Characterizing the Entities in Harmful Memes: Who is the Hero, the Villain, the Victim?

Shivam Sharma[1,4*], **Atharva Kulkarni**[2*], **Tharun Suresh**[3], **Himanshi Mathur**[3], **Preslav Nakov**[5], **Md. Shad Akhtar**[3] **and Tanmoy Chakraborty**[1]
[1]Indian Institute of Technology Delhi, India • [2]Carnegie Mellon University, USA
[3]Indraprastha Institute of Information Technology Delhi, India • [4]Wipro AI Labs, India
[5]Mohamed bin Zayed University of Artifcial Intelligence, UAE
{shivam.sharma, tanchak}@ee.iitd.ac.in, atharvak@cs.cmu.edu, preslav.nakov@mbzuai.ac.ae,
{tharun20119, himanshi18037, shad.akhtar}@iiitd.ac.in

Abstract

Memes can sway people's opinions over social media as they combine visual and textual information in an easy-to-consume manner. Since memes instantly turn viral, it becomes crucial to infer their intent and potentially associated harmfulness to take timely measures as needed. A common problem associated with meme comprehension lies in detecting the entities referenced and characterizing the role of each of these entities. Here, we aim to understand whether the meme glorifies, vilifies, or victimizes each entity it refers to. To this end, we address the task of *role identification of entities in harmful memes*, i.e., detecting who is the 'hero', the 'villain', and the 'victim' in the meme, if any. We utilize HVVMemes – a memes dataset on US Politics and Covid-19 memes, released recently as part of the CONSTRAINT@ACL-2022 shared-task. It contains memes, entities referenced, and their associated roles: hero, villain, victim, and other. We further design VECTOR (Visual-semantic role dEteCToR), a robust multi-modal framework for the task, which integrates entity-based contextual information in the multi-modal representation and compare it to several standard unimodal (text-only or image-only) or multi-modal (image+text) models. Our experimental results show that our proposed model achieves an improvement of 4% over the best baseline and 1% over the best competing stand-alone submission from the shared-task. Besides divulging an extensive experimental setup with comparative analyses, we finally highlight the challenges encountered in addressing the complex task of semantic role labeling within memes.

1 Introduction

Due to their pervasive nature, online social media platforms have emerged as a conducive medium for information exchange. Unfortunately, their democratic nature has fostered unabated dissemination

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

of hate speech (MacAvaney et al., 2019), misinformation (Wu et al., 2019), fake news (Aldwairi and Alwahedi, 2018), propaganda (Da San Martino et al., 2020), and other harmful content. Such information manifests itself in various ways, and more recently, in the form of *memes*. Though typically intended to burlesque and lampoon world events, political outlook, or daily routine, an ostensibly innocuous meme can readily become a multi-modal cause of distress with a dexterous blend of images and texts. Due to their viral nature and ability to circumvent censorship (Mina, 2014), social media instigators and hatemongers are increasingly using memes as a powerful medium for disseminating spiteful content. Therefore, investigating the dark side of memes has risen as a pertinent research problem both in industry and academia (Sharma et al., 2020a; Pramanick et al., 2021a).

Motivation. While there have been many studies that have analyzed memes through the lens of emotions (Sharma et al., 2020a), sarcasm (Kumar and Garg, 2019), hate speech (Zhou et al., 2021; Kiela et al., 2020), misinformation (Zidani and Moran, 2021), offensiveness (Suryawanshi et al., 2020),

*Equal contribution

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2149–2163
May 2-6, 2023

and harmfulness (Pramanick et al., 2021a,b), there has been much less focus on analyzing semantic roles present within memes. Understanding these roles is critical for comprehending memes and narrative framing of various social entities. Caricaturing these entities with nefarious motives can lead to misinformation propagation, social calumny, and hatred towards minority communities. In addition to the dark portrayals, memes sometimes depict the sorrowful state of certain entities, illustrate their heroism, etc. Consider the memes in Fig. 1. The meme in Fig. 1 (a) portrays Jill Stein and the Green Party as *heroes* for their feminist views. Fig. 1 (b) draws a comparison between Adolf Hitler and Donald Trump, thus portraying the latter as a *villain* for his anti-immigrant views. Fig. 1 (c), on the other hand, depicts the plight of Ola and Uber drivers, who are out of work and *victims* of the lock-downs due to the COVID-19 pandemic. Thus, through depictions of heroism, villainy, and victimization, memes act as an alluring means to spread entity-relevant information and opinions.

Challenges. Despite growing interest in analyzing memes, identifying the underlying connotations for the entities framed therein remains a challenging task (Sharma et al., 2022b). Memes are obscure due to their highly cryptic semantics, and satirical content (Sabat et al., 2019). Moreover, categorizing the entities as a hero, a villain, or a victim requires real-world, contextual, temporal, spatial, and commonsense knowledge, which makes the task highly complex and subjective even for humans. Therefore, off-the-shelf multi-modal models that stand out well on conventional visual-linguistic tasks often flounder for memes as they are presumably inept to comprehend and capture the veiled information and multi-modal nuances present in a meme (Kiela et al., 2020).

Our Contributions. In this work, we propose a powerful approach to tackle the novel task of identifying the roles (the hero, the villain, and the victim) of the entities present in memes.

We model the problem as a role identification task and report the results for several unimodal and multi-modal baselines to benchmark the task (and the dataset) and assess its feasibility. We then proffer VECTOR a vision and commonsense enriched version of DeBERTa (He et al., 2020) for the task at hand. As meme text often contains satirical content, which tends to contradict the meme image, it is necessary to consider mutual information from both modalities. Also, meme content is often stated in a non-obvious way, necessitating commonsense and world knowledge. Thus, our VECTOR attempts to infuse the relevant visual and commonsense knowledge with linguistic representations. Utilizing ConceptNet (Speer et al., 2017), we generate an entity-relevant knowledge graph to represent commonsense knowledge relevant to the meme. Moreover, we employ a distinct multi-modal information fusion strategy based on Optimal Transport. We appropriate Optimal Transport-based Kernel Embedding (OTKE) (Mialon et al., 2021) for cross-modal correspondence. This technique by Mialon et al. (2021) marries the concepts of optimal transport theory with kernel techniques to provide robust and adaptable cross-modal adaptation. Our qualitative analysis underscores the importance of vision and commonsense knowledge integration, as VECTOR outperforms several competitive baselines.

