Detecting the Role of an Entity in Harmful Memes: Techniques and Their Limitations

Rabindra Nath Nandi[1], **Firoj Alam**[2], **Preslav Nakov**[2]
[1]BJIT Limited, Dhaka, Bangladesh
[2]Qatar Computing Research Institute, HBKU, Doha, Qatar
rabindra.nath@bjitgroup.com, {falam,pnakov}@hbku.edu.qa

Abstract

Harmful or abusive online content has been increasing over time, raising concerns for social media platforms, government agencies, and policymakers. Such harmful or abusive content can have major negative impact on society, e.g., cyberbullying can lead to suicides, rumors about COVID-19 can cause vaccine hesitance, promotion of fake cures for COVID-19 can cause health harms and deaths. The content that is posted and shared online can be textual, visual, or a combination of both, e.g., in a meme. Here, we describe our experiments in detecting the roles of the entities (hero, villain, victim) in harmful memes, which is part of the CONSTRAINT-2022 shared task, as well as our system for the task. We further provide a comparative analysis of different experimental settings (i.e., unimodal, multimodal, attention, and augmentation). For reproducibility, we make our experimental code publicly available.[1]

1 Introduction

Social media have become one of the main communication channels for sharing information online. Unfortunately, they have been abused by malicious actors to promote their agenda using manipulative content, thus continuously plaguing political events, and the public debate, e.g., regarding the ongoing COVID-19 infodemic (Alam et al., 2021d; Nakov et al., 2022). Such type of content includes harm and hostility (Brooke, 2019; Joksimovic et al., 2019), hate speech (Fortuna and Nunes, 2018), offensive language (Zampieri et al., 2019; Rosenthal et al., 2021), abusive language (Mubarak et al., 2017), propaganda (Da San Martino et al., 2019, 2020), cyberbullying (Van Hee et al., 2015), cyber-aggression (Kumar et al., 2018), and other kinds of harmful content (Pramanick et al., 2021; Sharma et al., 2022b).

The propagation of such content is often done by coordinated groups (Hristakieva et al., 2022) using automated tools and targeting specific individuals, communities, and companies. There have been many research efforts to develop automated tools to detect such kind of content. Several recent surveys have highlighted these aspects, which include fake news (Zhou and Zafarani, 2020), misinformation and disinformation (Alam et al., 2021c; Nakov et al., 2021; Hardalov et al., 2022), rumours (Bondielli and Marcelloni, 2019), propaganda (Da San Martino et al., 2020), hate speech (Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017), cyberbullying (Haidar et al., 2016), offensive (Husain and Uzuner, 2021) and harmful content (Sharma et al., 2022b).

The content shared on social media comes in different forms: textual, visual, or audio-visual. Among other social media content, recently, *internet memes* became popular. Memes are defined as "a group of digital items sharing common characteristics of content, form, or stance, which were created by associating them and were circulated, imitated, or transformed via the Internet by many users" (Shifman, 2013). Memes typically consist of images containing some text (Shifman, 2013; Suryawanshi et al., 2020a,b). They are often shared for the purpose of having fun. However, memes can also be created and shared with bad intentions. This includes attacks on people based on characteristics such as ethnicity, race, sex, gender identity, disability, disease, nationality, and immigration status (Zannettou et al., 2018; Kiela et al., 2020). There has been research effort to develop computational methods to detect such memes, such as detecting hateful memes (Kiela et al., 2020), propaganda (Dimitrov et al., 2021a), offensiveness (Suryawanshi et al., 2020a), sexist memes (Fersini et al., 2019), troll memes (Suryawanshi and Chakravarthi, 2021), and generally harmful memes (Pramanick et al., 2021; Sharma et al., 2022a).

[1]https://github.com/robi56/harmful_memes_block_fusion

Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations, pages 43 - 54
May 27, 2022 ©2022 Association for Computational Linguistics

Harmful memes often target individuals, organizations, or social entities. Pramanick et al. (2021) developed a dataset where the annotation consists of (*i*) whether a meme is harmful or not, and (*ii*) whether it targets an individual, an organization, a community, or society. The CONSTRAINT-2022 shared task follows a similar line of research (Sharma et al., 2022c). The entities in a meme are first identified and then the task asks participants to predict which entities are glorified, vilified, or victimized in the meme. The task is formulated as *"Given a meme and an entity, determine the role of the entity in the meme: hero vs. villain vs. victim vs. other."* More details are given in Section 3.

Memes are multimodal in nature, but the textual and the visual content in a meme are sometimes unrelated, which can make them hard to analyze for traditional multimodal approaches. Moreover, context (e.g., where the meme was posted) plays an important role for understanding its content. Another important factor is that since the text in the meme is overlaid on top of the image, the text needs to be extracted using OCR, which can result in errors that require additional manual post-editing (Dimitrov et al., 2021a).

