Racist or Sexist Meme? Classifying Memes beyond Hateful

Haris Bin Zia
Queen Mary
University of London
United Kingdom
h.b.zia@qmul.ac.uk

Ignacio Castro
Queen Mary
University of London
United Kingdom
i.castro@qmul.ac.uk

Gareth Tyson
Queen Mary
University of London
United Kingdom
g.tyson@qmul.ac.uk

Abstract

Memes are the combinations of text and images that are often humorous in nature. But, that may not always be the case, and certain combinations of texts and images may depict hate, referred to as *hateful memes*. This work presents a multimodal pipeline that takes both visual and textual features from memes into account to (1) identify the protected category (e.g. race, sex etc.) that has been attacked; and (2) detect the type of attack (e.g. contempt, slurs etc.). Our pipeline uses state-of-the-art pre-trained visual and textual representations, followed by a simple logistic regression classifier. We employ our pipeline on the Hateful Memes Challenge dataset with additional newly created fine-grained labels for protected category and type of attack. Our best model achieves an AUROC of 0.96 for identifying the protected category, and 0.97 for detecting the type of attack. We release our code at https://github.com/harisbinzia/HatefulMemes

1 Introduction

An internet meme (or simply "meme" for the remainder of this paper) is a virally transmitted image embellished with text. It usually shares pointed commentary on cultural symbols, social ideas, or current events (Gil, 2020). In the past few years there has been a surge in the popularity of memes on social media platforms. Instagram, which is a popular photo and video sharing social networking service recently revealed that over 1 million posts mentioning "meme" are shared on Instagram each day.[1] We warn the reader that this paper contains content that is racist, sexist and offensive in several ways.

[1] https://about.instagram.com/blog/announcements/instagram-year-in-review-how-memes-were-the-mood-of-2020

Although memes are often funny and used mostly for humorous purposes, recent research suggests that they can also be used to disseminate hate (Zannettou et al., 2018) and can therefore emerge as a multimodal expression of online hate speech. Hateful memes target certain groups or individuals based on their race (Williams et al., 2016) and gender (Drakett et al., 2018), among many other protected categories, thus causing harm at both an individual and societal level. An example hateful meme is shown in Figure 1.

Figure 1: An example of a hateful meme. The meme is targeted towards a certain religious group.

At the scale of the internet, it is impossible to manually inspect every meme. Hence, we posit that it is important to develop (semi-)automated systems that can detect hateful memes. However, detecting hate in multimodal forms (such as memes) is extremely challenging and requires a holistic understanding of the visual and textual material. In order to accelerate research in this area and develop systems capable of detecting hateful memes, Facebook recently launched The Hateful Memes Challenge (Kiela et al., 2020). The challenge introduced a new annotated dataset of around 10K memes tagged for hatefulness (i.e. hateful vs. not-hateful). The baseline results show a substantial dif-

ference in the performance of unimodal and multimodal systems, where the latter still perform poorly compared to human performance, illustrating the difficulty of the problem.

More recently, a shared task on hateful memes was organized at the Workshop on Online Abuse and Harms[2] (WOAH), where the hateful memes dataset (Kiela et al., 2020) was presented with additional newly created fine-grained labels[3] for the protected category that has been attacked (e.g. race, sex, etc.), as well as the type of attack (e.g. contempt, slurs, etc.). This paper presents our multimodal pipeline based on pre-trained visual and textual representations for the shared task on hateful memes at WOAH. We make our code publicly available to facilitate further research.[4]

2 Problem Statement

There are two tasks with details as follows:

- *Task A*: For each meme, detect the Protected Category (PC). Protected categories are: race, disability, religion, nationality, sex. If the meme is not-hateful, the protected category is: pc_empty
- *Task B*: For each meme, detect the Attack Type (AT). Attack types are: contempt, mocking, inferiority, slurs, exclusion, dehumanizing, inciting violence. If the meme is not-hateful, the protected category is: attack_empty

Note, Tasks A and B are multi-label because memes can contain attacks against multiple protected categories and can involve multiple attack types.

3 Dataset

The dataset consists of 9,540 fine-grained annotated memes and is imbalanced, with large number of non-hateful memes and relatively small number of hateful ones. The details of different splits[5] are given in the Table 1 and the distribution of classes are given in Table 2. The majority of memes in the dataset are single-labeled. Figure 2 and Figure 3 present the distribution of memes with multiple protected categories and types of attacks respectively. For the evaluation, we use the standard AUROC metric.