Our contributions are summarized as follows:[1,2]

1. **Bench-marking HVVMemes:** We bench-mark the HVVMemes via ten baselines with various unimodal and multi-modal systems.
2. **Multi-modal system for identifying the hero, the villain, and the victim:** We develop VECTOR (Visual-semantic role dEteCToR), a knowledge enriched multi-modal system that integrates entity-based knowledge in the multi-modal representations.
3. **Extensive evaluation:** We report sizeable gains as part of our examination of VECTOR against state-of-the-art models and shared-task submissions.
4. **Detailed Analysis:** Along with the ablation investigations, we provide detailed qualitative and quantitative analysis.

2 Related Work

Online Harmfulness. Due to the exponential rise of harmful content on various social media platforms, the research community has piqued its curiosity toward related studies. Some of them are based on online trolling (Ortiz, 2020; Cook et al., 2018), cyber-bullying (Kim et al., 2021; Kowalski et al., 2014), cyber-stalking (Abu-Ulbeh et al., 2021), and hate speech (MacAvaney et al., 2019; Zhang and Luo, 2018). Other studies character-

[1]The source code is available at https://github.com/LCS2-IIITD/VECTOR-Visual-semantic-role-dEteCToR.

[2]The dataset can be downloaded from the official shared-task page: https://codalab.lisn.upsaclay.fr/competitions/906.

ize the correlation of racial and ethnic discrimination in the online and offline world (Relia et al., 2019). Cheng et al. (2017) examined the psycho-sociological outlook of online users toward online trolling behavior analysis. Few noteworthy investigations include characterizing homophily for self-harm due to eating disorders (Chancellor et al., 2016; Wang et al., 2017) using logistic regression and snow-ball sampling, and suicide-ideation (Burnap et al., 2015; Cao et al., 2019) via linguistic, structural, affective and socio-psychological features. For a significant period, most of these studies have been dominated by text-oriented investigation while obscuring knowledge about other modalities.

Characterising Online Targets. Another research direction focuses on aspects such as relevance, stance, hate speech, sarcasm, and dialogue acts within hateful exchanges on Twitter in conventional and multi-task settings (Zainuddin et al., 2017; Gautam et al., 2020; Ousidhoum et al., 2019). Zainuddin et al. (2018) addressed it by proposing neural networks with word embeddings. In contrast, the aspect-based sentiment was studied while addressing data sparsity, classification accuracy, and sarcastic content identification (Zainuddin et al., 2019). Shvets et al. (2021) demonstrated the efficacy of a generic concept extraction module for detecting the targets of hate speech. A few other studies on characterizing targets in harmful communication (Sap et al., 2020; Mathew et al., 2021) addressed social bias and hate speech explainability for targeted protected categories. Ma et al. (2018) used a hierarchical stack bidirectional gated recurrent units to detect targets and associated sentiments. A similar objective was studied in (Mitchell et al., 2013) but was formulated as sequence tagging in low-resource settings. Silva et al. (2016) used sentence structure to capture hate speech targets on social media to address detection and prevention. Most of these studies either focused on one primary designated target or emphasized detecting the association of sentiment while ignoring the affective spectrum. As observed in the literature (Shvets et al., 2021), such approaches may not generalize well across domains.

Studies on Memes. A significant influx of memes from online fringe communities, such as Gab, Reddit, and 4chan, to mainstream platforms, such as Twitter and Instagram, resulted in a massive epidemic of intended harm (Zannettou et al., 2018). Conventional visual features alongwith multimodal associativity was explored towards detecting memes in (Sharma. and Pulabaigari., 2020; Sharma et al., 2020b). Several datasets capturing offensiveness (Suryawanshi et al., 2020), hatefulness (Kiela et al., 2020; Gomez et al., 2020), and harmfulness (Pramanick et al., 2021b), have been curated. Detecting memetic harmfulness and targeted categories are discussed in (Pramanick et al., 2021b). Commonsense knowledge (Shang et al., 2021), web entities, racial cues (Pramanick et al., 2021b; Karkkainen and Joo, 2021), and other external cues have also been explored for detecting offense, harm, and hate speech in memes. Participatory events like the Facebook Hateful Meme Challenge (Kiela et al., 2020) have laid a strong foundation for community-level initiatives for detecting hate speech in memes. As part of this challenge, several interesting approaches utilising meta information, attentive interactions, and adaptive loss are attempted in the multimodal setting (Das et al., 2020; Sandulescu, 2020; Zhou et al., 2020; Lippe et al., 2020). Most of these efforts either address the detection tasks at various levels for harmfulness; see a recent survey (Sharma et al., 2022a) or design *ensemble* techniques lacking cost-optimality. However, as per our knowledge, no stand-alone approach reliably addresses the fine-grained task of understanding the roles of specific entities referred to within memes. We intend to address these aspects by seeking a robust and generalizable multimodal framework.

Domain	# Memes	# Entity References				
		Hero	Villain	Victim	Other	Total
COVID-19	3381	200	747	407	3065	4419
US Politics	3552	288	1641	544	3242	5715

Table 1: Summary of statistics for # memes and entities referenced within them in HVVMemes (Sharma et al., 2022b). Original train/val/test split ratio of 80:10:10 (%) for memes was preserved.

3 Dataset

We employ HVVMemes, a dataset released as part of CONSTRAINT@ACL-2022. It contains English memes on two topics: 3,552 memes about COVID-19 (C) and 3,381 memes related to US Politics (P). The dataset primarily captures connotative roles: hero, villain, and victim, for different entities referenced within memes. Table 1 shows a summary