Here, we address a task about entity role labeling for harmful memes based on the dataset released in the CONSTRAINT-2022 shared task; see the task overview paper for more detail (Sharma et al., 2022c). This task is different from traditional semantic role labeling in NLP (Palmer et al., 2010), where understanding *who* did *what* to *whom*, *when*, *where*, and *why* is typically addressed as a sequence labeling problem (He et al., 2017). Recently, this has also been studied for visual content (Sadhu et al., 2021), i.e., situation recognition (Yatskar et al., 2016; Pratt et al., 2020), visual semantic role labeling (Gupta and Malik, 2015; Silberer and Pinkal, 2018; Li et al., 2020), and human-object interaction (Chao et al., 2015, 2018).

To address the entity role labeling for a potentially harmful meme, we investigate textual, visual, and multimodal content using different pretrained models such as BERT (Devlin et al., 2019), VGG16 (Simonyan and Zisserman, 2015), and other vision–language models (Ben-younes et al., 2019). We further explore different textual data augmentation techniques and attention methods. For the shared task participation, we used only the image modality, which resulted in an underperforming system in the leaderboard.

Further studies using other modalities and approaches improved the performance of our system, but it is still lower (0.464 macro F1) than the best system (0.586). Yet, our investigation might be useful to understand which approaches are useful for detecting the role of an entity in harmful memes.

Our contributions can be summarized as follows:

- we addressed the problem both as sequence labeling and as classification;
- we investigated different pretrained models for text and images;
- we explored several combinations of multimodal models, as well as attention mechanisms, and various augmentation techniques.

The rest of the paper is organized as follows: Section 2 presents previous work, Section 3 describes the task and the dataset, Section 4 formulates our experiments, Section 5 discusses the evaluation results. Finally, Section 6 concludes and points to possible directions for future work.

2 Related Work

Below, we discuss previous work on semantic role labeling and harmful content detection, both in general and in a multimodal context.

2.1 Semantic Role Labeling

Textual semantic role labeling has been widely studied in NLP, where the idea is to understand who did what to whom, when, where, and why. Traditionally, the task has been addressed using sequence labeling, e.g., FitzGerald et al. (2015) used local and structured learning, experimenting with PropBank and FrameNet, and Larionov et al. (2019) investigated recent transformer models.

Visual semantic role labeling has been explored for images and video. Yatskar et al. (2016) addressed situation recognition, and developed a large-scale dataset containing over 500 activities, 1,700 roles, 11,000 objects, 125,000 images, and 200,000 unique situations. The images were collected from Google and the authors addressed the task as a situation recognition problem. Pratt et al. (2020) developed a dataset for situation recognition consisting of 278,336 bounding-box groundings to the 11,538 entity classes. Gupta and Malik (2015) developed a dataset of 16K examples in 10K images with actions and associated objects in the scene with different semantic roles for each action.

Yang et al. (2016) worked on integrating language and vision with explicit and implicit roles. Silberer and Pinkal (2018) learned frame–semantic representations of the images. Sadhu et al. (2021) approached the same problem for video, developing a dataset of 29K 10-second movie clips, annotated with verbs and semantics roles for every two seconds of video content.

2.2 Harmful Content Detection in Memes

There has been significant effort for identifying misinformation, disinformation, and malinformation online (Schmidt and Wiegand, 2017; Bondielli and Marcelloni, 2019; Zhou and Zafarani, 2020; Da San Martino et al., 2020; Alam et al., 2021c; Afridi et al., 2020; Hristakieva et al., 2022; Nakov et al., 2022). Most of these studies focused on textual and multimodal content. Compared to that, modeling the harmful aspects of memes has not received much attention.

Recent effort in this direction include categorizing hateful memes (Kiela et al., 2020), detecting antisemitism (Chandra et al., 2021), detecting the propagandistic techniques used in a meme (Dimitrov et al., 2021a), detecting harmful memes and the target of the harm (Pramanick et al., 2021), identifying the protected categories that were attacked (Zia et al., 2021), and identifying offensive content (Suryawanshi et al., 2020a). Among these studies, the most notable low-level efforts that advanced research by providing high-quality datasets to experiment with include shared tasks such as the *Hateful Memes Challenge* (Kiela et al., 2020), the SemEval-2021 shared task on detecting persuasion techniques in memes (Dimitrov et al., 2021b), and the troll meme classification task (Suryawanshi and Chakravarthi, 2021).

