10-2021

Disentangling hate in online memes

Ka Wei, Roy LEE
Singapore Management University, roylee@smu.edu.sg

Rui CAO
Singapore Management University, ruicao.2020@phdcs.smu.edu.sg

Ziqing FAN

Jing JIANG
Singapore Management University, jingjiang@smu.edu.sg

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

Citation

1

Disentangling Hate in Online Memes

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

Rui Cao*
ruicao.2020@phdcs.smu.edu.sg
Singapore Management University
Singapore, Singapore

Ziqing Fan*
ziqingfan0331@gmail.com
University of Electronic Science and
Technology of China
Chengdu, China

Jing Jiang
jingjiang@smu.edu.sg
Singapore Management University
Singapore, Singapore

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

ABSTRACT

Hateful and offensive content detection has been extensively explored in a single modality such as text. However, such toxic information could also be communicated via multimodal content such as online memes. Therefore, detecting multimodal hateful content has recently garnered much attention in academic and industry research communities. This paper aims to contribute to this emerging research topic by proposing DisMultiHate, which is a novel framework that performed the classification of multimodal hateful content. Specifically, DisMultiHate is designed to disentangle target entities in multimodal memes to improve the hateful content classification and explainability. We conduct extensive experiments on two publicly available hateful and offensive memes datasets. Our experiment results show that DisMultiHate is able to outperform state-of-the-art unimodal and multimodal baselines in the hateful meme classification task.

CCS CONCEPTS

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

KEYWORDS

hate speech, hateful memes, multimodal, social media mining

*Both authors contributed chiefly and equally to this research. The author order is incorrect but not revisable due to conference policy.

Disclaimer: *This paper contains violence and discriminatory content that may be disturbing to some readers. Specifically, Figures 1 and 2 and Tables 6 and 7 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.*

1 INTRODUCTION

The proliferation of social media has enabled users to share and spread ideas at a prodigious rate. While the information exchanges in social media platforms may improve an individual's sense of connectedness with real and virtual communities, these platforms are increasingly exploited for the propagation of hateful content that attacks or uses discriminatory languages targeting a person or a group based on their race, religion, gender, etc. [17, 43]. The spread of online hate speech has sowed discord among individuals or communities online and resulted in violent hate crimes. Therefore, it is a pressing issue to detect and curb online hateful content.

The existing research on automated hateful content detection has predominantly focused on text-based content [17, 43], neglecting multimodal content such as memes. *Memes*, which are images with short texts, are a popular communication form in online social media. Although the memes are typically humorous in nature, they are also increasingly used to spread hate. 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 [20, 28, 47]. The hateful memes could also be a greater threat to social peace than text-based online hate speeches due to their viral nature; users could re-post or share these hateful memes in multiple conversations and contexts.

To combat the spread of hateful memes, social networking platforms such as Facebook had recently released a large hateful meme dataset as part of a challenge to encourage researchers to submit solutions to perform hateful memes classification [28]. The research community has answered this challenge by exploring and proposing multimodal classification models [12, 31, 37, 49, 54, 57]. Several studies had proposed solutions that focused on innovating the fusion techniques to combine text and visual features extracted from

Figure 1: Example of hateful meme.

memes to perform the classification tasks [28, 47]. Others have explored fine-tuning large-scale pre-trained multimodal methods to perform hateful memes classification [12, 28, 31, 37, 49, 54, 57]. Nevertheless, these existing methods have limited explainability and cannot reason the context embedded in the hateful memes.

Hateful meme classification is a challenging 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 meme is the context illustrated in the memes, and often the target entities (e.g., race, religion, etc.) in the hateful message provide important contextual information. For the example in Figure 1, the target entity of the hateful content would be both gender and religion (i.e., female Muslim). Existing multimodal hateful meme classification models are unable to capture such target entity contextual information.

This paper aims to address the research gaps by proposing a novel framework, DisMultiHate, which learns and disentangles the representations of hate speech-related target entities, such as race and gender, in memes to improve the hateful content classification. Our framework includes a novel self-supervising training task that enables us to extract the target entities using disentangled latent representations. The disentangled representations serve as contextual information to improve hateful meme classification.

We summarize this paper's contribution[1] as follows: (i) We propose a novel multimodal hateful meme classification model called DisMultiHate, which disentangles the representations of hate speech-related target entities to improve performance and explainability of hateful meme classification. (ii) We conduct extensive experiments on two publicly available datasets. Our experiment results show that DisMultiHate consistently outperforms state-of-the-art methods in the hateful meme classification. (iii) We conduct case studies and demonstrated DisMultiHate's ability to identify target entities in hateful memes, thereby providing contextual explanations for the classification results.

2 RELATED WORK

2.1 Hate Speech Detection
With the proliferation of social media and social platforms, automatic detection of hate speech has received considerable attention from the data mining, information retrieval, and natural

language processing (NLP) research communities. Several text-based hate speech detection datasets [13, 39, 51] have been released. Previous works exploit both machine learning based methods [8, 10, 38, 50, 52] and deep learning based techniques to tackle the problem [2, 7, 15, 18, 22, 36, 55]. 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 Hate Speech
To address the gap in hate speech detection research, recent works have attempted to explore multimodal hateful content classification tasks such as detecting online hateful memes [12, 28, 31, 37, 49, 54, 57]. The flourish of multimodal hateful meme detection studies could be attributed to the availability of several hateful memes datasets published in recent year [20, 28, 47]. For instance, Facebook had proposed the *Hateful Memes Challenge*[2], which encouraged researchers to submit solutions to perform hateful memes classification [28]. A dataset consists of 10K memes were published as part of the challenge, and the memes are specially constructed such that unimodal methods cannot yield good performance in this classification task. Therefore, existing studies have motivated to adopted a multimodal approach to perform hateful memes classification.