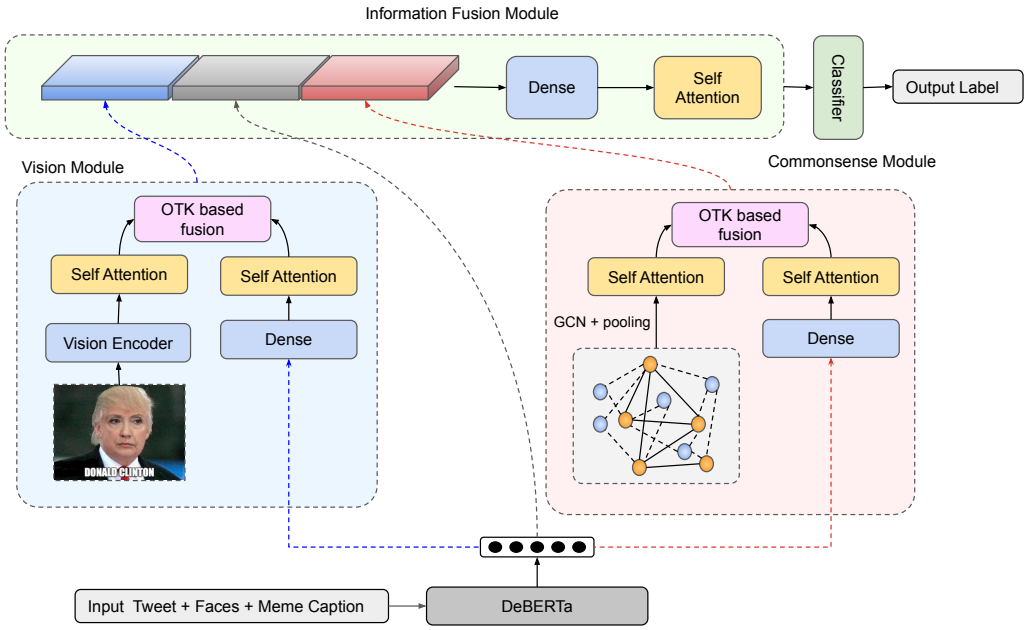


Figure 2: Schematic diagram of VECTOR. The *vision module* (left) consists of a ViT based image encoder and *commonsense module* (right) leverages GCN-based commonsense knowledge graph embedding. *Information fusion* module (top) fuses corresponding outputs toward final classification. OTK: Optimal Transport-based Kernel.

of HVVMemes.[3] In general, most of the entities referenced in the memes do not have any connotation associated (C: 3,065, P: 3,242) and are categorised as *other*. Amongst the key categories under consideration, *villain* has the most candidate references (C: 747, P: 1,641), followed by *victim* (C: 407, P: 544), and finally *hero* (C: 200, P: 288), within a total of 3,381 and 3,552 memes for COVID-19 and US Politics, respectively. This highlights the realistic trends on social media.

4 Proposed Approach

This section outlines our proposed model VECTOR (Visual-semantic role dEteCToR) and its varied components. As previously noted, role detection for memetic entities is challenging and requires real-world, contextual, and commonsense knowledge. Thus, we propose a neuro-symbolic approach that integrates commonsense-enriched modelling via graph (KG) structure into the language modelling-based architecture (Zhang et al., 2021). KG's can be considered as discrete symbolic knowledge, which we leverage along with multimodal neural modelling. As shown in Fig. 2, VECTOR houses two primary sub-modules. The Vision Module leverages cross-modal interaction between the visual-linguistic signals to grasp optimal contextual information. The Commonsense Module integrates commonsense cues through an entity-based knowledge graph. Lastly, the Information Fusion Module coalesces the information obtained via attention-based fusion. In the following subsections, we go over the specifics of each module.

Text Module: We use DeBERTa (He et al., 2020) as our backbone model as it gives the best results amongst the text-only baselines (see Table 2). Aside from the text encoded in the meme, additional verbal information can be gleaned from memes. Evidence by Blaier et al. (2021) suggests that utilizing meme captions improves hateful meme identification results. Furthermore, additional cues such as the person, the location, and the entities present in the meme are helpful for downstream tasks. Thus, we use such ancillary information along with the OCR text. For image captioning, we make use of the recently released OFA model (Wang et al., 2022). For face identification, we use the same technique as the one by Kun et al. (2022). The OCR text, the entity name, the image caption, and the identified face *labels* are concatenated and passed to the DeBERTa model. We take the final layer representation $Z \in \mathbb{R}^{l \times d}$ to fuse information from other sub-modules.

Vision Module: Meme contents often contain contradicting text and image pairs. Therefore, it

[3]For additional details, please refer to the shared-task paper (Sharma et al., 2022b).

is required to incorporate information from these modalities to understand memes. Instead of using the traditional cross-modal attention to facilitate interaction between the two modalities, we utilize an optimal transport-based kernel interaction (Mialon et al., 2021). To begin with, we use a Vision Transformer (Dosovitskiy et al., 2020) for generating the image representations $E_m \in \mathbb{R}^{l_m \times d_m}$. The text and the image representations undergo dimensionality reduction by non-linear transformation, followed by a self-attention layer (Vaswani et al., 2017) as given by equations 1. This spawns vectors $Z'_m, E'_m \in \mathbb{R}^{l \times d'}$. We concatenate these two vectors and pass them to the Optimal Transport-based Kernel Embedding (OTKE) layer to bring about cross-modal interaction. It transforms the feature vectors to a Reproducing Kernel Hilbert Space (RKHS) (Berlinet and Thomas-Agnan, 2004) followed by a weighted pooling scheme using weights determined by the transport plan between the set and a trainable reference. Such a fusion technique provides a theoretically grounded adaptive vector for the task. This yields vector $Z_m \in \mathbb{R}^{l \times d'}$, given by equation 2 below:

$$Z'_m = S\left(\frac{ZZ^T}{\sqrt{d}}\right)Z \quad E'_m = S\left(\frac{E_m E_m^T}{\sqrt{d_m}}\right)E_m \quad (1)$$

$$Z_m = OTKE([Z'_m : E'_m]) \quad (2)$$

Commonsense Module: Due to their cryptic nature, identifying the intent of a meme is challenging. Their satirical and non-obvious way of conveying a message often requires commonsense comprehension. Thus, in our commonsense module, we generate a commonsense knowledge graph based on the entities in the meme. Similarly to (Shang et al., 2021), we extract all the nouns and noun phrases in the meme OCR and the meme caption. We extract commonsense relation pairs having confidence > 2 from ConceptNet (Speer et al., 2017) for all the noun chunks from memes. Noun chunks, commonsense entities, and the meme entity in question form the nodes of the commonsense graph. Each noun chunk from the OCR text is connected. The same applies to the noun chunks in the meme caption. A special aggregator token connects OCR and caption-based nodes. Thus, for each entity $e_{ij} \in E_i = \{e_{i1}, e_{i2}, \ldots, e_{in}\}$, we have a commonsense graph $G_{ij} = (V_{ij}, E_{ij})$, where V_{ij} and E_{ij} are the nodes and the edges of the graph.

The graph having edges between various entities, noun chunks, and nouns, being undirected, represents a generic "connectivity". Therefore, an edge between nodes A and B indicates that they have some association. By doing this, we wanted to capture the commonsense-based 'proximity' of different entities/concepts within the vectorized space for common sense concepts. V_{ij} represents a set node (or vertices) constituting a commonsense graph corresponding to each entity.