Chandra et al. (2021) investigated antisemitism along with its types as a binary and a multi-class classification problem using pretrained transformers and convolutional neural networks (CNNs) as modality-specific encoders along with various multimodal fusion strategies. Dimitrov et al. (2021a) developed a dataset with 22 propaganda techniques and investigated the different state-of-the-art pretrained models, demonstrating that joint vision–language models performed better than unimodal ones. Pramanick et al. (2021) addressed two tasks: detecting harmful memes and identifying the social entities they target, using a multimodal model with local and global information.

Zia et al. (2021) went one step further than a binary classification of hateful memes, focusing on a more fine-grained categorization based on the protected category that was being attacked (i.e., race, disability, religion, nationality, sex) and the type of attack (i.e., contempt, mocking, inferiority, slurs, exclusion, dehumanizing, inciting violence) using the dataset released in the WOAH 2020 Shared Task.[2] Fersini et al. (2019) studied sexist memes and investigated the textual cues using late fusion. They also developed a dataset of 800 misogynistic memes covering different manifestations of hatred against women (e.g., body shaming, stereotyping, objectification, and violence), collected from different social media (Gasparini et al., 2021).

Kiela et al. (2021) summarized the participating systems in the Hateful Memes Challenge, where the best systems fine-tuned unimodal and multimodal pre-training transformer models such as VisualBERT (Li et al., 2019) VL-BERT (Su et al., 2020), UNITER (Chen et al., 2020), VILLA (Gan et al., 2020), and built ensembles on top of them.

The SemEval-2021 propaganda detection shared task (Dimitrov et al., 2021b) focused on detecting the use of propaganda techniques in the meme, and the participants' systems showed that multimodal cues were very important.

In the troll meme classification shared task (Suryawanshi and Chakravarthi, 2021), the best system used ResNet152 and BERT with multimodal attention, and most systems used pretrained transformers for the text, CNNs for the images, and early fusion to combine the two modalities.

Combining modalities causes several challenges, which arise due to representation issues (i.e., symbolic representation for language vs. signal representation for the visual modality), misalignment between the modalities, and fusion and transferring knowledge between the modalities. In order to address multimodal problems, a lot of effort has been paid to developing different fusion techniques such as (*i*) *early fusion*, where low-level features from different modalities are learned, fused, and fed into a single prediction model (Jin et al., 2017b; Yang et al., 2018; Zhang et al., 2019; Singhal et al., 2019; Zhou et al., 2020; Kang et al., 2020), (*ii*) *late fusion*, where unimodal decisions are fused with some mechanisms such as averaging and voting (Agrawal et al., 2017; Qi et al., 2019),

[2] http://github.com/facebookresearch/fine_grained_hateful_memes

and (*iii*) *hybrid fusion*, where a subset of the learned features are passed to the final classifier (early fusion), and the remaining modalities are fed to the classifier later (late fusion) (Jin et al., 2017a). Here, we use early fusion and joint learning for fusion.

3 Task and Dataset

Below, we describe the CONSTRAINT 2022 shared task and the corresponding dataset provided by the task organizers. More detail can be found in the shared task report (Sharma et al., 2022c).

3.1 Task

The CONSTRAINT 2022 shared task asked participating systems to detect the role of the entities in the meme, given the meme and a list of these entities. Figure 1 shows an example of an image with the extracted OCR text, implicit (image showing Salman Khan, who is not mentioned in the text), and explicit entities and their roles. The example illustrates various challenges: (*i*) an implicit entity, (*ii*) text extracted from the label of the vial, which has little connection to the overlaid written text, (*iii*) unclear target entity in the meme (*Vladimir Putin*). Such complexities are not common in the multimodal tasks we discussed above. The textual representation of the entities and their roles are different than for typical CoNLL-style semantic role labeling tasks (Carreras and Màrquez, 2005), which makes it more difficult to address the problem in the same formulation.

By observing these challenges, we first attempted to address the problem in the same formulation: as a sequence labeling problem by converting the data to CoNLL format (see Section 4.1). Then, we further tried to address it as a classification task, i.e., predict the role of each entity in a given meme–entity pair.

3.2 Data

We use the dataset provided for the CONSTRAINT 2022 shared task. It contains harmful memes, OCR-extracted text from these memes, and manually annotated entities with four roles: *hero*, *villian*, *victim*, and *other*. The datasets cover two domains: COVID-19 and US Politics. The COVID-19 domain consists of 2,700 training and 300 validation examples, while US Politics has 2,852 training and 350 validation examples. The test dataset combines examples from both domains, COVID-19 and US Politics, and has a total of 718 examples.

