. *arXiv:1908.09791 [cs, stat]*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. *arXiv:1909.11740 [cs]*.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online).

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence*, IJCAI-PRICAI '20, pages 4826–4832, Yokohama, Japan.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conf. of the North American Chapter of the Association for Comp. Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021a. Detecting propaganda techniques in memes. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '21, pages 6603–6617.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online.

Shaik Fharook. 2022. Are you a hero or a villain? a semantic role labelling approach for detecting harmful memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, CONSTRAINT '22, Dublin, Ireland.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. #MeTooMA: Multi-aspect annotations of tweets related to the MeToo movement. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):209–216.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision*, WACV '20, pages 1459–1467.

Diego Gomez-Zara, Miriam Boon, and Larry Birnbaum. 2018. Who is the hero, the villain, and the victim? Detection of roles in news articles using natural language techniques. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, page 311315, Tokyo, Japan.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of NAACL*, Seattle, Washington, USA.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654 [cs]*.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, WACV '21, pages 1548–1558.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33 of *NeurIPS '20*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *ICML '21*, pages 5583–5594.

Akshi Kumar and Geetanjali Garg. 2019. Sarc-M: Sarcasm detection in typo-graphic memes. In *Proceedings of the International Conference on Advances in Engineering Science Management & Technology*, ICAESMT '19, Dehradun, India.

Ludovic Kun, Jayesh Bankoti, and David Kiskovski. 2022. Logically at the CONSTRAINT 2022: Multimodal role labelling. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, CONSTRAINT '22, Dublin, Ireland.

David M.J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv:1908.03557 [cs]*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the 16th European Conference in Computer Vision*, volume 12375 of *ECCV '20*, pages 121–137, Glasgow, UK.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Dehong Ma, Sujian Li, and Houfeng Wang. 2018. Joint learning for targeted sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4737–4742, Brussels, Belgium.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):1–16.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

An Xiao Mina. 2014. Batman, Pandaman and the Blind Man: A case study in social change memes and internet censorship in China. *Journal of Visual Culture*, 13(3):359–375.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, Seattle, Washington, USA.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '21, pages 1–3.

Diego Molla and Aditya Joshi. 2019. Overview of the 2019 ALTA shared task: Sarcasm target identification. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, ALTA '19, pages 192–196, Sydney, Australia.

Syrielle Montariol, Étienne Simon, Arij Riabi, and Djamé Seddah. 2022. Fine-tuning and sampling strategies for multimodal role labeling of entities under class imbalance. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, CONSTRAINT '22, Dublin, Ireland.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France.

Martin Müller, Marcel Salathé, and Per E. Kummervold. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv:2005.07503 [cs]*.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, and Javier Beltrán. 2022. The CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection. In *Proceedings of the 44th European Conference on IR Research: Advances in Information Retrieval*, ECIR '22, pages 416–428, Berlin, Heidelberg.

Preslav Nakov, Vibha Nayak, Kyle Dent, Ameya Bhatawdekar, Sheikh Muhammad Sarwar, Momchil Hardalov, Yoan Dinkov, Dimitrina Zlatkova, Guillaume Bouchard, and Isabelle Augenstein. 2021. Detecting abusive language on online platforms: A critical analysis. *arXiv/2103.00153*.

Rabindra Nath Nandi, Firoj Alam, and Preslav Nakov. 2022. Detecting the role of an entity in harmful memes: Techniques and their limitations. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, CONSTRAINT '22, Dublin, Ireland.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP '20, pages 9–14, Online.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of ACL*, ACL-IJCNLP '21, pages 2783–2796.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, pages 502–518, Vancouver, Canada.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv:1910.02334*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 5477–5490, Online.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov, and Anna Feldman. 2021. Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 82–92, Online.

Lanyu Shang, Christina Youn, Yuheng Zha, Yang Zhang, and Dong Wang. 2021. KnowMeme: A knowledge-enriched graph neural network solution to offensive meme detection. In *Proceedings of the 2021 IEEE 17th International Conference on eScience*, eScience '21, pages 186–195.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 759–773.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022a. DISARM: Detecting the victims targeted by harmful memes. In *Findings of NAACL*, Seattle, Washington, USA.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022b. Detecting and understanding harmful memes: A survey. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, IJCAI-ECAI '22, Vienna, Austria.

Alexander Shvets, Paula Fortuna, Juan Soler, and Leo Wanner. 2021. Targets and aspects in social media hate speech. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 179–190, Online.

Pranaydeep Singh, Aaron Maladry, and Els Lefever. 2022. Combining language models and linguistic information to label entities in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, CONSTRAINT '22, Dublin, Ireland.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. Proceedings of the 15th Conference on Natural Language Processing, pages 352 – 363, München [u.a.].

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France.

Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *ICML '19*, pages 6105–6114, Long Beach, California, USA.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor. Newsl.*, 21(2):8090.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of NLPCC shared task 4: Stance detection in Chinese microblogs. In *Natural Language Understanding and Intelligent Applications*, pages 907–916, Cham.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3208–3216.

Nurulhuda Zainuddin, Ali Selamat, and Roliana Ibrahim. 2017. Twitter hate aspect extraction using association analysis and dictionary-based approach. In *New Trends in Intelligent Software Methodologies, Tools and Techniques - Proceedings of the 16th International Conference (SoMeT'17)*, volume 297 of *Frontiers in Artificial Intelligence and Applications*, pages 641–651, Kitakyushu City, Japan.

Nurulhuda Zainuddin, Ali Selamat, and Roliana Ibrahim. 2018. Evaluating aspect-based sentiment classification on Twitter hate speech using neural networks and word embedding features. In *New Trends in Intelligent Software Methodologies, Tools and Techniques*, pages 723–734.

Nurulhuda Zainuddin, Ali Selamat, and Roliana Ibrahim. 2019. Hate crime on Twitter: Aspect-based sentiment analysis approach. In *Advancing Technology Industrialization Through Intelligent Software Methodologies, Tools and Techniques*, pages 284–297.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 1415–1420, Minneapolis, MN, USA.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, pages 75–86, Minneapolis, MN, USA.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, page 188202, New York, NY, USA.

Mike Zhang, Roy David, Leon Graumans, and Gerben Timmerman. 2019. Grunn2019 at SemEval-2019 task 5: Shared task on multilingual detection of hate. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 391–395, Minneapolis, Minnesota, USA.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *Proceedings of the International Conference on Multimedia Expo Workshops*, ICMEW '21, pages 1–6.

Ziming Zhou, Han Zhao, Jingjing Dong, Jun Gao, and Xiaolong Liu. 2022. DD-TIG at Constraint@ACL2022: Multimodal understanding and reasoning for role labeling of entities in hateful memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, CONSTRAINT '22, Dublin, Ireland.

Sulafa Zidani and Rachel Moran. 2021. Memes and the spread of misinformation: Establishing the importance of media literacy in the era of information disorder. *Teaching Media Quarterly*, 9(1).