For each node in the commonsense graph, we generate an embedding of size d_g using the last layer representations from DeBERTa. In order to facilitate the interaction between the nodes, the commonsense graph goes through two rounds of graph convolutions (Kipf and Welling, 2017) followed by a max-pooling operation to spawn an aggregated graph embedding $E_g \in \mathbb{R}^{d'}$. Similarly to the vision module, the textual representations Z and the commonsense graph representation E_g undergo non-linear transformation and dimensionality reduction followed by self-attention (Vaswani et al., 2017). This generates contextual vectors $Z'_g, E'_g \in \mathbb{R}^{l \times d'}$, respectively. Finally, the commonsense knowledge is infused in the language representations using OKTE, generating the final vector $Z_g \in \mathbb{R}^{l \times d'}$.

Information Fusion Module: Each of the modules mentioned above integrates salient information in the language representations. The information fusion module aggregates the knowledge obtained from all other modules using attention-based mutual interaction (Vaswani et al., 2017). Concretely, we concatenate the vectors of Z, Z_m, and Z_g and pass them through a round of dimensionality reduction and non-linear transformations. Then, we use a self-attention mechanism so that the information obtained from each component interacts with one another. The final generated vector $Z'_c \in \mathbb{R}^{l \times d}$ is passed to a classifier with a softmax activation to predict the final labels.

5 Baselines

Unimodal Systems: We use a variety of text-based and image-based models as our baseline systems. Starting with the text baselines, we use **BERT** (Devlin et al., 2019), and variants thereof such as **DistilBERT** (Sanh et al., 2019), **RoBERTa** (Liu et al., 2019), **XLNet** (Yang et al., 2019), and **DeBERTa** (He et al., 2020). For the image-based baselines, we start with a representation based on **ResNet-50** (He et al., 2016), followed by **Vision Transformer (ViT)** (Dosovitskiy et al., 2020) and **SWIN** (Liu et al., 2021), which is hierarchical ver-

	Model Details	Hero			Villain			Victim			Other			Macro-F1			Acc.
		Prec	Rec	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	
Text-only	Dis. BERT	.154	.115	.132	.431	.400	.415	.292	.246	.267	.856	.881	.868	.433	.411	.420	.766
	BERT	.250	.115	.158	.484	.446	.464	.358	.254	.297	.864	.905	.884	.489	.430	.451	.791
	RoBERTa	.243	.173	.202	.473	.400	.433	.346	.386	.365	.866	.891	.879	.482	.463	.470	.782
	XLNet	.172	.096	.123	.491	.386	.432	.383	.272	.318	.852	.910	.880	.475	.416	.438	.788
	DeBERTa	.250	.250	.250	.469	**.591**	.523	.395	.395	.395	**.889**	.847	.868	.501	**.521**	.509	.776
	DeBERTa(l)	.268	.212	.237	**.604**	.440	.509	.543	**.500**	.521	.870	.922	**.895**	.518	.571	.540	**.818**
Vision-only	ResNet	0	0	0	.467	.303	.367	.667	.053	.097	.827	**.948**	.883	.490	.326	.337	.793
	ConvNeXT	0	0	0	.398	.329	.360	.333	.009	.017	.828	.924	.873	.390	.315	.313	.776
	ViT	0	0	0	.436	.380	.406	.250	.009	.016	.836	.926	.878	.380	.329	.325	.785
	BEiT	0	0	0	.413	.420	.416	.500	.009	.017	.838	.907	.871	.438	.334	.326	.775
	SWIN	0	0	0	.393	.326	.356	.231	.026	.047	.828	.919	.871	.363	.318	.319	.772
Multimodal	CLIP	0	0	0	.478	.311	.378	**.691**	.071	.104	.845	.951	.886	.393	.335	.342	.786
	ViLBERT	.078	.034	.045	.409	.513	.469	.398	.276	.321	.856	.843	.849	.428	.408	.421	.761
	V-BERT	.143	.077	.100	.466	.560	.508	.452	.333	.384	.886	.879	.882	.487	.462	.469	.790
	MMTrans.	0	0	0	.516	.477	.496	.447	.303	.361	.814	.878	.845	.437	.392	.405	.747
	MMBT	.103	.058	.074	.446	.537	.487	.414	.298	.347	.881	.872	.877	.458	.438	.447	.780
	VECTOR	**.444**	**.385**	**.412**	.553	.534	**.544**	.505	.456	**.479**	.883	.897	.890	**.568**	**.596**	**.581**	.813
	$\Delta_{(\text{VECTOR}-\text{V-BERT})}$	.301↑	.308↑	.312↑	.087↑	.026↓	.036↑	.053↑	.123↑	.095↑	.003↓	.018↑	.008↑	.081↑	.134↑	.112↑	.023↑

Table 2: Benchmarking results (0's indicate no correct prediction). $\Delta_{(X-Y)}$: Performance difference between models X and Y, # ↑ and # ↓: Absolute increment and decrement respectively.

sion of ViT that uses shifted windows. We also include the recently proposed **ConvNeXT** (Liu et al., 2022) and **BEiT** (Bao et al., 2021), which uses self-supervised pre-training of Vision Transformers.

Multi-modal Systems: We use various variants of multi-modal pre-trained systems from the MMF Framework. **MMF Transformer** is a library Transformer model that uses visual and language tokens with self-attention. **MMBT:** multi-modal Bi-transformer (Kiela et al., 2019) captures the two modalities' intra-modal and inter-modal dynamics. **ViLBERT** is a vision and language BERT (Lu et al., 2019), a strong model with the task-agnostic joint representation of images and text. **ViLBERT CC** is pre-trained on conceptual captions (Sharma et al., 2018) based pretext task. **Visual BERT** (Li et al., 2019), also pre-trained using MS COCO (Lin et al., 2014), implicitly aligns the input text and regions in the input image using self-attention. CLIP (Radford et al., 2021) leverages image–text contrastive pretraining.

6 Experiments

Experimental Details We present benchmarking results in Table 2, comparison with the shared-task submissions (c.f. Table 4) averaged over five independent runs, while for the ablation study in Table 3, we compare the *best* check-points for different VECTOR component-wise evaluations to assess the performance bounds, on the *diversified* test set. We use precision, recall, and F1 for individual classes, and macro-averaged for the overall assessment.[4]

[4]See Appendix A for additional experimental details.