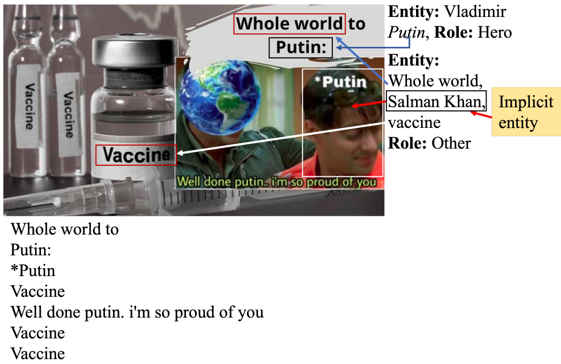

Figure 1: An example image showing the implicit (*Salman Khan*) and the explicit entities (from a text perspective) and their roles.

Class label	Train		Val		Test	
	Count	%	Count	%	Count	%
Hero	475	2	224	3	52	2
Villain	2,427	10	886	10	350	14
Victim	910	5	433	5	114	5
Others	13,702	83	6,937	82	1,917	79
Total	17,514		8,480		2,433	

Table 1: Distribution of the entity roles in the combined COVID-19 + US politics datasets.

For the experiments, we combined the two domains, COVID-19 and US Politics, which resulted in 5,552 training and 650 validation examples.

The class distribution of the entity roles, aggregated over all memes, in the combined COVID-19 + US Politics dataset is highly imbalanced as shown in Table 1. We can see that overall the role of *hero* represents only 2%, and the role of *victim* covers only 5% of the entities. We can further see that the vast majority of the entities are labeled with the *other* role. This skewed distribution adds additional complexity to the modeling task.

4 Experiments

Settings: We addressed the problem both as a sequence labeling and as a classification task. Below, we discuss each of them in detail.

Evaluation measures: In our experiments, we used accuracy, macro-average precision, recall, and F_1 score. The latter was the official evaluation measure for the shared task.

Provided JSON

{"OCR": "Bernie or Elizabeth?\nBe informed.Compare them on the issues that matter.\nIssue: Who makes the dankest memes?\n", "image": "covid_memes_18.png", "hero": [], "villain": [], "victim": [], "other": ["bernie sanders", "elizabeth warren"]}

Converted into IOB format

bernie	or	elizabeth	?	be	...	dankest	memes	?	sanders	warren
B-other	O	B-other	O	O	...	O	O	O	B-other	I-other

Figure 2: Example with text in BIO format.

4.1 Sequence Labeling

For the sequence labeling experiments, we first converted the OCR text and the entities to the CoNLL BIO-format. An example is shown in Figure 2. To convert them, we matched the entities in the text and we assigned the same tag (role label) to the token in the text. For the implicit entity that is not in the text, we added them at the end of the text and we assigned them the annotated role; we labeled all other tokens with the O-tag.

We trained the model using Conditional Random Fields (CRFs) (Lafferty et al., 2001), which has been widely used in earlier work. As features, we used part-of-speech tags, token length, tri-grams, presence of digits, use of special characters, token shape, w2vcluster, LDA topics, words present in a vocabulary list built on the training set, and in a name list, etc.[3] We ran two sets of experiments: (*i*) using the same format, and (*ii*) using only entities as shown in Figure 2.

4.2 Classification

For the classification experiments, we first converted the dataset into a classification problem. As it contains all examples with one or more entities, we reorganized the dataset so that an example contains an entity, OCR text, image, and entity role. Hence, the dataset size is now the same as the number of entity instances rather than memes. We ended up with 17,514 training examples, which is the number of training entities as shown in Table 1.

We then ran different unimodal and multimodal experiments: (*i*) only text, (*ii*) only meme, and (*iii*) text and meme together. For each setting, we also ran several baseline experiments. We further ran advanced experiments such as adding attention to the network and text-based data augmentation. Figure 3 shows our experimental pipeline for this classification task. For the unimodal experiments, we used individual modalities, and we trained them using different pre-trained models.

[3]More details about the feature set can be found at https://github.com/moejoe95/crf-vs-rnn-ner

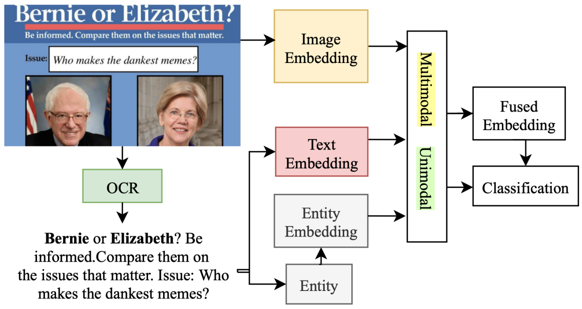

Figure 3: Diagram of our experimental pipeline.

Note that for the text modality, we ran several combinations of fusion (e.g., text and entity) experiments. For the multimodal experiments, we combined embedding from both modalities, and we ran the classification on the fused embedding, as shown in Figure 3.