The existing multimodal hateful memes classification approaches can be broadly categorized into two groups: (a) models that adopt early fusion techniques to concatenate text and visual features for classification [28, 47], and (b) models that directly fine-tune large scale pre-trained multimodal models [12, 28, 31, 37, 49, 54]. Recent studies have also attempted to use data augmentation [57, 58] and ensemble methods [42, 49] to enhance the hateful memes classification performance. Nevertheless, these existing methods have limited explainability and cannot reason the context embedded in the hateful memes. For example, hate speech should involve abusive content targeted at an individual or a group [1, 6, 13]. However, most hateful memes classification methods cannot identify the hate targets and unable to provide the context for the hateful content.

This paper aims to address the research gaps by proposing a novel framework, DisMultiHate, which learns and disentangles the representations of hate speech-related target entities, such as race and gender, in memes to improve the hateful content classification. Unlike [58], which augment general entities information as input, we extract the target entities using disentangled latent representations learned using self-supervised training. The disentangled representations serve as contextual information to provide some form of explanation to the hateful meme classification.

2.3 Disentangling Representation Learning
Bengio et al. [3] defined a disentangled representation as one where single latent units are sensitive to changes in a single generative factor while being relatively invariant to changes in other factors. Hence, [3] proposed to factorize and learn disentangled representations by training independent units to encode different aspects of input data. Various methods have also been proposed to perform representation disentanglement. For instance, existing studies have attempted to learn disentangled representations using supervised

Table 1: Distributions of FHM and MultiOFF datasets

Dataset	train	validation	test
FHM	hateful (3050), non-hateful (5450)	hateful (250), non-hateful (250)	hateful (500), non-hateful (500)
MultiOFF	offensive (187), non-offensive (258)	offensive (58), non-offensive (91)	offensive (58), non-offensive (91)

signals [21, 24, 26]. Other works have explored purely unsupervised approaches to disentangled factor learning by exploiting the information-theoretic aspect of the variational auto-encoders and adding a regularization term that minimizes the mutual information between different units of the representations [5, 9, 23].

Latent representation disentanglement has been applied in various tasks. For example, representation disentanglement had been applied to learn and disentangle users' preferences for different categories of items to improve recommendation [34, 35]. Bouchacourt et al. [4] had exploited the disentangled representation learning to group data based on disentangled visual semantics. Dupont [16] had disentangled representations to learn factors that correspond to handwriting numbers' various characteristics. This paper contributes the state-of-the-art in disentangled representation learning by disentangling the representation of target entities in online hateful memes. To the best of our knowledge, this is the first work that applies representation disentanglement on the hateful memes classification task.

3 PRELIMINARIES

3.1 Problem Definition

We define the problem of hateful memes multimodal classification as follows: Given an image I and a piece of text O consisting of a sequence of words, a classification model will predict the label of the multimodal meme (*hateful* or *non-hateful*). This binary classification task can also be regarded as a regression task, where a model predicts a confidence score $y \in \mathbb{R}$ ranging from zero to one, indicating the likelihood of the meme being hateful. Specifically, the meme would be regarded as hateful if the predicted score is above a threshold λ; otherwise, the meme is predicted to be non-hateful.

3.2 Datasets

We train and evaluate our proposed model on two popular and publicly-available hateful datasets: *Facebook hateful memes (FHM* and *MultiOff*. Table 1 shows the distributions of the datasets.

Facebook Hateful Memes (FHM) [28]: The dataset was constructed and released by Facebook as part of a challenge to crowdsource multimodal hateful meme classification solutions. The dataset contains $10K$ memes with binary labels(i.e., hateful or non-hateful). The Facebook challenge did not release the labels of the memes in the test split. Therefore, we utilize the *dev-seen* split as the *test*. Existing studies have also adopted the same dataset setting in their training and evaluations [31, 57, 58].

MultiOFF [47]: The dataset contains $1K$ memes related to the 2016 United States presidential election. The memes are labeled as *offensive* or *non-offensive*. Although the authors have labeled the memes based on offensiveness, we manually examined the memes

and found that the offensive memes could also be considered hateful as most of them contain abusive and discriminatory messages against a person or minority group.

3.3 Data Pre-processing

To improve our proposed method's reproducibility, we also provide an overview of the data pre-processing steps applied on the FHM and MultiOFF memes datasets. Specifically, the following steps were taken to pre-process the memes in the datasets:

- **Image resizing**: The datasets provided memes in all sizes. We resized the images proportionally to a minimum of 140 pixels and a maximum of 850 pixels. This ensures consistency of the visual input into our proposed model and baselines.
- **Text extraction and removal**: We extract and remove the text in the memes using open-source Python packages Easy-OCR[3] and MMEditing[4]. The texts are removed from the memes to facilitate better entity and demographic detection.
- **Entity detection**: To augment the memes with relevant external information, we leverage Google Vision Web Entity Detection API[5] to detect and caption the entities in the cleaned image. The detected entities provide contextual information on the memes.
- **Demographic detection**: Often, hate speeches are targeted at groups based on demographic information such as race and gender, and such information servers as important contextual information in hateful meme classification. To augment the memes with demographic information, we utilized the FairFace classifier [27] to detect and classify the faces in the images, then mapping the label back to the person's bounding box with the largest overlapped area with the face. Noted that the demographic information is only extracted when the meme contains human entities.