Comparative Analysis. ▷ **Unimodal Models:** Despite the evident glorification cues within the *Hero* references in memes, large pre-trained models are observed to depict limitations. This can be observed from Table 2, wherein significantly low performance is observed for strong models like DistilBERT, BERT, and XLNet, with F1 scores of 0.132, 0.158, and 0.123, respectively. On the other hand, RoBERTa enhances the class-specific F1 score by about 0.05 absolute points, which suggests the efficacy the exhaustive hyperparameter tuning can induce. Finally, DeBERTa-based unimodal systems yield the highest optimal F1 scores among all unimodal models evaluated. The scores observed are 0.250 and 0.237 for the base and the large variants, respectively.

For the *Villain* category, BERT can be observed (c.f. Table 2) to yield relatively better performance as against Hero detection, with an F1 score of 0.464. This is in contrast to the sub-par performances by DistilBERT, RoBERTa, and XLNet, yielding 0.415, 0.433, and 0.432 F1 scores, respectively. As with the Hero detection, DeBERTa-base and large models outperform the other models for both text and vision modalities.

Interestingly, none of the text-only models, except for DeBERTa large, could beat the F1 score of 0.395 for DeBERTa-base. The lower performance is primarily due to inadequate categorical representation within the dataset. Moreover, the inherent complexity in distinguishing villains from victims confuses the model. The modelling efficacy solicited for such a scenario is suggested by over 0.10 absolute point enhancement by DeBERTa-large, which has 4X more backbone parameters

Model Details	HER	VIL	VIC	OTH	F1	Acc.
Simple early-fusion (DeBERTa + ViT)	0.31	0.55	0.50	0.88	0.56	0.79
(a). + Meme (image) caption	0.32	0.52	0.48	0.89	0.55	0.80
(b). + Face labels	0.26	0.56	0.51	0.89	0.56	0.82
(c). [(a) + (b)] + Commonsense KG	0.36	0.54	0.45	0.89	0.56	0.81
(d). [(a) + (b) + (c)] + CAT	0.28	0.53	0.49	0.89	0.55	0.81
(e). [(a) + (b) + (c)] + OTK (VECTOR)	**0.38**	**0.57**	**0.53**	**0.90**	**0.60**	**0.83**

Table 3: Ablation study: Comparing VECTOR and its sub-modules over *best* model results. CAT: Fusion via concatenation, OTK: Fusion via Optimal Transport Kernel, KG: Knowledge Graph.

Rank	System	Prec.	Rec.	F1
-	VECTOR	0.568	0.596	0.581
1	Logically (Kun et al., 2022)	0.544	0.610	0.571
2	c1pher (Singh et al., 2022)	0.527	0.581	0.547
3	smontariol (Montariol et al., 2022)	0.580	0.450	0.485
4	zhouziming (Zhou et al., 2022)	0.480	0.450	0.462
5	IIITDWD (Fharook, 2022)	0.256	0.238	0.239
6	rabindra.nath (Nandi et al., 2022)	0.253	0.253	0.237

Table 4: Comparison to the results on the CONSTRAINT@ACL-2022 shared-task on HVVMemes.

than DeBERTa-base.

Other category, having the majority representation with over 6K unique references within memes, projects the highest F1 scores with an average of 0.879. Finally, the overall Macro-F1 scores reflect the category-wise trend observed, with DeBERTa large leading with an F1 score of 0.540. DeBERTa-base follows it with 0.509 and the rest with an average F1 score of 0.444.

For intuitive reasons, the visual modality, not indicative of complex role semantics, yields poor performance compared to text-only models. As a result, none of the image-only models (ConvNext, ViTBEiT, and SWIN) are observed to make any headway in correctly detecting a Hero reference in memes (c.f. Table 2). At the same time, except for the Villain category, a simple ResNet-based model outperforms, albeit with fine margins, the rest of the models within the categories Victim and Other, with F1 scores of 0.097 and 0.883, respectively. This highlights the efficacy of global image representations against tokenized (or patched) ones for factoring visual features, especially where there is not much visual-linguistic grounding to be leveraged. On average, the image-only models can project a paltry F1 score of 0.324.

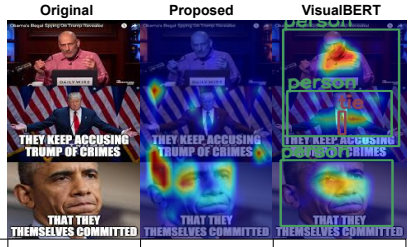
▷ **Multi-modal Models.** Several state-of-the-art multimodal systems, on average, are observed to yield a Macro-F1 score of 0.416, which lies between the text-only (0.444) and the image-only (0.324) models. Besides the category-wise performance trend being similar to the one observed for unimodal models, multimodal systems like VisualBERT and MMBT yield the highest Macro-F1 scores of 0.468 and 0.447, respectively. This is likely due to the joint attentive modelling, multimodal pre-training using standard datasets like MS COCO, and fine-tuning adopted by these models. Other competitive models like CLIP, ViLBERT, and MM Transformer achieve Macro-F1 scores of up to 0.342, 0.421, and 0.405, respectively. This either suggests that the image component induces additional noise within the models or that the existing multimodal systems do not effectively capture the complex pragmatics that semantic role-labeling in memes solicits. The former is less likely, as intuitively, memetic visuals do provide minor yet decisive semantics toward holistic assimilation of the meme's message. Our proposed approach VECTOR, is observed to address the required cross-modal association by achieving impressive F1 scores across different roles and a 0.581 Macro-F1 score, which induces an enhancement of almost 4% and 12 % over DeBERTa large (unimodal best) and VisualBERT (multimodal best), respectively.

Ablations Results. Table 3 depicts ablation results, wherein the basic early-fusion setup involving DeBERTa and ViT performs well for *villain* category with 0.55 F1 score. Adding meme-image captions enhances the *hero* predictions and overall accuracy marginally by 1% each. Face labels, although, significantly enhance the prediction of *villain* and *victim*, suggesting lexical and semantic utility via *face labels*. The overall performance remains unchanged due to compromised *hero* predictions. Besides yielding balanced scores, adding a commonsense module elevates hero prediction distinctly to 0.36, effectively indicating its utility, especially for the under-represented role category. Finally, VECTOR with OTK-based embedding yields optimal cross-modal correspondence, as observed from the best performances across roles and metrics evaluated, showcasing the constituting feature's utility towards addressing the given task. Although *not* averaged over multiple runs, the *best* model check-point for VECTOR is observed to yield an impressive overall macro-F1 score of 0.60, which is the best score observed across experiments, suggesting an upper-bound for VECTOR's performance.