4.2.1 Text Modality

For the text modality, we experimented using BERT (Devlin et al., 2019) and XLM-RoBERTa (Liu et al., 2019). We performed ten reruns for each experiment using different random seeds, and then we picked the model that performed best on the development set. We used a batch size of 8, a learning rate of 2e-5, a maximum sequence length of 128, three epochs, and categorical cross-entropy as the loss function. We used the Transformer toolkit to train the transformer-based models.

Using the text-only modality, we also ran a different combination of experiments using the text and the entities, where we used bilinear fusion to combine them. We discuss this fusion technique in more detail in Section 4.2.3.

4.2.2 Image Modality

For our experiments using the image modality, we extract features from a pre-trained model, and then we trained an SVM classifier using these features. In particular, we extracted features from the penultimate layer of the EfficientNet-b1 (EffNet) model (Tan and Le, 2019), which was trained using the ImageNet dataset. For training the model using the extracted features, we used SVM with its default parameter settings, with no further optimization of its hyper-parameter values. We chose EffNet as it was shown to achieve better performance for some social media image classification tasks (Alam et al., 2021a,b).

4.2.3 Multimodal: Text and Image

For the multimodal experiments, we used the BLOCK Fusion (Ben-younes et al., 2019) approach, which was originally proposed for question answering (QA). Our motivation is that an entity can be seen like a question about the meme context, asking for its role as an answer. In a QA setting, there are three elements: (*i*) a context (image or text), (*ii*) a question, and (*iii*) a list of answers. The goal is to select the right answer from the answer list. Similarly, we have four types of answers (i.e., roles). The task formation is that for an entity and a context (image or text), we need to determine the role of the entity in that context.

BLOCK fusion is a multi-modal framework based on block-superdiagonal tensor decomposition, where tensor blocks are decomposed into blocks of smaller sizes, with the size characterized by a set of mode-n ranks (De Lathauwer, 2008). It is a bilinear model that takes two vectors $x^1 \in R^I$ and $x^2 \in R^J$ as input and then projects them to a K-dimensional space with tensor products: $y = \mathcal{T} \times x^1 \times x^2$, where $y \in R^K$. Each component of y is a quadratic form of the inputs, $\forall k \in [1; K]$:

$$y_k = \sum_{i=1}^{I} \sum_{j=1}^{J} \mathcal{T}_{ijk} x_i^1 x_2^j \tag{1}$$

BLOCK fusion can model bilinear interactions between groups of features, while limiting the complexity of the model, but keeping expressive high dimensional mono-model representations (Ben-younes et al., 2019). We used BLOCK fusion in different settings: (*i*) for image and entity, (*ii*) for text and entity, and (*iii*) for text, image with entity.

Text and entity: We extracted embedding representation for the entity and the text using a pre-trained BERT model. We then fed both embedding representations into linear layers of 512 neurons each. The output of two linear layers is taken as input to the trainable block fusion network. Then, a regularization layer and linear layer are used before the final layer.

Image and entity: To build embedding representations for the image and the entity, we used a vision transformer (ViT) (Dosovitskiy et al., 2021) and BERT pretrained models. The output of two different modalities was then used as input to the block fusion network.

Image, text, and entity: In this setting, we first built embedding representations for the text and the image using a pretraind BERT and ViT models, respectively. Then, we concatenated these representations (text + image) and we passed them to a linear layer with 512 neurons. We then extracted embedding representation for the target entity using the pretraind BERT model. Afterwards, we merged the text + image and the entity representations and we fed them into the fusion layer. In this way, we combined the image and the text representations as a unified context, aiming to predict the role of the target entity in this context.

In all the experiments, we uses a learning rate of $1e^{-6}$, a batch size of 8, and a maximum length of the text of 512.

4.2.4 Additional Experiments

We ran two additional sets of experiments using attention mechanism and augmentation, as using such approaches has been shown to help in many natural language processing (NLP) tasks.

Attention: In the entity + image block fusion network, we used block fusion to merge the entity and the image representations. Instead of using the image representation directly, we used attention mechanism on the image and then we fed the attended features along with the entity representation into the entity + image block. To compute the attention, we used the PyTorchNLP library.[4] In a similar fashion, we applied the attention mechanism to the text and to the combined text + image representation.

Augmentation: Text data augmentation has recently gained a lot of popularity as a way to address data scarceness and class imbalance (Feng et al., 2021). We used three types of text augmentation techniques to balance the distribution of the different class: (*i*) synonym augmentation using WordNet, (*ii*) word substitution using BERT, and (*iii*) a combination thereof. In our experiments, we used the NLPAug data augmentation package.[5] Note that we applied six times augmentation for the *hero* class, twice for the *villain* class, and three times for the *victim* class. These numbers were empirically set and require further investigation in future work.

