Uncertainty-Aware Cross-Modal Alignment for Hate Speech Detection

Chuanpeng Yang[1,2], **Fuqing Zhu**[1,2]*, **Yaxin Liu**[1,2], **Jizhong Han**[1], **Songlin Hu**[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
{yangchuanpeng, zhufuqing, liuyaxin, hanjizhong, husonglin}@iie.ac.cn

Abstract

Hate speech detection has become an urgent task with the emergence of huge multimodal harmful content (*e.g.*, memes) on social media platforms. Previous studies mainly focus on complex feature extraction and fusion to learn discriminative information from memes. However, these methods ignore two key points: 1) the misalignment of image and text in memes caused by the modality gap, and 2) the uncertainty between modalities caused by the contribution degree of each modality to hate sentiment. To this end, this paper proposes an uncertainty-aware cross-modal alignment (UCA) framework for modeling the misalignment and uncertainty in multimodal hate speech detection. Specifically, we first utilize the cross-modal feature encoder to capture image and text feature representations in memes. Then, a cross-modal alignment module is applied to reduce semantic gaps between modalities by aligning the feature representations. Next, a cross-modal fusion module is designed to learn semantic interactions between modalities to capture cross-modal correlations, providing complementary features for memes. Finally, a cross-modal uncertainty learning module is proposed, which evaluates the divergence between unimodal feature distributions to to balance unimodal and cross-modal fusion features. Extensive experiments on five publicly available datasets show that the proposed UCA produces a competitive performance compared with the existing multimodal hate speech detection methods.

Keywords: hate speech detection, uncertainty-aware, cross-modal alignment

1. Introduction

The proliferation of social media has revolutionized the way ideas are shared and propagated, fostering the exchange of opinions across individuals, diverse cultures, and social communities at an unprecedented pace. While offering unparalleled convenience to users, social media platforms have also become conduits for the rapid dissemination of hate speech, especially in the wake of significant events like the Russian-Ukrainian conflict and COVID-19 (Pramanick et al., 2021a). Hate speech directly or indirectly attacks people based on the race, religion or other characteristics, and disseminates discriminatory statements toward social groups through platforms (Kiela et al., 2020). Such hate speech is sowing the seeds of disunity, fuelling violence and criminality in conflict areas. Therefore, detecting and curbing hate speech is a particularly urgent research issue.

Early works on hate speech detection mainly focus on analyzing text content (Waseem and Hovy, 2016; Kim et al., 2010; Malmasi and Zampieri, 2017), where the logical and semantic coherence are typically verified based on trivial indicators such as grammatical errors. Nowadays, there are various forms of hate speech (such as memes) widely present on social media platforms, and the above unimodal approaches are no longer sufficient to effectively respond. Memes, a prevalent form of user-generated content on social media platforms, have emerged as a popular means of expressing hate sentiment. Typically, a meme is an image embedded with a short piece of text that is humorous in nature. Nevertheless, what may appear as an innocuous meme can swiftly morph into a vessel for multimodal hate speech through the strategic combination of images and text, particularly in the context of contemporary political and socio-cultural divisions. The diverse and interactive nature of multimodal information renders conventional unimodal-based detection methodologies insufficient for identifying hate speech. Therefore, combining multiple modal information for reasoning is the critical factor in detecting multimodal hate speech.

Recent multimodal hate speech detection studies focus on innovative fusion technologies (Kiela et al., 2020; Lee et al., 2021) and fine-tuning large-scale pretrained multimodal models (Das et al., 2020; Lippe et al., 2020; Muennighoff, 2020; Velioglu and Rose, 2020; Zhang et al., 2020; Zhou et al., 2021). Besides, some works also attempt to utilize data augmentation (Zhu, 2020; Cao et al., 2022) and ensemble strategies (Yang et al., 2022, 2023). Despite the above studies have produced the promising progress, there are still the following limitations: 1) *The misalignment of image and text.* Most works focus on capturing critical features such as entities and demographic information, while ignoring the issue of misalignment in memes. 2) *The uncertainty between modalities.* Existing methods ex-

* Corresponding author

LREC-COLING 2024, pages 16973–16983
20-25 May, 2024.

cessively rely on multimodal fusion features, where the inherent uncertainty between modalities has not been explicitly considered, resulting in inferior performance. And the inherent uncertainty is directly reflected in the contribution degree of each modality to hate sentiment.