Comparison to Previous Work. Table 4 showcases the best-performing systems from

	Original	Proposed	VisualBERT
Hero	–	–	–
Villain	[jeffrey epstein, donald trump]	[donald trump]	[jeffrey epstein, donald trump]
Victim	[13 year old girl]	[13 year old girl]	[13 year old girl]
Other	–	[jeffrey epstein]	–

Figure 3: Error analysis for VECTOR and VisualBERT.



	Original	Proposed	VisualBERT
Hero	–	–	–
Villain	[barack obama]	[barack obama]	[barack obama, donald trump, daily wire]
Victim	[donald trump]	[donald trump]	–
Other	[daily wire]	[daily wire]	–

Figure 4: Interpretability analysis for VECTOR; proposed model and VisualBERT; best multimodal baseline.

CONSTRAINT@ACL-2022's shared-task. Since all of the shared-task submissions used ensemble techniques of multiple models, we present their best individual model results for a fair comparison with our proposed model: VECTOR. The macro-F1 score varies by 0.334 points across the participants, emphasizing model selection. We can draw parallels between Table 2 and Table 4 with Logically, c1pher, smontariol, and zhouziming using DeBERTa, RoBERTa, VisualBERT, and VisualBERT, respectively. They exhibit marginal improvements using loss-weighting techniques and additional classification layers. Our model VECTOR performs more consistently, especially with the class-wise scores of *Hero* (c.f. Table 2), establishing its efficacy across the category types and limited categorical data representation. Besides class-wise consistency, VECTOR outperforms other shared-task submissions as a stand-alone system. Also, despite being relatively complex, VECTOR produces consistent class-wise and better overall performance, suggesting marginal yet robust modelling capacity facilitated by VECTOR's vision + common-sense module's interaction with the textual signals.

Statistical Significance. We performed a bootstrapping significance test w.r.t the proposed model (VECTOR), and the previous best solution by team Logically (Kun et al., 2022) via random sampling with replacement strategy, with N=1000 over 1000 simulations, and observed a 'p-value' of 0.0410. This indicates a subtle yet encouraging confidence margin in the model predictions. This could be likely due to better predictions across the four categories, including 'hero', which most of the other models compared are empirically observed to struggle at. These aspects corroborate the semantic-role label-agnostic characteristics of the proposed model.

Error Analysis. The example depicted in Fig. 3 insinuates 'donald trump' and 'jeffrey Epstein's as *villains* while *victimizes* a '13 year old girl'. Visually, there isn't much to consider towards adjudicating the former two entities as villains, except the expressions of 'jeffrey Epstein's exuding somewhat *sinister* looks. Whereas *vilifying* connotation is implied primarily by the embedded text. Now, although VECTOR predicts the roles of 'donald trump' and '13 year old girl' correctly as *villain* and *victim*, respectively, it fails to detect 'jeffrey epstein as a *villain* and categorizes it as an *other*. This example highlights the limitations of VECTOR in terms of its inherent modality-specific biases. A ViT-based image encoder, due to its disparate patch-wise processing, and self-attention across the input patches, leads to noisy visual attention. On the other hand, VisualBert-based predictions replete with pre-trained common-sense semantics are better positioned for this case to capture the required semantic indicators, a portion of *epstein's* facial expression in this case. Also, as the dataset is well-stocked with examples where 'Donald trump' is *vilified*, both models being compared assign the role of *villain* to 'donald trump'.

7 Interpretability

The predictive capacity of VECTOR can be assessed by comparing its interpretability with that of VisualBERT, the best-performing baseline. We investigate a reasonably complex example depicted in Fig 4, for which VECTOR accurately predicts all the entity roles, but VisualBERT wrongly predicts all of them, as *villains*. Attention map for VECTOR reveals that it primarily attends to the face of 'barack obama', while leveraging other contextual cues, likely via the common-sense and visual description-

based signals, leading to a correct assignment of roles. On the other hand, VisualBERT, having a Faster R-CNN-based image encoder, primarily pre-trained for the common object-detection task, fails to uniquely attend to different semantically referred entities within the meme image and additionally adds noisy features related to objects like a *tie*, without any additional contextual knowledge. This validates VECTOR's superior discriminatory capacity and interpretability over VisualBERT.

8 Generalizability

Since, the test set consists of all unique memes, there is a higher chance that there are many entities as part of it that are not seen during the training stage. Table 2 attests to the VECTOR's potential to succeed across different roles. This suggests that VECTOR can adapt to the domain-specific nuances offered by the variety and complexity associated with visual-semantic roles within memes. Through the results observed in this work, we attempt to highlight that existing *standard* approaches either have inconsistent performances across the roles: hero, villain, victim, and other, or yield low scores for particular role categories (c.f. Table 2), which corroborates their limitations. On the other hand, VECTOR yields a balanced performance and generalizes reasonably well for the least represented class in the dataset (*hero*).

9 Conclusion

This paper addresses a recently proposed task of identifying the roles of entities in harmful memes and discusses its challenges. We further presented numerous unimodal and multi-modal baselines to benchmark HVVMemes. Moreover, we proffer VECTOR, a contextual knowledge-enriched multi-modal framework that bolsters the multi-modal representations with entity-based external knowledge using a cross-modal attention scheme. VECTOR shows noteworthy improvements over the baselines, thus justifying contextual knowledge inclusion. As for future investigations, we plan to conceive a more symbolic system with graph-based entity linking, commonsense knowledge, and visual concepts.

10 Limitations

As noted in the discussion dedicated to *Error Analysis*, several entities tend to dominate specific roles within the dataset due to the realistic representation of the harmful referencing in memes. This not only biases the model against their generalizability but also poses challenges towards modelling entity-independent role detection hypotheses for a diverse set of entities. This especially calls for building models regularized to address such biases and more participatory initiatives toward curating better and more large-scale datasets.

11 Ethics and Broader Impact

Reproducibility. We present detailed hyper-parameter configurations in Appendix A.

User Privacy. The information depicted/used does not include any personal information. Copyright aspects are attributed to the dataset source.

Biases. As per the authors, any biases found in the dataset are unintentional (Sharma et al., 2022b), and by conducting the study on this dataset we do not intend to cause harm to any group or individual. We acknowledge that detecting harmfulness can be subjective, and thus it is inevitable that there would be biases in gold-labelled data or in the label distribution. This is addressed by the dataset curators by using general keywords about US Politics, and also by following a well-defined schema, which sets explicit definitions for annotation.