[4] http://github.com/PetrochukM/PyTorch-NLP

[5] https://github.com/makcedward/nlpaug

Exp.	Acc	P	R	F1
All tokens	0.51	0.32	0.21	0.24
Only entities	0.77	0.40	0.27	0.25

Table 2: Evaluation results on the test set for the sequence labeling reformulation of the problem.

5 Results and Discussion

Below, we first discuss our sequence labeling and classification experiments. We then perform some analysis, and finally, we put our results in a broader perspective in the context of the shared task.

5.1 Sequence Labeling Results

Table 2 shows the evaluation results on the test set for our sequence labeling reformulation of the problem. We performed two experiments: one where we used as input the entire meme text (i.e., all tokens), and another one where we used the concatenation of the target entities only. We can see that the latter performed marginally better, but overall the macro-F1 score is quite low in both cases.

5.2 Classification Results

Table 3 shows the evaluation results on the test set for our classification reformulation of the problem. We computed the *majority class* baseline (row 0), which always predicts the most frequent label in the training set. Due to time limitations, our official submission used the image modality only, which resulted in a very low macro-F1 score of 0.23, as shown in row 1. For our text modality experiments, we used the meme text and the entities. We experimented with BERT and XLM-RoBERTa, obtaining better results using the former. Using the BLOCK fusion technique on unimodal (text + entity) and multimodality (text + image + entity) yielded sizable improvements. The combination of image + text (rows 6 and 9) did not yield much better results compared to using text only (row 4). Next, we added attention on top of block fusion, which improved the performance, but there was no much difference between the different combinations (rows 7–9). Considering only the text and the entity, we observe an improvement using text augmentation. Among the different augmentation techniques, there was no performance difference between WordNet and BERT, and combining them yielded worse results.

	Exp.	Acc	P	R	F1
	Baseline				
0	Majority	0.79	0.20	0.25	0.22
	Image modality				
1	*EffNet feat + SVM*	*0.72*	*0.24*	*0.25*	*0.23*
	Text modality				
2	BERT	0.76	0.42	0.36	0.37
3	XLM-RoBERTa	0.75	0.38	0.32	0.32
	Multimodality/Fusion				
	BLOCK fusion				
4	Entity + Text	0.74	0.44	0.43	0.43
5	Entity + Image	0.74	0.39	0.39	0.39
6	Entity + (Text + Image)	0.75	0.43	0.42	0.41
	Attention				
7	Entity + Text	0.72	0.42	0.48	0.44
8	Entity + Image	0.71	0.42	0.48	0.44
9	Entity + (Text + Image)	0.71	0.42	0.49	0.44
	Augmentation				
10	Entity + Text (WordNet aug)	0.76	0.48	0.46	**0.46**
11	Entity + Text (BERT aug)	0.74	0.46	0.46	**0.46**
12	Entity + Text (Mix aug)	0.77	0.49	0.41	0.43

Table 3: Evaluation results on the test set for our classification reformulation of the problem. Our official submission for the shared task is shown in *italic*.

5.3 Role-Level Analysis

Next, we studied the impact of using attention and data augmentation on the individual entity roles: *hero*, *villain*, *victim*, and *other*.

Table 4 shows the impact of using attention on (a) entity + image (left side), and (b) entity + [image + text] (right side) combinations. We can observe a sizable gain for the *hero* (+0.09), the *villain* (+0.06), and the *victim* (+0.07) roles in the former case (a). However, for case (b), there is an improvement for the *victim* role only; yet, this improvement is quite sizable: +0.16.

Table 5 shows the impact of data augmentation using WordNet or BERT on the individual roles. We can observe sizable performance gains of +0.11 for the *hero* role, and +0.04 for the *villain* role, when using WordNet-based data augmentation. Similarly, BERT-based data augmentation yields +0.12 for the *hero* role, and +0.02 for the *villain* role. However, the impact of either augmentation on the *victim* and on the *other* role is negligible.

	E+I, w/o Att.			E+I, w/ Att.			E+[I+T], w/o Att.			E+[I+T], w/ Att.		
Role	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Hero	0.06	0.02	0.03	0.09	0.15	**0.12**	0.22	0.12	0.15	0.09	0.21	0.12
Villain	0.35	0.44	0.39	0.40	0.51	**0.45**	0.39	0.54	0.45	0.39	0.54	0.45
Victim	0.30	0.25	0.28	0.33	0.39	**0.35**	0.23	0.18	0.20	0.31	0.45	**0.36**
Other	0.86	0.84	0.85	0.88	0.81	0.84	0.87	0.84	0.85	0.89	0.77	0.82

Table 4: Role-level results on the test set with (w/) or without (w/o) attention between the context (text, image) and the entity. (E: Entity, I: Image, Att.: Attention, T: Text)

	No Aug.			Aug. WordNet			Aug. BERT		
Role	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Hero	0.21	0.12	0.15	0.33	0.21	**0.26**	0.30	0.25	**0.27**
Villain	0.36	0.49	0.42	0.41	0.52	**0.46**	0.39	0.51	**0.44**
Victim	0.31	0.27	0.29	0.30	0.27	0.29	0.29	0.27	0.28
Other	0.87	0.83	0.85	0.87	0.84	0.86	0.87	0.83	0.85

Table 5: Role-level results on the test set for the entity + text combination with and without augmentation.