The pre-processed datasets serve as input for the training of our proposed model discussed in the next section.

4 PROPOSED MODEL

Figure 2 illustrates the architectural framework of our proposed DisMultiHate model. Broadly, DisMultiHate consists of three main modules: (a) *data pre-processing*, (b) *text representation learning*, and (c) *visual representation learning*. The details of the data pre-processing module are discussed in Section 3.3. The goal of the text representation learning module is to learn a disentangled latent representation of the combined textual information output from the data pre-processing module. The details of the text representation learning module will be discussed in Section 4.1. The visual representation learning module aims to learn a disentangled latent representation based on the meme's image. The details of the visual representation learning module will be discussed in Section 4.2.

A core element in the two learning modules is the process of disentangling the target information from the text and visual representations. Specifically, the latent text and visual representations are projected into a disentangled latent space $\mathcal{D}$, where each latent unit of the disentangled representation represents a probability for

---

[3]https://github.com/JaidedAI/EasyOCR
[4]https://github.com/open-mmlab/mmediting
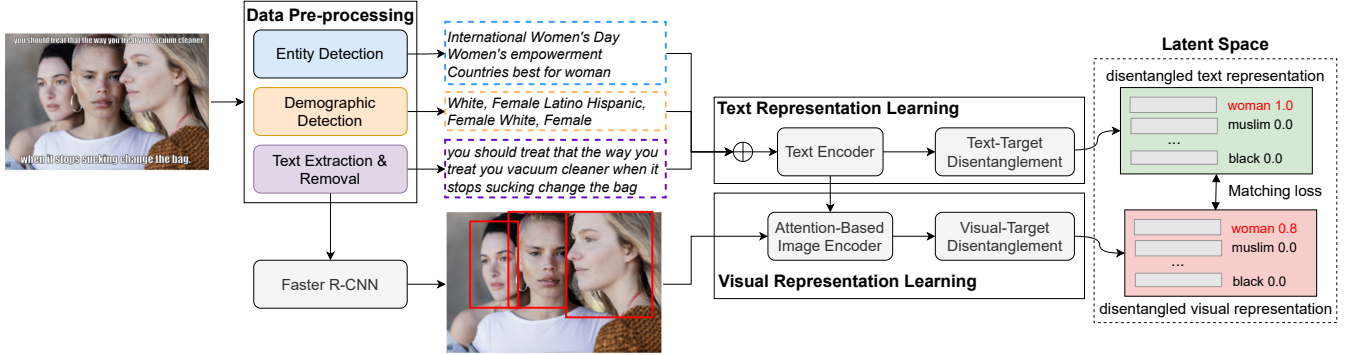[5]https://cloud.google.com/vision/docs/detecting-web

Figure 2: Architecture of our DisMultiHate model.

a certain category of hate (i.e., religion, gender and race, etc.). For a multimodal meme, disentangled targets from the textual modality and the visual modality should be consistent with each other. Therefore, we introduce a self-supervised matching loss, which constrains the disentangled visual representation to be similar to the disentangled text representation. Finally, the learned text and visual representations will be fed into a regression layer to predict the likelihood of the meme being hateful. The classification process will be discussed in Section 4.3.

4.1 Text Representation Learning

This module is designed to learn a disentangled latent representation of the textual information extracted from a meme. The input of the module is the concatenation of the text information output of the data pre-processing step. Specifically, we concatenated the text extracted from the meme, detected entities, and demographic information. Formally, we denote the concatenated text information as $\mathbf{O} = \{\mathbf{o}_j\}_{j=1}^{M}$, where $\mathbf{o}_j \in \mathbb{R}^{|\mathcal{V}|}$ is the one-hot vector representation for the j-th word in the text's word sequence, M is the length of the text, and $\mathcal{V}$ is the vocabulary.

Text Encoder. The concatenated text information $\mathbf{O}$ is first fed into a text encoder to generate latent text representations. Since the input text involves words from various domains such as religion, politics, or military, a powerful text encoder is required to capture the semantics in textual information. Bidirectional Encoder Representations from Transformers (BERT) [14], which has demonstrated its superiority in various natural language processing (NLP) tasks, is an ideal text encoder for our task. We initialize the BERT with pre-trained weights and fine-tune it with our task. Using the BERT text encoder, we generate the textual representations as follow:

$$[\mathbf{s}, \mathbf{C}] = \text{BERT}([\mathbf{w}_{[CLS]}, \mathbf{O}]), \qquad (1)$$

where $\mathbf{w}_{[CLS]} \in \mathbb{R}^{|\mathcal{V}|}$ denotes the one-hot representations for the "[CLS]" token, $[\cdot, \cdot]$ is the concatenation operation and $\mathbf{C} = \{\mathbf{c}_j\}_{j=1}^{M}$ is the set of textual representations, and $\mathbf{c}_j \in \mathbb{R}^{u}$ is the representation for the j-th word in the input text $\mathbf{O}$. Similar to [14], we utilize the representation of the "[CLS]" token as the sentence representation $\mathbf{s} \in \mathbb{R}^{u}$.