Misuse Potential. Our approach can be potentially used for ill-intended purposes, such as biased targeting of individuals/communities/organizations, etc. that may or may not be related to demographics and other information within the text. Intervention with human moderation would be required to ensure that this does not occur.

Intended Use. We make use of the existing dataset in our work in line with the intended usage prescribed by its creators and solely for research purposes. This applies in its entirety to its further usage as well. We do not claim any rights to the dataset used or any part thereof. We believe that it represents a useful resource when used appropriately.

Environmental Impact. Finally, large-scale models require a lot of computations, which contribute to global warming (Strubell et al., 2019). However, in our case, we do not train such models from scratch; rather, we fine-tune them on a relatively small dataset.

Acknowledgements

The work was supported by Wipro research grant.

References

Waheeb Abu-Ulbeh, Maryam Altalhi, Laith Abualigah, Abdulwahab Almazroi, Putra Sumari, and Amir Gandomi. 2021. Cyberstalking victimization model using criminological theory: A systematic literature review, taxonomies, applications, tools, and validations. *Electronics*, 10:1670.

Monther Aldwairi and Ali Alwahedi. 2018. Detecting fake news in social media networks. *Procedia Computer Science*, 141:215–222.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*.

Alain Berlinet and Christine Thomas-Agnan. 2004. *Reproducing Kernel Hilbert Space in Probability and Statistics*.

Efrat Blaier, Itzik Malkiel, and Lior Wolf. 2021. Caption enriched samples for improving hateful memes detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9350–9358, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext &; Social Media*, HT '15, page 75–84, New York, NY, USA.

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. 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 1718–1728, Hong Kong, China.

Stevie Chancellor, Zhiyuan (Jerry) Lin, and Munmun De Choudhury. 2016. "this post will just get taken down": Characterizing removed pro-eating disorder social media content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1157–1162, New York, NY, USA.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1217–1230, New York, NY, USA.

Christine Cook, Juliette Schaafsma, and Marjolijn Antheunis. 2018. Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media & Society*, 20(9):3323–3340.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. 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).

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. In *Proceedings of the NeurIPS Workshop on Visually Grounded Interaction and Language*, ViGIL '19, Vancouver, Canada.

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*.

Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J. Wisniewski, and Munmun De Choudhury. 2021. A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings ACM Hum. Comput. Interact.*, 5(CSCW2):1–34.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15, San Diego, California, USA.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Robin Kowalski, Gary Giumetti, Amber Schroeder, and Micah Lattanner. 2014. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin*, 140.

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*, pages 24–34, Dublin, Ireland. Association for Computational Linguistics.

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*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, ECCV '14, pages 740–755, Zurich, Switzerland.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv:2012.12871*.

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 preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Conference on Neural Information Processing Systems*, NeurIPS '19, pages 13–23, Vancouver, Canada.

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.

Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, and Julien Mairal. 2021. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *International Conference on Learning Representations*.

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.

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.

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.

Stephanie M. Ortiz. 2020. Trolling as a collective form of harassment: An inductive study of how online users understand trolling. *Social Media + Society*, 6(2):2056305120928512.

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.

Kunal Relia, Zhengyi Li, Stephanie H. Cook, and Rumi Chunara. 2019. Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 u.s. cities. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):417–427.

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*.

Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble. *arXiv:2012.13235*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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.

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. 2020a. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 759–773.

Chhavi Sharma. and Viswanath Pulabaigari. 2020. A curious case of meme detection: An investigative study. In *WEBIST*, pages 327–338.

Chhavi Sharma, Viswanath Pulabaigari, and Amitava Das. 2020b. Meme vs. non-meme classification using visuo-linguistic association. pages 353–360.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pages 2556–2565, Melbourne, Australia.

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

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md Shad Akhtar, and Tanmoy Chakraborty. 2022b. Findings of the constraint 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 1–11.

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.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):687–690.

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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 3645–3650, Florence, Italy.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052.

Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 91–100, New York, NY, USA.

Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor. Newsl.*, 21(2):80–90.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

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 188–202, New York, NY, USA.

Jing Zhang, Bo Chen, Lingxi Zhang, Xirui Ke, and Haipeng Ding. 2021. Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *AI Open*, 2:14–35.

Z. Zhang and L. Luo. 2018. Hate speech detection: a solved problem? the challenging case of long tail on twitter. *Semantic Web*.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pages 354–367. Springer.

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).

Modality	Model	BS	#Epochs	LR	V-Enc	T-Enc	#Param
UM	BERT	8	15	1e-5	-	bert	25M
	DistilBERT	8	15	1e-5	-	distilbert-base	66M
	XLNet	8	15	1e-5	-	xlnet	116M
	RoBERTa	8	15	1e-5	-	roberta-base	123M
	DeBERTa	8	15	1e-5	-	deberta-base	86M
	DeBERTa-Large	8	15	1e-5	-	deberta-large	304M
	ResNet	8	15	1e-5	resnet	-	25M
	ConvNeXT	8	15	1e-5	convnet	-	50M
	ViT	8	15	1e-5	vit	-	86M
	SWIN	8	15	1e-5	swin	-	88M
	BEiT	8	15	1e-51	beit	-	71M
MM	MMFT	16	20	0.001	ResNet-152	bert	170M
	CLIP	16	20	0.0001	ViT	clip	149M
	MMBT	16	20	0.0001	ResNet-152	bert	169M
	ViLBERT*	16	10	0.0001	Faster RCNN	bert	112M
	V-BERT*	16	10	0.0001	Faster RCNN	bert	247M
	VECTOR	8	15	1e-5	vit	deberta-large	123M

Table 5: Hyperparameters summary. [BS→Batch Size; LR→Learning Rate; V/T-Enc→Vision/Text-Encoder; vit→vit-base-patch16-224-in21k; bert:→bert-base-uncased; xlnet→xlnet-base-uncased; resnet→resnet50].

