6-2022

Multimodal zero-shot hateful meme detection

Jiawen ZHU

Roy Ka-Wei LEE

Wen Haw CHONG
Singapore Management University, whchong.2013@phdis.smu.edu.sg

Multimodal Zero-Shot Hateful Meme Detection

Jiawen Zhu
Singapore University of Technology and Design
Singapore, Singapore
jiawen_zhu@sutd.edu.sg

Roy Ka-Wei Lee
Singapore University of Technology and Design
Singapore, Singapore
roy_lee@sutd.edu.sg

Wen-Haw Chong
Singapore Management University
Singapore, Singapore
whchong.2013@phdis.smu.edu.sg

ABSTRACT

Facebook has recently launched the hateful meme detection challenge, which garnered much attention in academic and industry research communities. Researchers have proposed multimodal deep learning classification methods to perform hateful meme detection. While the proposed methods have yielded promising results, these classification methods are mostly supervised and heavily rely on labeled data that are not always available in the real-world setting. Therefore, this paper explores and aims to perform hateful meme detection in a zero-shot setting. Working towards this goal, we propose Target-Aware Multimodal Enhancement (TAME), which is a novel deep generative framework that can improve existing hateful meme classification models' performance in detecting unseen types of hateful memes. We conduct extensive experiments on the Facebook hateful meme dataset, and the results show that TAME can significantly improve the state-of-the-art hateful meme classification methods' performance in seen and unseen settings.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Computer vision representations**.

KEYWORDS

hateful memes, multimodal, social media mining

ACM Reference Format:
Jiawen Zhu, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. Multimodal Zero-Shot Hateful Meme Detection. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3501247.3531557

Disclaimer: *This paper contains violence and discriminatory content that may be disturbing to some readers. Specifically, the analyses contain actual examples of hateful memes and hate speech targeting particular groups. These examples are very offensive and distasteful. However, we have made the hard decision to display these actual hateful examples to provide context on the toxicity of malicious content that we are dealing with. Besides making technical contributions in this paper, we hope the distasteful examples used could also raise awareness of the vulnerable groups targeted in hate speeches in the real-world.*

WebSci '22, June 26–29, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9191-7/22/06…$15.00
https://doi.org/10.1145/3501247.3531557

1 INTRODUCTION

Motivation. Online hateful content, which attacks or uses discriminatory languages targeting individuals or groups based on protected characteristics such as religion, race, and gender, is of growing social concern. The increase in hateful online content has shown discord among communities and resulted in violent hate crimes. To tackle this issue, social media platforms and researchers have proposed automated methods to detect and curb hateful online content. For instance, Facebook recently launched a hateful meme detection challenge to combat the spread of hateful memes [17]. The challenge motivated the development of new multimodal hateful meme classification methods [20, 23, 32, 38]. These hateful memes often target certain communities and or individuals based on race, religion, gender, or physical attributes, by portraying them in a derogatory manner [14, 17, 30].

Detecting hateful memes is a challenging multimodal problem that requires a holistic understanding of the image, text, and context of both modalities. While humans can intrinsically understand the combined meaning of texts and images in memes, machines have difficulty performing such a complex task. Consider an example of a hateful meme illustrated in Figure 1; when we examine the text and image as independent features, the content seems normal and benign. However, when we interpret the meme as a whole, the underlying message is very offensive and hateful. Another key element that helped us understand hateful memes is the context illustrated in the memes. The target entities (e.g., race, religion, etc.) in the hateful message often provide important contextual information. For the example in Figure 1, the target entity of the hateful content would be African or African American in the context of slavery.

Several studies had proposed fusion techniques to combine the memes' text and visual features for hateful meme classification [17, 30]. Others have explored fine-tuning large-scale pre-trained multimodal methods to perform the task [17, 23, 32, 38]. Nevertheless, existing methods heavily rely on labelled data that are not always available in the real-world setting. In particular, a model trained on existing hateful memes with seen target types (e.g., African American) will not generalize well to detecting new hateful memes with unseen target types (e.g., transgender).

Research Objectives. We fill this research gap by proposing a new machine learning task: *zero-shot hateful meme detection*. This evaluates how models performed when applied to detecting hateful memes with unseen target types. Towards this goal, we also propose a novel deep generative framework, Target-Aware Multimodal Enhancement (TAME), to generate target-aware latent

Figure 1: Example of hateful meme.

representations for memes with seen and unseen target types. We show that various base models can be plugged into this framework with improved hateful meme classification performance for both normal and zero-shot settings. Specifically, TAME[1] combines the strengths of Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) to learn discriminative latent representations that generalize well to both seen and unseen target types. VAE is designed to be a regularized Auto-Encoder, where the latent space is constrained to the prior distribution by optimizing the lower bound of the data likelihood. GAN takes the samples as a proxy to minimize the distribution divergence between the generated and real data through a two-player min-max game. Inspired by [3], we propose a joint VAE-GAN generative architecture where instead of matching VAE's encoded distribution of training examples to the prior distribution, we map the prior random vector into the encoded latent space by adversarial training based on GAN. The underlying intuition of TAME is to train a VAE-GAN generative framework using hateful memes of seen types and generate a target-aware representation for hateful memes with unseen target types. Through the generative framework, TAME ultimately maps the visual-linguistic and semantic features of memes into a common latent space, whereby semantic features are derived from seen target types during model training.