5.4 Official Submission

For our official submission for the task, we used the image modality system from line 1 in Table 3, which was quite weak, with a macro-F1 score of 0.23. Our subsequent experiments and analysis pointed to several promising directions: (*i*) combining the textual and the image modalities, (*ii*) using attention, (*iii*) performing data augmentation. As a result, we managed to improve our results to 0.46. Yet, this is still far behind the F1-score of the winning system: 0.5867.

6 Conclusion and Future Work

We addressed the problem of understanding the role of the entities in harmful memes, as part of the CONSTRAINT-2022 shared task. We presented a comparative analysis of the importance of different modalities: the text and the image. We further experimented with two task reformulations —sequence labeling and classification—, and we found the latter to work better. Overall, we obtained improvements when using BLOCK fusion, attention between the image and the text representations, and data augmentation.

In future work, we plan to combine the sequence and the classification formulations in a joint multimodal setting. We further want to experiment with multi-task learning using other meme analysis tasks and datasets. Last but not least, we plan to develop better data augmentation techniques to improve the performance on the low-frequency roles.

Acknowledgments

The work is 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

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young Koo Lee. 2020. A multimodal memes classification: A survey and open research issues. In *Proceedings of the 5th International Conference on Smart City Applications*, SCA '20, pages 1451–1466, Online. Springer.

Taruna Agrawal, Rahul Gupta, and Shrikanth Narayanan. 2017. Multimodal detection of fake social media use through a fusion of classification and pairwise ranking systems. In *Proceedings of the 25th European Signal Processing Conference*, EUSIPCO '17, pages 1045–1049. IEEE.

Firoj Alam, Tanvirul Alam, Md Hasan, Abul Hasnat, Muhammad Imran, Ferda Ofli, et al. 2021a. MEDIC: a multi-task learning dataset for disaster image classification. *arXiv:2108.12828*.

Firoj Alam, Tanvirul Alam, Muhammad Imran, and Ferda Ofli. 2021b. Robust training of social media image classification models for rapid disaster response. *arXiv:2104.04184*.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San

Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021c. A survey on multimodal disinformation detection. *arXiv preprint arXiv: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. 2021d. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics*, EMNLP (Findings) '21, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 AAAI Conference on Artificial Intelligence*, AAAI '19, Honolulu, Hawaii, USA. AAAI Press.

Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55.

Sian Brooke. 2019. "Condescending, Rude, Assholes": Framing gender and hostility on Stack Overflow. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180, Florence, Italy. Association for Computational Linguistics.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the 9th Conference on Computational Natural Language Learning*, CoNLL '05, pages 152–164, Ann Arbor, Michigan, USA.

Mohit Chandra, Dheeraj Reddy Pailla, Himanshu Bhatia, AadilMehdi J. Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "Subverting the Jewtocracy": Online antisemitism detection using multimodal deep learning. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, pages 148–157. ACM.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to detect human-object interactions. In *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision*, WACV '18, pages 381–389, Lake Tahoe, Nevada, USA. IEEE.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, ICCV '15, pages 1017–1025, Santiago, Chile. IEEE.

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. In *Proceedings of the European Conference on Computer Vision*, ECCV '20, pages 104–120, Cham. Springer International Publishing.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI '20, pages 4826–4832, Online. IJCAI.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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 '19, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Lieven De Lathauwer. 2008. Decompositions of a higher-order tensor in block terms—part ii: Definitions and uniqueness. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1033–1066.

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*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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 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, Online. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. Task 6 at SemEval-2021: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International*

Conference on Learning Representations, ICLR '21, Online.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist MEME on the web: A study on textual and visual cues. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction*, ACIIW '19, pages 226–231, Cambridge, UK. IEEE.

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Semantic role labeling with neural network factors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 960–970, Lisbon, Portugal. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2021. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *arXiv:2106.08409*.

Saurabh Gupta and Jitendra Malik. 2015. Visual semantic role labeling. *arXiv:1505.04474*.