Text-target disentanglement. The latent text representation generated using the BERT encoder captures rich information on the

meme's semantics. However, for our hateful meme classification task, we are interested in contextual information present in the latent text representation, specifically the targets of the hateful content. For example, in Figure 2, we aim to identify "gender" as the target entity of the hateful message. Therefore, we need to design a mechanism to disentangle the target information in the latent text representation $\mathbf{s}$.

We first transform the the sentence representation $\mathbf{s}$ into a latent space using a linear projection layer:

$$\mathbf{s}_{\text{p}} = \mathbf{W}_{\text{s}}\mathbf{s} + \mathbf{b}_{\text{s}}, \qquad (2)$$

where $\mathbf{W}_{\text{s}} \in \mathbb{R}^{|\mathcal{D}| \times u}$ and $\mathbf{b}_{\text{s}} \in \mathbb{R}^{|\mathcal{D}|}$ are parameters to be learnt.

The goal of latent space disentanglement is to minimize the overlap of information among latent units in the vector. There are many methods that perform latent space disentanglement using regularization terms to minimize the mutual information between latent units [5, 9, 41]. For our task, we aim to disentangle the projected text representation such that each unit in the latent vector represents a type of hateful meme targets. Noted that we assume that each meme is likely to concentrate on a certain category of hate (i.e., religion, gender, race, etc.) in most cases. To achieve this objective, we adopt a similar approach to [35], where we maximize the likelihood of the target present in the latent text representation while minimizing for the absent targets. Such a discontinuous arg max operation can be fulfilled by applying *Straight-Through Gumbel-Softmax* (STGS) function [25] over $\mathbf{s}_{\text{p}}$. Specifically, a continuous vector $\mathbf{z} \in \mathbb{R}^{|\mathcal{D}|}$ is first sampled from the Gumbel-Softmax distribution based on $\mathbf{s}_{\text{p}}$:

$$u_k \sim \text{Uniform}(0, 1), \qquad (3)$$

$$g_k = -\log(-\log(u_k)), \qquad (4)$$

$$z_k = \frac{\exp(\log(s_{\text{p}}^k) + g_k)/\tau}{\sum_{k=1}^{||\mathcal{D}||} \exp(\log(s_{\text{p}})^k + g_k)/\tau}, \qquad (5)$$

where s_{p}^k is the k-th element in $\mathbf{s}_{\text{p}}$. In the forward pass, the STGS function then transforms the continuous vector sampled from the Gumbel-Softmax distribution into a one-hot vector [25]:

$$l_{\text{s}}^k = \begin{cases} 1 & k = \arg\min_m z_m \\ 0 & else \end{cases} \qquad (6)$$

Finally, the exploitation of STGS to generate the disentangled text representation can be simplified as:

$$\mathbf{l}_s = \text{Gumbel-Softmax}(\mathbf{s}_p), \qquad (7)$$

where $\mathbf{l}_s = \{s_p^k\}_{k=1}^{|\mathcal{D}|}$ as generated by Equation 3. For example, in Figure 2, the disentangled text representation would be a one-hot latent vector where 1 is assigned to the latent unit that represents 'woman' (i.e., target). The disentangled text representation $\mathbf{l}_s$ will be used in the self-supervised matching with the visual disentangled latent representation in section 4.2.

Via learning a disentangled text representation in the projected latent space, we update the text representation $\mathbf{s}$ through the back-propagation process, thus fine-tuning $\mathbf{s}$ to be more representative of the target information. We then use the updated $\mathbf{s}$ for hateful meme classification, as discussed in Section 4.3.

4.2 Visual Representation Learning

After learning the text representation, we focus on learning the disentangled latent representation in the visual modality. The input of this module is the image features pre-processed using Faster R-CNN [40]. Formally, we define the set of image features as $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{N}$, where $\mathbf{v}_i \in \mathbb{R}^d$ is the feature for the i-th detected region using *Faster R-CNN* [40] and N is the number of detected region.

Attention-Based Image Encoder To enable better interaction between visual and textual modality, we adopted the multi-head attention proposed in [48] to learn an attended latent visual representation of the meme. Specifically, we leverage the textual representations $\mathbf{C}$ generated by BERT text encoder as guidance to attend the feature map $\mathbf{V}$ and generate the attended visual representation using the attention-based image encoder. The attended visual representation is computed as follow:

$$\mathbf{F}_t = \text{softmax}\left(\frac{\mathbf{W}_{q,t}\mathbf{C}(\mathbf{W}_{k,t}\mathbf{V})^T}{\sqrt{q}}\right)\mathbf{W}_{v,t}\mathbf{V} \qquad (8)$$

$$\tilde{\mathbf{F}} = \text{Concate}(\{\mathbf{F}_t\}_{t=1}^{q}) \qquad (9)$$

$$\mathbf{G} = \mathbf{W}_{f,1}(\text{Relu}(\mathbf{W}_{f,2}\tilde{\mathbf{F}}) + \mathbf{b}_{f,2}) + \mathbf{b}_{f,1} \qquad (10)$$

$$\tilde{\mathbf{V}} = \tilde{\mathbf{F}} + \mathbf{G} \qquad (11)$$

$$\mathbf{v}_{att} = \sum_{m=1}^{M} \tilde{\mathbf{v}}_m, \qquad (12)$$

where q denotes the number of times the dot-attention is computed in the multi-head attention. Specifically, the t-th attended image feature $\mathbf{F}_t \in \mathbb{R}^{\frac{u}{q} \times M}$ is generated as shown in Equation 8, where $\mathbf{W}_{c,t} \in \mathbb{R}^{\frac{u}{q} \times u}$, $\mathbf{W}_{k,t} \in \mathbb{R}^{\frac{u}{q} \times d}$ and $\mathbf{W}_{v,t} \in \mathbb{R}^{\frac{u}{q} \times d}$ are parameters involved in the $t-th$ computation. The attended image features in different aspects are concatenated in row and generate $\tilde{\mathbf{F}} \in \mathbb{R}^{u \times M}$. Similar to [48], a residual connection is applied to the attended image features as illustrated in Equation 10 and 11. Finally, weighted average pool over the attended image features $\tilde{\mathbf{V}}$ results in the attended latent visual representation $\mathbf{v}_{att}$.