Contributions. We summarize our contribution as follows:

(1) We propose a zero-shot hateful meme detection task to evaluate the performance of existing hateful meme classification methods in detecting hateful memes with unseen target types. To the best of our knowledge, this is the first paper to propose hateful meme detection in zero-shot settings.
(2) We propose a deep generative framework, TAME, which is a complementary module that could be incorporated into existing hateful meme classification methods to enhance their performance in normal and zero-shot settings.
(3) We conducted extensive experiments on the Facebook hateful meme dataset and demonstrated TAME's effectiveness in improving the state-of-the-art hateful meme classification methods' performance in normal and zero-shot learning settings.

[1]Code are available at https://gitlab.com/bottle_shop/safe/TAME

2 RELATED WORK

2.1 Hate Speech Detection.

Automated hate speech detection in social media is a widely studied research area that have received quite a lot of attention in recent years. Several text-based hate speech detection datasets [8, 27, 34] have been released. Previous works exploit both machine learning based methods [6, 7, 26, 33, 36] and deep learning based techniques to detect hate speech in social media [1, 2, 5, 10, 11, 16, 22, 25, 37]. The existing automated hate speech detection method has yielded good performance. However, most of the existing studies have focused on text-based hateful content, neglecting the rich multimedia user-generated content.

2.2 Multimodal Hateful Meme Detection.

The flourish of multimodal hateful meme detection studies could be attributed to the availability of several hateful memes datasets published in recent year [14, 17, 30]. For instance, Facebook had proposed the *Hateful Memes Challenge*, which encouraged researchers to submit solutions to perform hateful memes classification [17]. A dataset consisting of 10K memes was published as part of the challenge. The memes are specially constructed such that unimodal methods cannot yield good performance in this classification task. Therefore, existing studies have adopted multimodal approaches to perform hateful memes classification.

The existing multimodal hateful memes detection approaches can be broadly categorized into two groups: (a) models that adopt early fusion techniques to concatenate text and visual features for classification [17, 30], and (b) models that directly fine-tune large scale pre-trained multimodal models [17, 20, 23, 32]. Recent studies have also attempted to use data augmentation [40] and ensemble methods [29, 32] to enhance the hateful memes classification performance and explaining the classification results **??**. Nevertheless, the existing hateful meme classification methods heavily rely on labeled data that are not always available in the real-world setting. This paper aims to fill this research gap by evaluating existing hateful meme classification methods' performance in detecting hateful memes that target unseen protected characteristics (i.e., zero-shot setting). A novel deep generative framework, TAME, is also proposed to enhance the state-of-the-art methods' hateful meme detection performance in normal and zero-shot settings.

2.3 Multimodal Zero-Shot Learning.

Multimodal Zero-shot Learning [4, 35, 39] is an emerging research area. Multimodal Zero-shot Learning aims to learn high-quality multimodal representation for classification or recognition and transfer knowledge from seen categories to unseen ones. A common multimodal zero-shot learning approach typically utilizes an auxiliary semantic space, where each sample has a particular semantic representation [4, 35, 39]. This paper extends the applications of multimodal zero-shot learning to a new task on zero-shot multimodal hateful meme classification.

2.4 Deep Generative Model.

Deep generative models have shown great potential for data generation [15, 18] . Popular Deep generative models include Variational Autoencoder (VAE) [18] and Generative Adversarial Network (GAN) [15]. VAE directly models the relationship between real and reconstructed data through element-wise reconstruction but has limited ability to complex data structures [3]. GAN is capable of capturing global information, but the training process is not stable [19]. To address these limitations, recent works have explored combining these two generative models [3, 19]. Inspired by these generative methods, our proposed TAME framework adopts a joint VAE-GAN generative architecture to learn target-aware representations that are generalizable to hateful memes with unseen target types.

3 PROPOSED MODEL

3.1 Problem Statement

We propose a zero-shot learning setting for hateful meme classification. Specifically, the setting assumes a specific target type (i.e., protected characteristic) is unseen during targeting. Let $\mathfrak{D}_s = (x_s, y_s, s_s)|x_s \in X_s, y_s \in Y_s, s_s \in S_s$ be a collection of seen set, where $x_s \in \mathbb{R}^D$ denotes the D-dimensional seen visual-linguistic features in the multi-model feature space X, $Y_s = \{0, 1\}$ represents the hateful label set (hateful and non-hateful) of all seen memes in the hateful label space Y. Each seen hateful meme also has the hateful target types (i.e., religion, nationality, etc.) labeled. The semantic features of the seen hateful target types are denoted as $s_s \in \mathbb{R}^{d_s}$, where d_s is the dimension in the semantic space S.

In the normal setting, hateful memes of all target types are seen in the training stage. However, in the zero-shot learning setting, only hateful memes of seen target types are leveraged in the training stage. For the testing stage, we explore two settings, namely, the *conventional zero-shot learning* and *generalized zero-shot learning* settings. In the conventional zero-shot learning test setting, the test set only contains hateful memes from unseen target types. Conversely, in the generalized zero-shot learning test setting, the test set includes hateful memes with both seen and unseen target types. We discuss the details of the training and test sets in Section 4.