Batoul Haidar, Maroun Chamoun, and Fadi Yamout. 2016. Cyberbullying detection: A survey on multilingual techniques. In *Proceedings of the 2016 European Modelling Symposium*, EMS '2016, pages 165–171, Pisa, Italy. IEEE.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '2022, Seattle, Washington, USA.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

Kristina Hristakieva, Stefano Cresci, Giovanni Da San Martino, Mauro Conti, and Preslav Nakov. 2022. The spread of propaganda by coordinated communities on social media. In *Proceedings of the 14th International ACM Conference on Web Science*, WebSci '2022, Barcelona, Spain. ACM.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017a. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, MM '17, pages 795–816, California, USA. ACM.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017b. Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, 19(3):598–608.

Srecko Joksimovic, Ryan S. Baker, Jaclyn Ocumpaugh, Juan Miguel L. Andres, Ivan Tot, Elle Yuan Wang, and Shane Dawson. 2019. Automated identification of verbally abusive behaviors in online discussions. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 36–45, Florence, Italy. Association for Computational Linguistics.

SeongKu Kang, Junyoung Hwang, and Hwanjo Yu. 2020. Multi-modal component embedding for fake news detection. In *Proceedings of the 14th International Conference on Ubiquitous Information Management and Communication*, IMCOM '20, pages 1–6, Taichung, Taiwan. IEEE.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, et al. 2021. The hateful memes challenge: competition report. In *Proceedings of the 35th International Conference on Neural Information Processing Systems: Competition and Demonstration Track*, NeurIPS '21, pages 344–360, Online.

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*, NIPS '20, NY, USA. Curran Associates Inc.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying*, TRAC'2018, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, California, USA. Morgan Kaufmann Publishers Inc.

Daniil Larionov, Artem Shelmanov, Elena Chistova, and Ivan Smirnov. 2019. Semantic role labeling with pretrained language models for known and unknown predicates. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '2019, pages 619–628, Varna, Bulgaria. INCOMA Ltd.

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*.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 2557–2568, Online. Association for Computational Linguistics.

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*.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, WALO '17, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

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 *Advances in Information Retrieval*, CLEF '2022, pages 416–428. Springer International Publishing.

Preslav Nakov, Husrev Taha Sencar, Jisun An, and Haewoon Kwak. 2021. A survey on predicting the factuality and the bias of news media. *arXiv:2103.12506*.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics*, EMNLP (Findings) '21, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *Proceedings of the European Conference on Computer Vision*, ECCV '20, pages 314–332, Online. Springer.

Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *Proceedings of the IEEE International Conference on Data Mining*, ICDM '19, pages 518–527, Beijing, China. IEEE.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification. In *Findings of the Association for Computational Linguistics*, ACL-IJCNLP (Findings) '21, pages 915–928, Online. Association for Computational Linguistics.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '21, pages 5589–5600, Online. IEEE.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022a. DISARM: Detecting the victims targeted by harmful memes. In *Findings of North American Chapter of the Association for Computational Linguistics*, EMNLP (Findings) '22, Seattle, Washington, USA. Association for Computational Linguistics.

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.

Shivam Sharma, Tharun Suresh, Atharva Jitendra, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022c. Findings of the constraint 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, CONSTRAINT '22, Dublin, Ireland. Association for Computational Linguistics.

Limor Shifman. 2013. *Memes in digital culture*. MIT press.

Carina Silberer and Manfred Pinkal. 2018. Grounding semantic roles in images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 2616–2626, Brussels, Belgium. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15, San Diego, CA, USA.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. SpotFake: A multi-modal framework for fake news detection. In *Proceedings of the 2019 IEEE fifth international conference on multimedia big data*, BigMM '19, pages 39–47. IEEE.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *Proceedings of the 8th International Conference on Learning Representations*, ICLR '20, Addis Ababa, Ethiopia. OpenReview.net.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, Kyiv. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020a. 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. European Language Resources Association.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020b. A dataset for troll classification of TamilMemes. In *Proceedings of the 5th Workshop on Indian Language Data: Resources and Evaluation*, WILDRE '20, pages 7–13, Marseille, France. European Language Resources Association.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, ICLR '19, pages 6105–6114, CA, USA.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '15, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, Bulgaria.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Chai. 2016. Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '16, pages 149–159, San Diego, California. Association for Computational Linguistics.

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. 2018. TI-CNN: Convolutional neural networks for fake news detection. *arXiv:1806.00749*.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, San Juan, PR, USA. IEEE.

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

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*, IMC '18, pages 188–202, Boston, USA. ACM.

Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1942–1951, Nice, France. ACM.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-aware multi-modal fake news detection. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, PAKDD '20, pages 354–367, Singapore. Springer.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40.

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? Classifying memes beyond hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms*, WOAH '21, pages 215–219, Online. Association for Computational Linguistics.