	# memes		
split	**hateful**	**not-hateful**	***total***
train	3007	5493	8500
dev (seen)	246	254	500
dev (unseen)	199	341	540

Table 1: Dataset statistics.

	classes	**train**	**dev (seen)**	**dev (unseen)**
PC	pc_empty	5495	254	341
	religion	1078	95	77
	race	1008	78	63
	sex	746	56	46
	nationality	325	26	20
	disability	255	22	17
AT	attack_empty	5532	257	344
	mocking	378	35	29
	dehumanizing	1318	121	104
	slurs	205	6	4
	inciting violence	407	26	23
	contempt	235	10	6
	inferiority	658	49	35
	exclusion	114	13	8

Table 2: Distribution of classes in splits.

4 Model & Results

This section describes our model, the visual & textual embeddings, as well as the results.

4.1 Embeddings

We use the following state-of-the-art pre-trained visual and textual representations:

- CLIP[6]: OpenAI's CLIP (Contrastive Language Image Pre-Training) (Radford et al., 2021) is a neural network that jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) examples. We use pre-trained CLIP image encoder (hereinafter CIMG) and CLIP text

[2] https://www.workshopononlineabuse.com

[3] https://github.com/facebookresearch/fine_grained_hateful_memes

[4] https://github.com/harisbinzia/HatefulMemes

[5] Note, at the time of writing, the gold annotations were available only for train, dev (seen) and dev (unseen) sets. We used train for training, dev (seen) for hyperparameter tuning and dev (unseen) to report results. We also report the results on a blind test set as released by the organizers of WOAH.

[6] https://github.com/OpenAI/CLIP

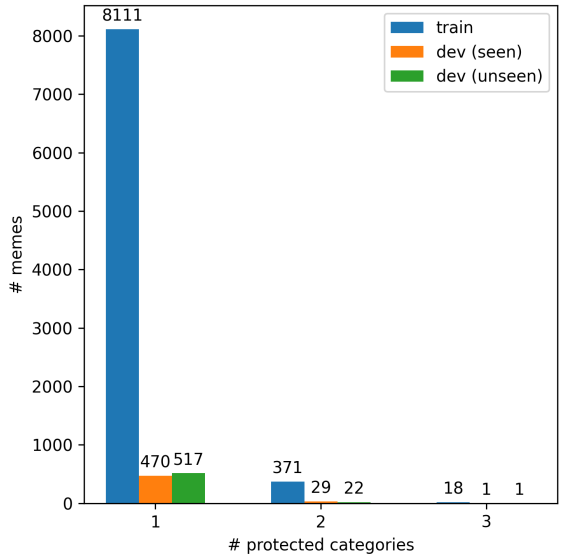

Figure 2: Count of memes with multiple protected categories.

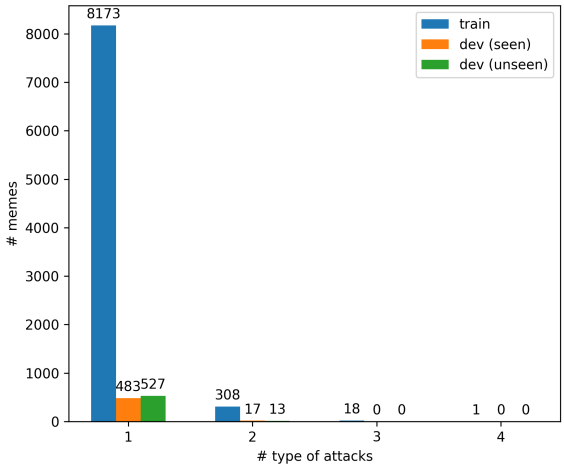

Figure 3: Count of memes with multiple attack types.

encoder (hereinafter CTXT) to embed meme images and text respectively.

- LASER[7]: Facebook's LASER (Language Agnostic SEntence Representations) (Artetxe and Schwenk, 2019) is a BiLSTM based seq2seq model that maps a sentence in any language to a point in a high-dimensional space with the goal that the same statement in any language will end up in the same neighborhood. We use LASER encoder to obtain embeddings for the meme text.

- LaBSE: Google's LaBSE (Language agnostic BERT Sentence Embedding) (Feng et al., 2020) is a Transformer (BERT) based embedding model that produces language-agnostic cross-lingual sentence embeddings. We use the LaBSE model to embed meme text.

4.2 Pipeline

Exploiting the above models, we employ a simple four step pipeline as shown in Figure 4:

1. We extract text from the meme.

2. We embed the meme image and the text into visual and textual representations (Section 4.1).

3. We concatenate the visual and textual embeddings.

4. We train a multi-label Logistic Regression classifier using scikit-learn (Pedregosa et al., 2011) to predict the protected category attacked in the meme (Task A) and the type of attack (Task B).

4.3 Results

The results are shown in Table 3, where we contrast various configurations of our classifier. We observe that the vision-only classifier, which only uses visual embeddings (CIMG), performs slightly better than the text-only classifier, which only uses textual embeddings (CTXT, LASER or LaBSE). The multimodal models outperform their unimodal counterparts. Our best performing model is multimodal, trained on the concatenated textual (CTXT, LASER and LaBSE) and visual (CIMG) embeddings.[8] Class-wise performance of best model is given in Table 4.