3.2 Target-Aware Multimodal Enhancement

Figure 2 illustrates our proposed Target-Aware Multimodal Enhancement (TAME). The idea is to leverage on training memes with seen target types to synthesize highly general features that can subsequently be derived for test memes with unseen target types (hence without accompanying semantic information).

We first extract the visual-linguistic features x_s of the meme using a visual-linguistic model. Our framework accommodates any visual-linguistic model. In the training set, the hateful memes are also labeled with the hateful targets (e.g., religion). On these target labels, we use a pre-trained Word2Vec[13] model to extract the semantic features s_s, which we then input into a Variational AutoEncoder (VAE) to output the latent variable $\hat{z} = E(s_s) \in \mathbb{R}^d$. Concurrently, we sample $z \sim \mathbf{N}(0, 1) \in \mathbb{R}^d$, which is an arbitrary representation (i.e., noise) drawn from a Gaussian distribution. Both $\hat{z}$ and z are concatenated with the visual-linguistic features x_s and input into the generator of a Generative Adverserial Network (GAN). The generator outputs two types of target-aware representations; $s'_s = G(z, x_s)$, which is based on Gaussian noise and visual-linguistic features, and $\hat{s}_s = G(\hat{z}, x_s)$, which is based on the VAE output latent variable and visual-linguistic features. The generated target-aware representations and semantic features are input into an adversarial discriminator so as to guide the generation of s'_s and $\hat{s}_s$ to resemble the ground truth semantic features. Additionally, we use a pre-trained hate-target classifier as supervisory signal to ensure that s'_s is capturing features relating to the observed hate target. Finally, the target-aware representation s'_s is used to train a MLP hateful meme classifier.

Intuitively, through weight sharing and joint learning of $\hat{s}_s$ generation based on the VAE output, the generator is better able to endow s'_s with semantic information. This is core to the TAME framework because during testing, hateful memes with unseen targets do not have semantic features s_u for generating $\hat{s}_u$. Instead useful semantic information is propagated via s'_s, which we then combine with s'_u for hateful meme classification. Subsequently, we describe the key components of TAME framework in detail.

Joint VAE-GAN Generative Architecture. At the core of TAME framework is a VAE-GAN architecture, comprising an encoder, a generator, and an adversarial discriminator. The encoder captures the inherent attributes of semantic features and maps the semantic features into a latent space. The generator acts as the decoder of the VAE, decoding its latent variables and visual-linguistic features into the semantic feature space for reconstruction. The VAE structure minimizes the objective $\mathcal{L}_{VAE} = \mathcal{L}_{KL} + \mathcal{L}_{cyc}$, comprising the the Kullback-Leibler divergence loss $\mathcal{L}_{KL}$ and the cycle consistency loss L_{cyc}. $\mathcal{L}_{KL}$ is expressed as:

$$\mathcal{L}_{KL} = KL(p(\hat{z}|s; \theta_E)||\mathbf{N}(0, 1)) \tag{1}$$

where $\hat{z} \sim p(\hat{z}|s; \theta_E)$ is the latent representation generated by the encoder with parameters θ_E. Optimizing Eq. 1, minimizes the divergence between the distribution of the latent vector $\hat{z}$ and the unit Gaussian distribution $\mathbf{N}(0, 1)$. Thus the generator has to utilize latent vectors that are more generalizable and less specific to seen target types. Concurrently $\mathcal{L}_{cyc}$ ensures the decoded representation $\hat{s}$ is close to the ground truth semantic features s, written as:

$$\mathcal{L}_{cyc} = -\|p(\hat{s}|\hat{z}, x; \theta_G)\|_1 \tag{2}$$

where $\|\cdot\|_1$ is L_1 loss and θ_G are the generator parameters.

By optimizing $\mathcal{L}_{VAE}$, the encoder captures the inherent attributes of the ground truth semantic features, which helps the generator decode a target-aware representation that has similar distribution to the ground truth semantic features. Nevertheless, the VAE may not adequately capture high-level global information. For capturing both detailed and global information of input data, we further employ a GAN for the joint feature learning. The generator and the adversarial discriminator learn the features distribution using an adversarial learning architecture. Specifically, the generator synthesizes target-aware representation that has similar distribution to the ground truth semantic features using the visual-linguistic features and noise vector from Gaussian distribution as input. Finally, the adversarial discriminator $\mathcal{D}$ tries to distinguish the real semantic features (ground truth) and generated target-aware representations. We express the optimizing objective