Visual-Target Disentanglement Although the attended visual representation is generated with an attention mechanism, there is no explicit guidance or supervision signal that forces the model to focus on the image regions that are more relevant to the contextual

information (i.e., target entities of hateful memes). For instance, in Figure 2, the visual representation should ideally focus on the image region with the three women and to be aware that the focused region infers the "gender" as the target of the hateful meme. To make the visual representation more "target-aware", we design a latent space matching mechanism, which aims to disentangle the target information from the visual representation and constraint the disentangled visual representation to be consistent with the disentangled latent text representation.

Similar to the text-target disentanglement, we first project the visual representation $\mathbf{v}_{att}$ into the latent space with a linear projection layer:

$$\mathbf{v}_p = \mathbf{W}_a \mathbf{v}_{att} + \mathbf{b}_a, \qquad (13)$$

where $\mathbf{W}_a \in \mathbb{R}^{|\mathcal{D}| \times u}$ and $\mathbf{b}_a \in \mathbb{R}^{|\mathcal{D}|}$ are parameters to be learnt.

To disentangle the target information in the visual representation, we introduce a matching loss that constraints $\mathbf{v}_p$ to be similar to the disentangled text representation $\mathbf{l}_s$:

$$\mathcal{L}_{\text{Match}} = \sum_{k=1}^{|\mathcal{D}|} l_s^k log(v_p^k) + (1 - l_s^k)log(1 - v_p^k), \qquad (14)$$

where v_p^k is the k-th unit in the visual latent representation $\mathbf{v}_p$. The matching loss serves as a supervision signal to disentangle the target information in the latent visual representation. The underlying intuition is that the disentangled text representation $\mathbf{l}_s$ is trained to disentangle the target in the textual information, and constraining the disentangled visual representation to be similar to the disentangled text representation would enable the target latent unit to be maximized in disentangled visual representation. For example, in Figure 2, the matching loss will take guidance from the disentangled text representation and enforce the disentangled visual representation to maximize the latent unit representing "gender".

Similarly to the text representation learning, the visual representation $\mathbf{v}_{att}$ would be updated through the back-propagation process, fine-tuning $\mathbf{v}_{att}$ to be more representative of the target information. The updated $\mathbf{v}_{att}$ would be used for the hateful meme classification discussed in Section 4.3.

4.3 Classification

To perform hateful meme classification, we leverage a regression layer to generate a *hateful score*. Specifically, if the hateful score is above a threshold, it will be regarded as hateful, otherwise non-hate. By learning the disentangled textual and visual representation and minimizing the matching loss between them in the latent space, the sentence representation $\mathbf{s}$ and the attended image feature $\mathbf{v}_{att}$ are both fine-tuned to be more representative of the target information. Finally, we concatenate $\mathbf{s}$ and $\mathbf{v}_{att}$, and feed the concatenated representation to a regression layer for the score prediction:

$$y = \sigma(\mathbf{w}_r^T[\mathbf{s}, \mathbf{v}_{att}] + b_r), \qquad (15)$$

where $\mathbf{w}_r \in \mathbb{R}^{2u}$ and $b_r \in \mathbb{R}$ are parameters to be learnt. Following [28], we set the threshold λ as 0.5: if the score y is above the threshold, it will be predicted as a hateful meme, otherwise, non-hate.

Loss Function. We optimize the following loss when training our model using mini-batch gradient descent:

$$\mathcal{L}_{\theta} = \mathcal{L}_{\theta,\text{Predict}} + \mu \mathcal{L}_{\theta,\text{Match}}, \qquad (16)$$

where θ denotes parameters of the model, $\mathcal{L}_{\theta,\text{Match}}$ is the matching loss in the disentangled latent space, computed from the sum of matching loss over all training samples (see Equation 14); and $\mathcal{L}_{\theta,\text{Predict}}$ is the loss from prediction and μ is the hyper-parameter balancing the relative importance of both loss types. The prediction loss is defined as:

$$\mathcal{L}_{\theta,\text{Predict}} = \sum_{s=1}^{|\mathcal{T}|} \hat{y}_s log(y_s) + (1 - \hat{y}_s)log(1 - y_s), \qquad (17)$$

where $\mathcal{T}$ is the training set and $\hat{y}_s$ is the ground-truth label and y_s is the predicted score of the s-th training sample.

4.4 Implementation Details

We set the dimension size to 2048 for image features extracted from Faster R-CNN [40]. All textual information is tokenized using BERT [14] default tokenizer. The length of the sequence is set to 64, and shorter sentences are padded while longer sentences are truncated. Weights in the BERT-based model are used to initialize the text encoder. The hyperparameter μ for balancing prediction loss and matching loss is set to be 0.05 on the FHM dataset and 0.03 on the MultiOFF dataset. All implementations are done in Pytorch, and we adopt AdamW [32] as the optimization algorithm to train our model. The size of the minibatch is set to 64.