A Implementation Details and Hyperparameter Values

We train all the models using PyTorch on an actively dedicated NVIDIA Tesla V100 GPU, with 32 GB dedicated memory, CUDA-11.2, and cuDNN-8.1.1 installed. For the unimodal models, we import all the pre-trained weights from the TORCHVISION.MODELS,[5] a sub-package of the PyTorch framework. We randomly initialise the remaining weights. Sharma et al. (2022b) re-annotate the HarMeme dataset (Pramanick et al., 2021b) by collecting annotator responses for different roles different entities within memes take. The re-annotated memes may or may not have harmful implications, contrary to the distinction modelled as part of original curation. However, they portray various entities within different contexts implying glorification, vilification, and victimisation. For most of our experiments, we use Adam optimiser (Kingma and Ba, 2015) with a learning rate of $1e^{-4}$ or $1e^{-5}$, a weight decay of $1e^{-5}$ and a Cross-Entropy (CE) loss as the objective function. We optimized our models to obtain hyper-parameter settings (c.f. Table 5) and early-stop to preserve our best state convergence. On average, it took approx. 2:30 hours to train a typical multi-modal neural model on a dedicated GPU system.

B Additional details of HVVMemes

B.1 Edge Case

Most memes are intended to project harmless mockery toward various sections of society. Such memes do not imply heroes, villains, or victims; instead, they disseminate harmless humour and trivial opinions. Therefore, Sharma et al. (2022b) do not presume any implications regarding these connotations and categorise them as 'other', a fourth *neutral* category, unless expressed otherwise in the meme. A depiction of such a scenario can be observed in Fig. 5. In this meme, it is unclear if Donald Trump is being vilified for being reluctant to use a mask or if the meme expresses a benign attempt at mocking his physical appearance with sarcasm. Additionally, since no background information would suffice to facilitate its complete assimilation, it is categorised as *other*.

[5] http://pytorch.org/docs/stable/torchvision/models.html

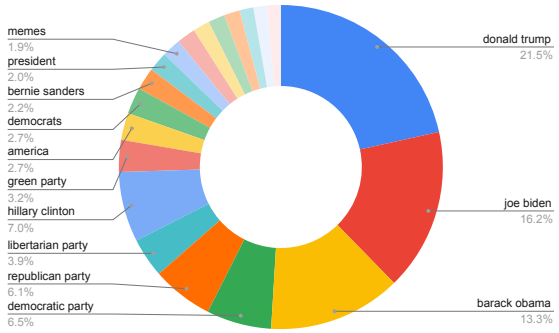


Figure 6: Entity distribution - US Politics.

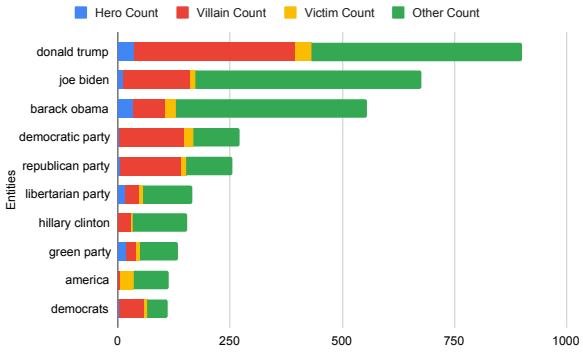


Figure 7: Entity-role distribution - US Politics.

B.2 Analysing Different Entities and Roles

Figure 5: Edge case

Regardless of the connotations and domains in which various entities are referred, only a handful of entities/topics dominate pivotal referencing in memes. The entities in *COVID-19* memes predominantly focus on *Coronavirus*, *Donald Trump*, *mask*, *COVID-19*, and *work from home*. The entities of *Donald Trump*, *Joe Biden*, *Barack Obama*,

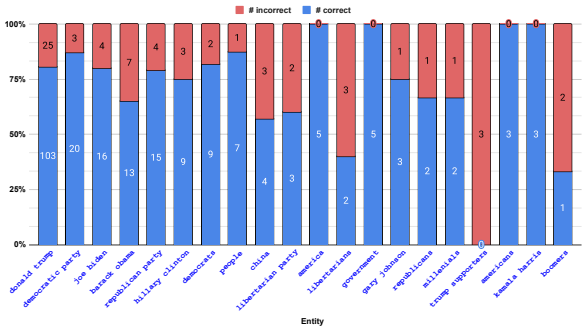


Figure 8: Entity role prediction summary for HVVMemes's (combined) test set. X-axis: Top 20 entities referenced; Y-axis: True positive rate (Recall). Values depicted at the center of each portion for each bar depicts corresponding total counts.

Democratic Party, and *Republican Party* crowd the US Politics memes, as shown in Fig. 6. Such trends highlight the key figures and real-world topics that dominated memetic communication on social media during the period this dataset was compiled, which coincided with the onset of the global COVID-19 pandemic and contemporary US Politics. In COVID-19 memes, we observe that entities like *Donald Trump* and *China* are referenced almost equally within the vilifying and neutral contexts. In contrast, entities like *Corona beer*, *introverts*, and *Tom Hanks*, are invariably referenced in neutral contexts through irony, satire, or benign humour. Similar trends are observed in US Politics memes (Fig. 7), wherein entities like *Donald Trump*, *Democratic Party*, *Republican Party*, and *Democrats* observe almost equivalent referencing as *villain* and *others*. Interestingly, as shown in Fig. 7, US Politics related memes depict a higher propensity towards vilification than that of COVID-19, as most of the prominent entities encountered are vilified at least once in the dataset.

C Entity Role Prediction Analysis.

Careful analysis of the role predictions for various entities elicits the correlation between the role-wise distribution and the test set predictions for different entities. This correlation can be observed from Fig. 8, wherein entities like *Donald Trump, Democratic Party, Joe Biden, Republican Party, Democrats*, etc., which have a *true positive rate* (recall) of atleast 75%, are specifically the ones that have relatively balanced role-wise distribution for the roles of *villain* and *other* within HVVMemes, (see Fig. 4 (main content). Interestingly, a lower but role-wise balanced representation within HVVMemes, does not appear to deter VECTOR from yielding an impressive recall of 90% for an entity such as *Democrats*, which have a total of approximately 125 samples in the training set. On the other hand, for entities like *Barack Obama, China, Libertarian Party, Libertarian*, etc., the memetic portrayal is significantly skewed-in as an *other*, suggesting distinct role-wise imbalance within memes that VECTOR failed to accommodate, highlighting its limitations. Further, there are entities like *America, Government, Americans, Kamala Harris*, etc., that register a 100% recall. Role-wise distributions for such cases highlight a distinct majority of neutral connotations within both training and test splits via *other* category, essentially suggesting possible biases within the role-prediction modelling setup. Entity *Trump Supporters* is also observed to be present within the training set via another similar referencing *Donald Trump Supporters*, which has a 1:5 ratio of villain/victim:other. This is complemented by the predominant yet balanced referencing of an independent entity *Donald Trump*. This could lead to the model being confused for such entities as is depicted in Fig. 8.