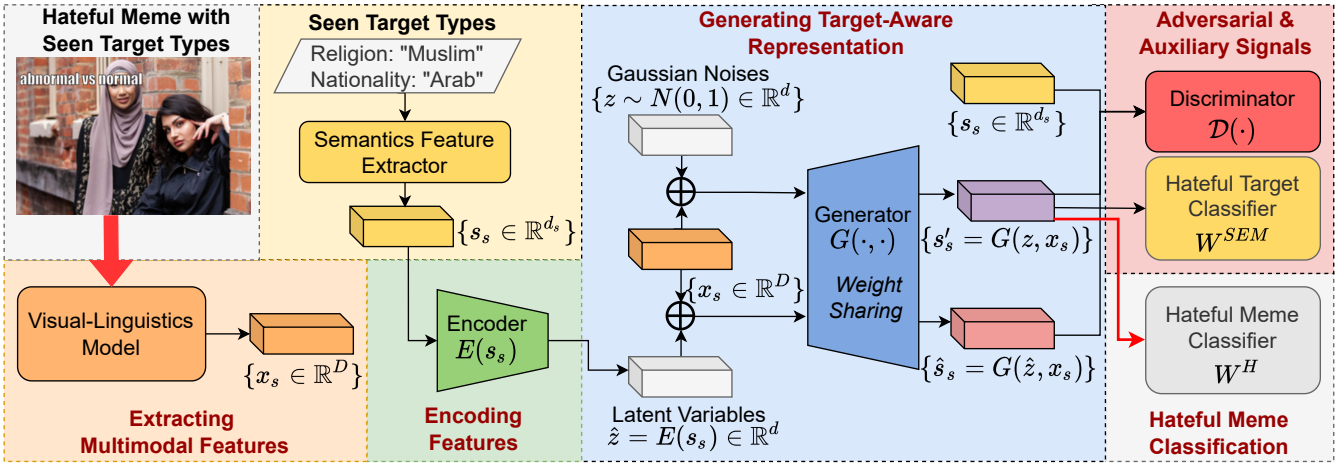


Figure 2: Target-Aware Multimodal Enhancement (TAME) Framework

of generator as follow:

$$\mathcal{L}_{adv} = -\mathcal{L}_{mse}[log\mathcal{D}(G(x,z))] - \mathcal{L}_{mse}[log\mathcal{D}(G(x,\hat{z}))] \quad (3)$$

where $\mathcal{L}_{mse}[\cdot]$ is Mean-Squared loss. Note that unlike $\hat{z}$, z is an arbitrary representation directly drawn from a Gaussian distribution, which is used as the input for the GAN along with the visual-linguistic features.

Finally, the adversarial discriminator outputs 1 when the input is a real semantic representation and outputs 0 when the input is a synthetic representation decoded by the generator. The following loss is minimized:

$$\begin{aligned}\mathcal{L}_{dis} &= -\mathcal{L}_{mse}[log\mathcal{D}(s)] - \mathcal{L}_{mse}[1 - log\mathcal{D}(G(x,z))] \\ &\quad - \mathcal{L}_{mse}[1 - log\mathcal{D}(G(x,\hat{z}))]\end{aligned} \quad (4)$$

Essentially, the generator in TAME decodes two types of target-aware representations. It generates the target-aware representation $\hat{s}_s$ based on the latent variable $\hat{z}$ from VAE and visual-linguistic features x_s, and it generates the target-aware representation s'_s based on the arbitrary representation z and visual-linguistic features x_s. Adopting a single generator to decode the two target-aware representations allows weight sharing across the two generation tasks. This enables the generator to capture rich information to synthesize realistic and discriminative features, which ultimately alleviates the domain shift problem and contributes to the zero-shot learning task.

Multi-Task Auxiliary Signal. To improve the learning of target-aware representation, we add an auxiliary supervised signal to ensure that s'_s captures features related to the hateful target semantics. Specifically, we adopt a multi-task learning approach to utilize the generated target-aware representation s'_s to train two separate classifiers for hateful meme and hateful target classification tasks. The first classifier W^{SEM} aims to preserve the network's capability of recognizing hateful targets in memes. The objective function of W^{SEM} is:

$$\mathcal{L}_{cls^{SEM}} = -\mathcal{L}_{ce}[logP(s|\hat{s},\theta_{SEM})] - \mathcal{L}_{ce}[logP(s|s',\theta_{SEM})] \quad (5)$$

where $\mathcal{L}_{ce}[\cdot]$ is the Cross Entropy loss and θ_{SEM} is the parameters of W^{SEM}. The second classifier W^H aims to predict the hateful label of a given meme. We define the learning objective of W^H as follows:

$$\mathcal{L}_{cls^H} = -\mathcal{L}_{ce}[logP(y|s',\theta_H)] \quad (6)$$

where θ_H is the parameters of W^H.

Joint Optimization. The final objective function is as follows:

$$\mathcal{L}_{gen} = \mathcal{L}_{KL} + \lambda_{cyc}\mathcal{L}_{cyc} + \mathcal{L}_{adv} + \lambda_{cls^H}\mathcal{L}_{cls^H} + \lambda_{cls^{SEM}}\mathcal{L}_{cls^{SEM}} \quad (7)$$

where λ_{cyc}, λ_{cls^H} and $\lambda_{cls^{SEM}}$ are hyper-parameters to control the relative contributions of different components.

3.3 Zero-Shot Hateful Meme Classification

After training the TAME framework, the generator can synthesize target-aware representations based on given visual-linguistic features. Given an arbitrary representation z drawn from a Gaussian distribution and the visual-linguistic feature of an hateful meme x_u with unseen target type, the target-aware representation of the hateful meme, s'_u, will be generated by the generator as follow:

$$s'_u = G(x_u, z) \quad (8)$$

Through this approach, the TAME generates the target-aware representation even though we do not have any semantic information of the unseen hateful target types. The generated s'_u contain class-level semantic information compared with unseen instance-level visual-linguistic features. Finally, in the testing stage, s'_u will be input to hateful classifier W^H for hateful meme classification.