5 Conclusion & Future Work

This paper has presented our pipeline for the multi-label hateful memes classification shared task organized at WOAH. We show that our multimodal classifiers outperform unimodal classifiers. Our best multimodal classifier achieves an AUROC of 0.96 for identifying the protected category, and 0.97 for detecting the attack type. Although we trained our classifier on language agnostic representations, it was only tested on a dataset of English memes. As a future direction, we plan to extend our work

[7] https://github.com/facebookresearch/LASER

[8] On a blind test set of 1000 memes our best model achieves an AUROC of 0.90 for Task A and 0.91 for Task B

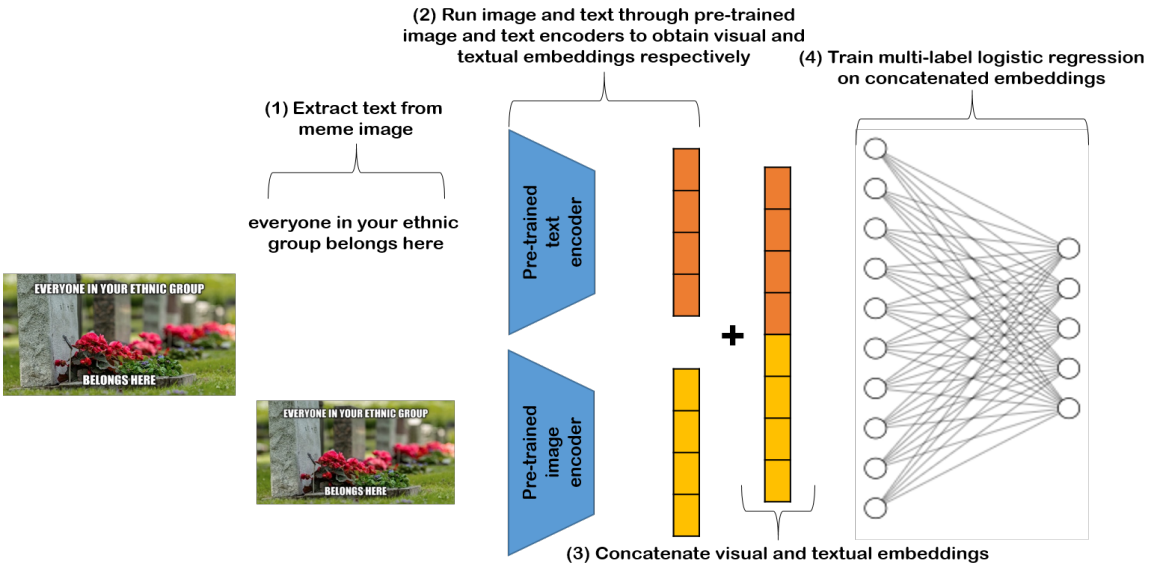

Figure 4: Multimodal pipeline for multi-label meme classification.

		AUROC	
Type	**Embedding**	**Task A**	**Task B**
Unimodal	CTXT	0.56	0.67
	LASER	0.88	0.91
	LaBSE	0.89	0.92
	CIMG	0.93	0.94
Multimodal	CIMG + CTXT	0.95	0.96
	CIMG + LASER	0.94	0.95
	CIMG + LaBSE	0.94	0.95
	CIMG + CTXT + LASER + LaBSE	**0.96**	**0.97**

Table 3: Model performance.

	classes	**Precision**	**Recall**	**F1**
PC	pc_empty	0.74	0.82	0.78
	religion	0.78	0.61	0.69
	race	0.57	0.49	0.53
	sex	0.85	0.61	0.71
	nationality	0.65	0.75	0.70
	disability	0.94	0.88	0.91
AT	attack_empty	0.74	0.82	0.78
	mocking	0.77	0.79	0.78
	dehumanizing	0.68	0.44	0.53
	slurs	0.80	1.00	0.89
	inciting violence	0.67	0.61	0.64
	contempt	1.00	0.33	0.50
	inferiority	0.73	0.31	0.44
	exclusion	1.00	1.00	1.00

Table 4: Class-wise performance of best model.

to multilingual settings, where we evaluate the performance of our classifier on multilingual memes.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jessica Drakett, Bridgette Rickett, Katy Day, and Kate Milnes. 2018. Old jokes, new media–online sexism and constructions of gender in internet memes. *Feminism & Psychology*, 28(1):109–127.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Paul Gil. 2020. Examples of memes and how to use them.

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. *arXiv preprint arXiv:2005.04790*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,

D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Amanda Williams, Clio Oliver, Katherine Aumer, and Chanel Meyers. 2016. Racial microaggressions and perceptions of internet memes. *Computers in Human Behavior*, 63:424–432.

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*, pages 188–202.