5 EXPERIMENT

In this section, we will first describe the settings of experiments conducted to evaluate our DisMultiHate model. Next, we discuss the experiment results and evaluate how DisMultiHate fare against other state-of-the-art baselines. We also conduct case studies to qualitatively analyze DisMultiHate's ability to identify the targets of hateful memes. Finally, we perform error analysis and discuss the limitations of the DisMultiHate model.

5.1 Evaluation Setting

Dataset. We evaluate DisMultiHate and the baseline models on the two publicly available hateful memes classification datasets presented in Section 3.2. The train-validation-test split of the dataset are illustrated in Table 1.

Evaluation Metrics. We adopt the evaluation metrics proposed in the hateful meme dataset papers [28, 47]. Specifically, for the evaluation on FHM dataset [28], we use Area Under the Receiver Operating Characteristic curve (AUROC) and accuracy score as the evaluation metrics. Suryawanshi et al. [47] had only reported the F1, precision, and recall on the hateful class when they proposed the MultiOFF dataset [47]. However, we noted that due to class imbalance in hate speech classification, most existing studies [17, 43] have preferred to use weight metrics to evaluate the classification task. Thus, we adopt weighted F1, weighted precision, and weighted recall as the evaluation metrics for the MultiOFF dataset.

Baselines. We benchmark DisMultiHate against the state-of-the-art multimodal methods that were evaluated on the FHM and MultiOFF datasets. We have also included a unimodal baseline for

Table 2: Experimental results on FHM dataset.

Model	Acc.	AUROC
BERT (unimodal)	58.3	64.7
ViLBERT	62.2	71.1
VisualBERT	62.1	70.6
ViLBERT-CC	61.4	70.1
VisualBERT-COCO	65.1	74.0
ERNIE-VIL	69.0	78.7
UNITER	68.6	78.0
VILLNA	71.2	78.5
VL-BERT	71.4	78.8
DisMultiHate (w/o disentangle)	73.6	81.4
DisMultiHate	**75.8**	**82.8**

comparison. Specifically, we applied the pre-trained BERT [14] to extract text features from the meme's text and feed the extracted text features to a fully connected layer for classification.

For FHM dataset, we compare with the four best performing multimodal models reported in the original dataset paper [28], namely: **ViLBERT** [33], **ViLBERT-CC** (i.e., ViLBERT pre-trained on Conceptual Captions [44]), **VisualBERT** [29], and **VisualBERT-COCO** (i.e., VisualBERT pre-trained on COCO [30]). There are many interesting solutions proposed by the Facebook hateful memes classification challenge participants [31, 37, 54, 56–58]. For our evaluation, we benchmark against the methods explored by the top-performing participant[6]. Specifically, we benchmark against the methods proposed in Zhu's exploration [58]: **ERNIE-Vil** [53], **UNITER** [11], **VILLNA** [19], and **VL-BERT** [46]. We have reproduced the model using the code[7] released in [58]. We also adopt the same data augmentation method proposed in [58] to enhance the models. Specifically, all the reproduced models are augmented with entity tags retrieved using Google Vision Web Entity Detection, and **VL-BERT** is also further enhanced with demographic information extracted using FairFace [27].

For MultiOFF dataset, we compare with the multimodal methods reported in the dataset paper [47], namely: **StackedLSTM+VGG16**, **BiLSTM+VGG16**, and **CNNText+VGG16**. For these multimodal baselines, the researchers first extract the image features using VGG16 [45] pre-trained on the ImageNetdataset, and extract text features using various text encoders (e.g., BiLSTM). Subsequently, the extracted image and text features are concatenated before feeding into a classifier for hateful meme classification. As the MultiOFF dataset is relatively new, few studies have benchmarked on this dataset. Therefore, as additional baselines, we reproduced the methods proposed by Zhu [58] on the MultiOFF dataset.

5.2 Experimental Results

We have also included a variant of the DisMultiHate, which performed the hateful meme classification without disentangling the target information. Interestingly, we observe that even without the target disentanglement module, DisMultiHate had outperformed the baselines, demonstrating the strength of our data pre-processing approach on augmenting the model with entity and demographics information. DisMultiHate without target disentanglement has also outperformed the VL-BERT model, which was also augmented with

---

[6]https://ai.facebook.com/blog/hateful-memes-challenge-winners/
[7]https://github.com/HimariO/HatefulMemesChallenge

Table 3: Experimental results on MultiOFF dataset.

Model	F1	Precision	Recall
BERT (unimodal)	56.4	56.1	57.7
StackedLSTM+VGG16	46.3	37.3	61.1
BiLSTM+VGG16	48.0	48.6	58.4
CNNText+VGG16	46.3	37.3	61.1
ERNIE-VIL	53.1	54.3	63.7
UNITER	58.1	57.8	58.4
VILLNA	57.3	57.1	57.6
VL-BERT	58.9	59.5	58.5
DisMultiHate (w/o disentangle)	60.8	61.4	62.7
DisMultiHate	**64.6**	**64.5**	**65.1**

Table 4: Ablation study on FHM dataset.