4 EXPERIMENT

4.1 Experimental Settings

Dataset. We train and evaluate TAME on the popular Facebook hateful meme detection challenge dataset [17]. The original dataset contains about 10K memes with binary labels (i.e., non-hateful and hateful). Recently, Facebook also released a fine-grain version[2] of the dataset with the target types of the hateful meme annotated (i.e.,

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

Table 1: Datatset splits for various settings.

Split	Train	Test (GZSL)	Test (ZSL)
All seen (Normal)	non-hateful: 5,495 hateful: 2616	non-hateful: 254 hateful: 216	-
Race unseen	non-hateful: 5,495 hateful: 1,815		non-hateful: 254 hateful: 254
Disability unseen	non-hateful: 5,495 hateful: 2,432		non-hateful: 254 hateful: 200
Nationality unseen	non-hateful: 5,495 hateful: 2,425		non-hateful: 254 hateful: 210
Religion unseen	non-hateful: 5,495 hateful: 1,728		non-hateful: 254 hateful: 254
Sex unseen	non-hateful: 5,495 hateful: 2,064		non-hateful: 254 hateful: 254

race, disability, nationality, sex, religion). Existing studies mostly adopted the normal dataset split proposed in [17], where all target types of hateful memes are seen in the training set. We extend the normal dataset split setting to the conventional and generalized zero-shot learning settings. Table 1 shows the train-test splits for the different experimental settings. Specifically, when testing for an unseen hateful target type, hateful memes with the same target type are removed in the training set. For example, when testing for *Race* being the unseen target type, hateful memes targeting races are removed in the training set.

For the generalized zero-shot setting, we created one test set which contains both the non-hateful and the hateful memes with all target types (including the unseen target type). For the conventional zero-shot setting, we created multiple test sets. Each test set contains non-hateful memes and hateful memes with the specific unseen target type. Noted that we deliberately balance the test sets for the conventional zero-shot test setting. Evaluation on imbalanced test sets is more challenging and intended for future work.

Baselines. As TAME is a model-agnostic framework, we implemented our proposed framework on three commonly used visual-linguistic models, namely, VisualBERT [21], VilBERT [24], and LXmert [31]. We also benchmark the implementations against these base models without TAME enhancement.

Implementation Details We implemented all experiments using PyTorch[28]. We extract the RoI visual features for meme images using the pre-trained Faster R-CNN object detector[12]. The OCR textual information of memes is tokenized using BERT[9] default tokenizer. The semantic features of seen classes are extract via Word2Vec[13] model with dimension size $R_{d_s} = 300$. We have also attempted to leverage pre-trained language models such as BERT to extract the semantic features. However, we found that Word2Vec was able to yield better performance. During training, the dimension of visual-linguistic features extracted by pre-trained VisualBERT and LXmert is $R_D = 768$, while for pre-trained VilBERT, the dimension of visual-linguistic features is $R_D = 1024$. We adopt AdamW and SGD as optimizers of generative model and discriminator, respectively. We randomly chose five seeds and computed the average of all experimental results. The hyperparameter λ_{cyc}, λ_{cls^H} and $\lambda_{cls^{SEM}}$ are 10.0, 5.0 and 5.0, respectively.

Evaluation Metrics. We adopt the evaluation metrics proposed in the Facebook hateful meme challenge [17]. Specifically, for all experiment settings, we report the models' Area Under the Receiver Operating Characteristic curve (AUROC) and accuracy scores (Acc).

Table 2: AUROC & Acc for normal and generalized zero-shot learning setting.

Setting	Metric	VisualBERT		VilBERT		LXmert	
		Base	+TAME	Base	+TAME	Base	+TAME
All seen (normal)	AUROC	67.43	**69.89**	70.99	**73.58**	65.10	**67.08**
	Acc	62.77	**64.11**	65.32	**67.02**	62.49	**64.40**
Race unseen	AUROC	62.99	**64.33**	69.18	**70.20**	61.85	**62.45**
	Acc	59.68	**61.60**	64.89	**66.60**	61.70	**62.13**
Disability unseen	AUROC	66.65	**67.35**	69.54	**70.83**	63.34	**65.65**
	Acc	59.47	**62.34**	64.04	**66.81**	60.21	**62.55**
Nationality unseen	AUROC	66.80	**68.10**	69.47	**72.49**	65.97	**66.55**
	Acc	61.60	**62.85**	64.04	**66.81**	62.18	**62.98**
Religion unseen	AUROC	60.14	**63.89**	63.79	**68.00**	60.81	**63.60**
	Acc	58.73	**59.47**	61.91	**63.19**	58.94	**59.36**
Sex unseen	AUROC	63.75	**65.02**	69.71	**71.60**	63.04	**63.85**
	Acc	60.00	**61.32**	63.83	**65.74**	60.43	**61.06**

Table 3: AUROC & Acc for conventional zero-shot learning settings.

Setting	Metric	VisualBERT		VilBERT		LXmert	
		Base	+TAME	Base	+TAME	Base	+TAME
Race unseen	AUROC	51.68	**54.80**	42.16	**42.18**	46.06	**49.70**
	Acc	50.39	**51.57**	48.23	**49.80**	48.43	**50.20**
Disability unseen	AUROC	48.73	**49.99**	47.38	**53.52**	50.45	**53.64**
	Acc	53.52	**54.64**	56.83	**58.13**	54.85	**55.08**
Nationality unseen	AUROC	56.06	**57.84**	58.22	**59.62**	53.93	**60.72**
	Acc	57.76	**58.33**	56.90	**59.05**	54.74	**59.48**
Religion unseen	AUROC	56.98	**58.69**	58.03	**58.34**	56.30	**58.30**
	Acc	50.79	**54.53**	54.72	**56.13**	51.77	**52.57**
Sex unseen	AUROC	47.38	**48.78**	38.14	**44.79**	50.53	**53.38**
	Acc	45.87	**47.24**	45.87	**48.62**	51.57	**52.95**

4.2 Experiment Results

Comparison with baselines. Table 2 shows the AUROC and Acc results for the normal setting and the generalized zero-shot learning setting. The highest figures are highlighted in **bold**. Overall, we observe that the visual-linguistic models enhanced with TAME outperform the baseline models in both normal and generalized zero-shot learning settings. More specifically, TAME improved the baseline models by about 2% in all experiment settings. The improved performance demonstrated TAME's strength in detecting hateful memes with seen and unseen target types. The improvement is consistent across all baseline models and demonstrated TAME's adaptability to different models. This also suggests that TAME has the ability to synthesise useful general features for hateful meme classification.

Table 3 shows the experiment results for the conventional zero-shot learning settings. Note that this is a much more challenging setting than the generalized zero-shot setting since all test memes are of unseen target types. Attesting to the difficulty of the task, all models suffer great drops in performance, with some results being close to random guess. Nonetheless, our focus is on whether TAME is able to improve the performance of the model it is applied on.

We observe that TAME significantly improves visual-linguistic baselines' performance in both evaluation metrics. TAME improved the baseline models by about 3% in both AUROC and Acc when detecting hateful memes with various unseen target types. As we are interested in the models' performance in detecting hateful memes, we further examine their performance by computing the *Recall* for recovering hateful memes in the conventional zero-shot test sets. The results are shown in Figure 3. Interestingly, we observe that the visual-linguistic models enhanced with TAME are able to detect hateful memes with unseen target types better. For instance, TAME is able to enhance the Recall of VilBERT by about 12% when detecting unseen nationality-related hateful memes. TAME's promising zero-shot learning results can be attributed to the learned

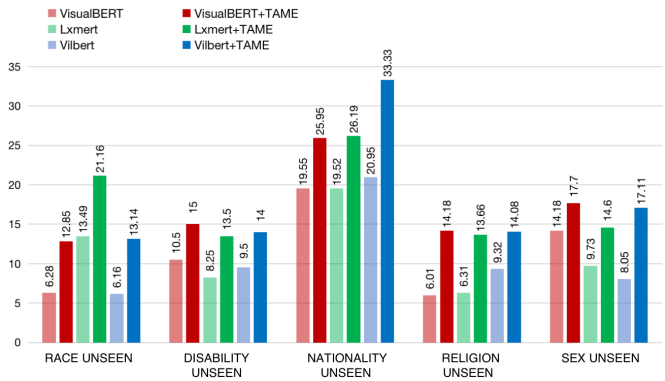


Figure 3: Recall for hateful meme with unseen target types

Table 4: Ablation study in normal setting and generalized zero-shot learning setting, *wo* indicates that the specific loss or modules are excluded during training.

Description	Normal		GZSL	
	AUROC	Acc	AUROC	Acc
Base model (*wo TAME*)	70.99	65.32	69.47	64.04
TAME (*wo* $L_{cls^{SEM}}$)	64.15	62.13	65.63	63.62
TAME (*wo* L_{cyc})	72.07	65.81	71.36	66.02
TAME (*wo* L_{cls^H})	57.50	60.43	64.31	61.48
TAME (*wo* L_{adv})	70.73	65.23	70.37	64.89
TAME (*wo* VAE)	70.42	65.38	70.19	66.66
TAME (ALL)	**73.58**	**67.02**	**72.49**	**66.81**

target-aware representation, which may have captured more salient features of hateful memes that could be generalized to detect hateful memes of unseen target types.

Ablation Study As there are a number of components in our TAME framework, we conduct an ablation study to examine the contribution of each component to the hateful meme classification task. Table 4 show the results of the ablation study on normal and generalized zero-shot learning settings. VilBERT is used as the base model and we select nationality as the unseen hateful target type. We observe that TAME with all components included yield the best performance in both settings. Interestingly, we noted that without $L_{cls^{SEM}}$, the model's performance decreases drastically as the model is unable to recognize the hateful targets. The VAE and L_{adv} also play important roles in the TAME framework's good performance by generating target-aware representations.