Model	Acc.	AUROC
DisMultiHate (w/o Entity)	60.6	68.2
DisMultiHate (w/o Demo)	72.8	80.8
DisMultiHate (w/o Entity+Demo)	62.0	70.3
DisMultiHate	75.8	82.8

Table 5: Ablation study on MultiOFF dataset.

Model	F1	Precision	Recall
DisMultiHate (w/o Entity)	62.0	64.0	63.8
DisMultiHate (w/o Demo)	60.5	61.0	61.1
DisMultiHate (w/o Entity+Demo)	62.0	62.4	63.1
DisMultiHate	64.6	64.5	65.1

entity and demographics. A possible reason for the performance could be due to DisMultiHate's ability to learn better textual and visual representations for hateful meme classification. Specifically, the visual representation was attended with the textual information, thereby enhancing the visual features with some form of contextual information. Nevertheless, we noted that the performance of target disentanglement further improves the classification results, suggesting the importance of target information in the hateful meme classification task.

Table 2 and 3 show the experimental results on the FHM and MultiOFF datasets respectively. In both tables, the highest figures are highlighted in **bold**. We observed that DisMultiHate outperformed the state-of-the-art baselines in both datasets. DisMultiHate has significantly outperformed the baselines proposed in the original dataset papers. For instance, DisMultiHate has outperformed VisualBERT-COCO by more than 10% (Acc) on the FHM dataset and outperformed BiLSTM+VGG16 by more than 16% (F1) on the MultiOFF dataset. DisMultiHate has also outperformed the best baseline by 4% (Acc) and 5% (F1) on the FHM and MultiOFF, respectively. We noted that the multimodal methods had outperformed the BERT unimodal baselines in the FHM dataset. However, for the experiment on MultiOFF dataset, we observe that the BERT unimodal baseline is able to achieve competitive performance and outperformed the multimodal baselines proposed in the dataset paper [47]. A possible explanation could be the caption and text information in the MultiOFF memes already contain hateful content. Thus, the unimodal baseline using textual features is able to achieve good performance.

Ablation Study. We also conduct an ablation study to examine the usefulness of entity and demographic information augmented in our DisMultiHate method. Table 4 and 5 show the results of the ablation study on FHM and MultiOFF datasets, respectively. We observed that DisMultiHate model augmented with both entity and demographic information had yielded the best performance.

More interestingly, for the FHM dataset, we observed that without augmenting demographic information yields better performance than without augmenting entity information. However, a different observation is made for DisMultiHate performance on the MultiOFF dataset. Specifically, DisMultiHate not augmented with entity information yields better performance than without augmenting demographic information. The ablation study highlights that the model is highly amenable and adapts to different datasets with varying characteristics.

5.3 Case Study

The ability to disentangle target information in memes is a core contribution in our DisMultiHate model, and we aim to evaluate this aspect of the model qualitatively. Working towards this evaluation goal, we design an experiment to retrieve relevant memes for a given target query. Specifically, for a given text query (e.g., "woman"), we first generate its disentangled latent text representation, l_q, using the process described in Section 4.1. Next, we compute the cosine similarity between l_q and the disentangled latent visual representation v_p of all memes in the FHM dataset. Finally, we retrieved the top k memes with the highest similarity scores with the given target query. Intuitively, if DisMultiHate model is able to disentangle the target information in memes, we should be able to infer the query target from the retrieved memes qualitatively. For example, given the text query "woman", we should expect the top k retrieved memes to include woman-related memes.

Table 6 shows the retrieved memes for a given target. Specifically, we retrieve the two most relevant hateful and non-hateful memes for the given target query. We can intuitively infer that the retrieved memes are relevant to the given query. For example, the retrieved memes for the query "Muslim Woman" are observed to contain Muslim women in hajib. Interestingly, for the query "Black Man", we observe that the second meme is retrieved even though the image is in black and white, and it is not obvious that there are African Americans in the image. However, DisMultiHate is still able to disentangle the "Black Man" target in the meme by using relevant textual information such as "*dark*" and "*pick cotton*" to infer contextual information on the slavery of African American. A similar observation is observed for the "Woman" target query, where the second meme does not contain any image of a woman but an ape. However, the second meme is also relevant to the target query as DisMultiHate disentangle the "Woman" target in the meme by using relevant textual information such as "*Michelle Obama*" to infer contextual information on insulting the individual's physical appearance(i.e., a woman) with a picture of an ape. In summary, the case studies presented in Table 6 has demonstrated DisMultiHate's ability to disentangle the target in memes using a combination of textual and visual information captured in the memes. Similar observations were also made for other potential hate speech target queries (e.g., Hispanic, Asian, transgender, etc.).

Table 6: Retrieved memes from FHM dataset for a given target. The headers indicate the correctly predicted labels of the retrieved memes.

Target	Hateful		Non-Hateful	
Woman				
Black Man				

Table 7: Error analysis of wrongly classified memes from FHM dataset.

Meme				
Actual Label	Non-Hateful	Non-Hateful	Hateful	Hateful
Predicted Label	Hateful	Hateful	Non-Hateful	Non-Hateful
Disentangled Target	Woman	Catholics	Black Man	Muslim Woman

5.4 Error Analysis

Besides analyzing DisMultiHate quantitatively performance over the state-of-the-art baselines, we are also interested in examining the classification errors of DisMultiHate. Table 7 illustrates four selected examples of DisMultiHate's wrongly classified memes. For example, DisMultiHate has classified the first meme to be hateful when the actual label of the meme is non-hateful. A possible reason for this error could be the mention of the keyword "black" and the disentangled target being "woman", which misled the model to make a wrong prediction.