Visualization of the Learned Representations A key characteristic of the TAME is that the model is able to generate target-aware representations. We postulate that the learned target-aware representation should represent the memes better and ultimately improve hateful meme classification. To empirically examine the quality of the learned target-aware representations, we visualize the learned representations and compare them with the features extracted from the visual-linguistic model. Figure 4 (a) and (b) show the t-SNE visualization of visual-linguistic features and target-aware representations learned by VilBERT and VilBERT+TAME models in normal setting, respectively. The points in grey represent the non-hateful memes, while the other colored points represent the hateful memes with different target types. We observe that the hateful memes are clustered better according to their hateful target types with the target-aware representations. The hateful and non-hateful memes are also more well separated in the target-aware representations than the visual-linguistic features.

We further examine the difference between visual-linguistic features and target-aware representations in the zero-shot learning setting where we select nationality as the unseen hateful target. Figure 4 (c) and (d) show the t-SNE visualization of visual-linguistic features and target-aware representations learned by VilBERT and VilBERT+TAME models in zero-shot learning setting, respectively. Similarly, we observe that the hateful and non-hateful memes are more well separated in the target-aware representations than the visual-linguistic features. Interestingly, the target-aware representations of the hateful memes that target nationalities are also clustered closer together in the target-aware representations even though such memes are not observed in training.

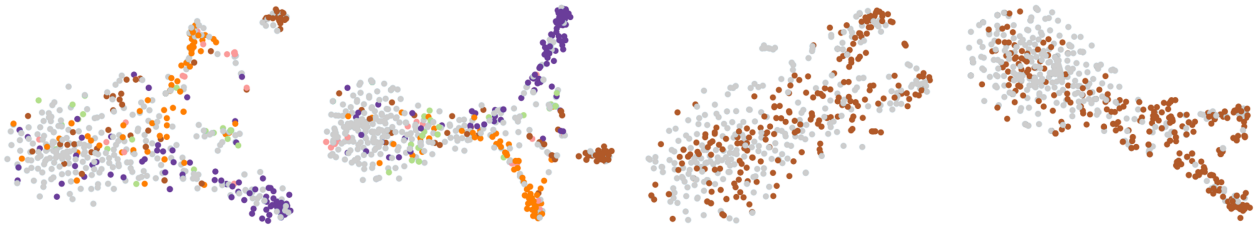
Case Studies We empirically examine some of the zero-shot hateful meme classification results. Table 5 shows some examples of hateful memes with various unseen target types that are detected by VilBERT+TAME but missed out by the VilBERT model. A potential reason for VilBERT+TAME good performance may be attributed to TAME's ability to learn general or transferable knowledge in the target-aware representations. Specifically, we notice that the first meme targeting religion contains an image of guns, which may also appear in hateful memes with other target types. Similarly, the second word contains hateful works such as "f*ck", which may be used in other hateful memes targeting seen types. These common hateful attributes might be better learned in the target-aware representations, thus improving its performance. Conversely,

Table 6 shows the example memes that are wrongly classified by both VilBERT+TAME and VilBERT. These memes seem to require specific context on the target types themselves, making it challenging for TAME to correctly classify them in the zero-shot setting. For instance, the first meme in Table 6 is expressing prejudice against the Catholic faith, specifically on their tithing practises. Detecting hateful memes will require additional context that goes beyond identifying targets in memes.

Finally, we also provide examples that are wrongly classified by VilBERT+TAME but correctly classified by the VilBERT model in Table 7. Similarly, these memes seem to require specific context on the target types, and the limited semantic information caused in TAME to make incorrect classifications. For instance, the first meme on the left in Table 7 is targeting "*Muslims*" by using a slur that suggests an unnatural sexual relationship with goats. However, VilBERT+TAME did not have the contextual knowledge to make the right prediction. The base model might have made correct random guesses in these examples; we observed that it made incorrect predictions in similar unseen target types memes in Table 6.

5 DISCUSSION AND TAKEAWAYS

Our extensive experiments show that TAME is able to enhance the hateful meme classification performance of visual language models in both conventional and generalized zero-shot settings. Although the improvement on the models enhanced with TAME seem small, we noted that the hateful meme detection is performed under very challenging settings. The analysis on recall for hateful memes with unseen target types also show that models enhanced with TAME



(a) (b) (c) (d)

Figure 4: t-SNE visualization of visual-linguistic features and target-aware representations learned by VilBERTand VilBERT+TAME models in normal setting (a & b) and zero-shot learning setting (c & d).

Table 5: Correctly classified hateful memes with different unseen target types

	Religion Unseen	Nationality Unseen	Sex Unseen	Race Unseen	Disability Unseen
					
Groundtruth	Hateful	Hateful	Hateful	Hateful	Hateful
VilBERT	Non-Hateful	Non-Hateful	Non-Hateful	Non-Hateful	Non-Hateful
VilBERT +TAME	Hateful	Hateful	Hateful	Hateful	Hateful

Table 6: Incorrectly classified hateful memes with different unseen target types

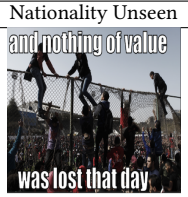
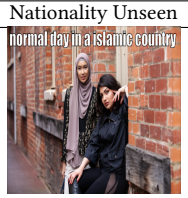
	Religion Unseen	Nationality Unseen	Sex Unseen	Race Unseen	Disability Unseen
					
Groundtruth	Hateful	Hateful	Hateful	Hateful	Hateful
VilBERT	Non-Hateful	Non-Hateful	Non-Hateful	Non-Hateful	Non-Hateful
VilBERT +TAME	Non-Hateful	Non-Hateful	Non-Hateful	Non-Hateful	Non-Hateful

Table 7: Incorrectly classified hateful memes with different unseen target types, while base model classified correctly

	Religion Unseen	Nationality Unseen	Sex Unseen	Race Unseen	Disability Unseen
					
Groundtruth	Hateful	Non-Hateful	Hateful	Hateful	Hateful
VilBERT	Hateful	Non-Hateful	Hateful	Hateful	Hateful
VilBERT +TAME	Non-Hateful	Hateful	Non-Hateful	Non-Hateful	Non-Hateful

have significant improvement over the base models, demonstrating TAME's ability to enhance the hateful meme detection by increasing true positives (i.e., correctly flagging hateful memes). The t-SNE visualization (Figure 4) demonstrated TAME's ability to learn better meme latent representation where the memes sharing similar target types are observed to cluster closer in the embedding space.

Nevertheless, through our case studies, we identify the limitations of TAME; contextual information is still lagging for flagging

certain hateful memes that require context and reasoning beyond target identification. Furthermore, the zero-shot learning setting may also lead to overfitting to seen hateful categories. The overfitting issue may be significantly reduced in with few-shot learning. This leaves room for future research to investigate and innovate better solutions to tackle hateful meme classification in zero-shot settings and explore hateful meme detection in other settings.

6 CONCLUSION

This paper proposed a zero-shot hateful meme detection task to evaluate existing hateful meme classification methods' performance in detecting hateful memes with unseen target types. To the best of our knowledge, this is the first paper to propose hateful meme detection in zero-shot settings. To address the newly proposed task, we also proposed a deep generative framework, TAME, which enhances existing hateful meme classification methods' performance in normal and zero-shot settings. Extensive experiments were conducted using the Facebook hateful meme dataset, and the results demonstrated TAME's effectiveness in improving the state-of-the-art hateful meme classification methods' performance in normal and zero-shot learning settings.

For future work, we aim to consider more fine-grain unseen targets in zero-shot hateful meme classification tasks. For instance, we can differentiate the different targets (e.g., "*Chinese*", "*Indian*") in each target types (e.g., "*Nationality*"). We will also explore more advanced methods to perform conventional and generalized zero-shot hateful meme detection.

REFERENCES

[1] Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2021. Angrybert: Joint learning target and emotion for hate speech detection. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 701–713.

[2] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 759–760.

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. 2017. CVAE-GAN: fine-grained image generation through asymmetric training. In *IEEE ICCV*. 2745–2754.

[4] Nihar Bendre, Kevin Desai, and Peyman Najafirad. 2021. Generalized Zero-Shot Learning Using Multimodal Variational Auto-Encoder With Semantic Concepts. In *IEEE ICIP*. 1284–1288.

[5] Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. DeepHate: Hate speech detection via multi-faceted text representations. In *12th ACM Conference on Web Science*. 11–20.

[6] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*. 13–22.

[7] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 71–80.

[8] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[10] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. 29–30.

[11] Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*. 85–90.

[12] Ross Girshick. 2015. Fast R-CNN. arXiv:1504.08083 [cs.CV]

[13] Yoav Goldberg and Omer Levy. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722* (2014).

[14] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *IEEE ICCV*. 1470–1478.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *NeurIPS* 27 (2014).

[16] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All You Need is" Love" Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. 2–12.

[17] 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* (2020).

[18] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv* (2013).

[19] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. In *ICML*. PMLR, 1558–1566.

[20] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling Hate in Online Memes. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5138–5147.

[21] 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* (2019).

[22] Ken-Yu Lin, Roy Ka-Wei Lee, Wei Gao, and Wen-Chih Peng. 2021. Early Prediction of Hate Speech Propagation. In *2021 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 967–974.

[23] 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* (2020).

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv* (2019).

[25] Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words?. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 299–303.

[26] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.

[27] Ji Ho Park and Pascale Fung. 2017. One-step and Two-step Classification for Abusive Language Detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*. 41–45.

[28] Adam Paszke, Sam Gross, and Francisco Massa et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703 [cs.LG]

[29] Vlad Sandulescu. 2020. Detecting Hateful Memes Using a Multimodal Deep Ensemble. *arXiv* (2020).

[30] Shardul Suryawanshi and Chakravarthi et al. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. 32–41.

[31] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv* (2019).

[32] Riza Velioglu and Jewgeni Rose. 2020. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge. *arXiv* (2020).

[33] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142.

[34] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.

[35] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. 2018. Feature Generating Networks for Zero-Shot Learning. In *IEEE CVPR*.

[36] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 1980–1984.

[37] Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*. Springer, 745–760.

[38] Yi Zhou and Zhenhao Chen. 2020. Multimodal Learning for Hateful Memes Detection. *arXiv* (2020).

[39] Jiawen Zhu, Xing Xu, Fumin Shen, Roy Ka-Wei Lee, Zheng Wang, and Heng Tao Shen. 2020. Ocean: A Dual Learning Approach For Generalized Zero-Shot Sketch-Based Image Retrieval. In *IEEE ICME*. 1–6.

[40] Ron Zhu. 2020. Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution. *arXiv* (2020).