Our error analysis also reveals some issues with the FHM dataset. For instance, the second meme is annotated as non-hateful in the dataset. However, upon closer examination of the meme, we could infer some form of discrimination towards the Catholics and Priest, and our DisMultiHate has predicted the meme to be hateful. Another issue of the FHM dataset is the potential noise in the dataset. For example, DisMultiHate has wrongly classified the meme as non-hateful when the content is obviously communicating otherwise. We have checked the FHM dataset and found similar memes (i.e., a meme with a running black man) annotated as non-hateful.

DisMultiHate has also wrongly predicted the last meme to be non-hateful as none of the textual keyword, or image features provided the context information that it is hateful. Some form of advance reasoning would be required to understand the hateful context presented in this meme. We could explore adding advanced reasoning modules to classify such memes that require deeper reasoning for future work.

6 CONCLUSION

In this paper, we proposed a novel framework, DisMultiHate, which learns and disentangles the representations of hate speech-related target entities, such as race and gender, in memes to improve the hateful content classification. We evaluated DisMultiHate on two publicly available datasets, and our extensive experiments have shown that DisMultiHate outperformed the state-of-the-art baselines. We have conducted case studies to empirically demonstrated DisMultiHate's ability to disentangle target information in the memes. We have also performed error analysis and discussed some of the limitations of the DisMultiHate model. We will incorporate a more advanced reasoning module in the model for future works and test the model on more hateful meme datasets. Through applying DisMultiHate to disentangle the target in hateful memes, we also hope to raise awareness of the vulnerable groups targeted in hate speeches in real-world datasets.

ACKNOWLEDGEMENT

REFERENCES

[1] Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrovic. 2021. Angry-BERT: Joint Learning Target and Emotion for Hate Speech Detection. *arXiv preprint arXiv:2103.11800* (2021).

[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] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[4] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. 2018. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[5] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2018. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599* (2018).

[6] Cambridge. [n.d.]. *Hate Speech*. Retrieved February 11, 2020 from https://dictionary.cambridge.org/dictionary/english/hate-speech

[7] 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.

[8] 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.

[9] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. 2018. Isolating sources of disentanglement in VAEs. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2615–2625.

[10] 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.

[11] 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 *European Conference on Computer Vision*. Springer, 104–120.

[12] Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting Hate Speech in Multi-modal Memes. *arXiv preprint arXiv:2012.14891* (2020).

[13] 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*.

[14] 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 2019*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186.

[15] 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.

[16] Emilien Dupont. 2018. Learning disentangled joint continuous and discrete representations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 708–718.

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

[18] 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.

[19] 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. *arXiv preprint arXiv:2006.06195* (2020).

[20] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1470–1478.

[21] Ross Goroshin, Michael Mathieu, and Yann LeCun. 2015. Learning to linearize under uncertainty. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*. 1234–1242.

[22] 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.

[23] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net.

[24] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. 2011. Transforming auto-encoders. In *International conference on artificial neural networks*. Springer, 44–51.

[25] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net.

[26] Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. 2015. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011*

[27] Kimmo Kärkkäinen and Jungseock Joo. 2019. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913* (2019).

[28] 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* (2020).

[29] 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 preprint arXiv:1908.03557* (2019).

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[31] 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 preprint arXiv:2012.12871* (2020).

[32] Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR* abs/1711.05101 (2017). arXiv:1711.05101

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265* (2019).

[34] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning Disentangled Representations for Recommendation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 5712–5723.

[35] Jianxin Ma, Chang Zhou, Hongxia Yang, Peng Cui, Xin Wang, and Wenwu Zhu. 2020. Disentangled Self-Supervision in Sequential Recommenders. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 483–491.

[36] 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.

[37] Niklas Muennighoff. 2020. Vilio: State-of-the-art Visio-Linguistic Models applied to Hateful Memes. *arXiv preprint arXiv:2012.07788* (2020).

[38] 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.

[39] 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.

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence* 39, 6 (2016), 1137–1149.

[41] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. 2020. Learning disentangled representations via mutual information estimation. In *European Conference on Computer Vision*. Springer, 205–221.

[42] Vlad Sandulescu. 2020. Detecting Hateful Memes Using a Multimodal Deep Ensemble. *arXiv preprint arXiv:2012.13235* (2020).

[43] 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*. 1–10.

[44] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[45] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[46] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* (2019).

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

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. 5998–6008.

[49] Riza Velioglu and Jewgeni Rose. 2020. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge. *arXiv preprint arXiv:2012.12975* (2020).

[50] 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.

[51] 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.

[52] 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.

[53] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934* (2020).

[54] Weibo Zhang, Guihua Liu, Zhuohua Li, and Fuqing Zhu. 2020. Hateful Memes Detection via Complementary Visual and Linguistic Networks. *arXiv preprint arXiv:2012.04977* (2020).

[55] 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.

[56] Xiayu Zhong. 2020. Classification of Multimodal Hate Speech–The Winning Solution of Hateful Memes Challenge. *arXiv preprint arXiv:2012.01002* (2020).

[57] Yi Zhou and Zhenhao Chen. 2020. Multimodal Learning for Hateful Memes Detection. *arXiv preprint arXiv:2011.12870* (2020).

[58] Ron Zhu. 2020. Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution. *arXiv preprint arXiv:2012.08290* (2020).