Figure 1: Examples illustrate two challenges encountered by current research works. Left: the misalignment between images and texts. Right: the inherent uncertainty between modalities.

The misalignment and uncertainty are widely present in multimodal hate speech. We show some representative samples in Figure 1. In the left part, the misalignment of image and text in memes is illustrated. The top meme expresses discrimination against the disabled, but only when the *leg* in the text corresponds to the *prosthetic limb* in the image can the model accurately identify the hate tendency in it. Similarly, for the meme below, only by aligning the *Asians* in the text with the *exaggerated eye-opening movements* in the image can the potential hate of nationality be identified. The above cases show that the misalignment between images and texts caused by the modality gap should be taken seriously. In the right part, the inherent uncertainty between modalities is illustrated. The text in the top meme tells an incredible story but contains an image of two smiling people. The text and image present strong cross-modal uncertainty due to completely opposite sentiment tendencies. The multimodal fusion features can provide additional discriminative information and a more comprehensive representation of memes, thereby identifying hate information against religion in memes. On the contrary, the text and image in the meme below express consistent sentiment tendency, which is able to identify the meme as non-hateful. However, the introduction of the cross-modal fusion features may cause interaction between *black* in the text and *woman* in the image, which makes the model wrongly classify it as sexist. The above cases indicate that when the cross-modal uncertainty is weak, the unimodal feature representation is sufficient to identify the hate tendency. Instead, when cross-modal uncertainty is strong, cross-modal fusion features can provide essential complementary information for memes. Therefore, the misalignment and uncertainty should be formulated in a unified manner to further discriminate hate speech in memes.

To alleviate the issues mentioned above, this paper proposes an uncertainty-aware cross-modal alignment (UCA) framework for multimodal hate speech detection. Specifically, we first utilize the cross-modal feature encoder to capture image and text feature representations of memes. Then, a cross-modal alignment module is applied to reduce semantic gaps between modalities by representing subspace alignment. Next, a cross-modal fusion module is designed to learn semantic interactions between modalities to capture cross-modal correlations. Finally, a cross-modal uncertainty learning module is proposed to estimate the uncertainty between modalities by learning from the distributional divergence of unimodal features.
The main contributions are summarized as follows:

- An uncertainty-aware cross-modal alignment (UCA) framework is proposed for modeling the misalignment and uncertainty between modalities in multimodal hate speech detection.
- The uncertainty between modalities is assessed by gauging the divergence of feature distributions, enabling adaptive control over the balance of cross-modal and unimodal features in memes.
- Extensive experiments on five publicly available datasets demonstrates that the proposed UCA yields competitive performance when compared to existing multimodal hate speech detection methods.

2. Related Work

2.1. Unimodal Hate Speech Detection

As social media platforms continue to flourish, the automated detection of hate speech has garnered substantial attention from research communities in data mining, information retrieval, and natural language processing. Researchers from diverse fields have delved into this challenging task (Fortuna et al., 2018), contributing numerous benchmark datasets (Mandl et al., 2019; Ross et al., 2017). Previous methods predominantly rely on feature engineering (Malmasi and Zampieri, 2018; Mehdad and Tetreault, 2016; Waseem and Hovy, 2016; Kim et al., 2010; Malmasi and Zampieri, 2017) to extract and organize low-level features, such as n-grams and emotional features. Currently, DNN-based methods have achieved comparable performance by aggregating potential semantic features (Zhang et al., 2018; Tekiroğlu et al., 2020). Furthermore, some studies have considered the bias and interpretability of hate models. For example, Vaidya *et al*.(Vaidya et al., 2020) enhance model interpretability and mitigate unintended bias by employing multi-task learning to predict text toxicity alongside target group labels. Mathew *et al*.(Mathew et al., 2021) utilize dataset rationales as supplementary information for fine-tuning BERT (Devlin et al., 2019) to tackle bias and enhance explainability. Despite the significant experimental progress and commercial applications of existing hate speech detection methods, they primarily focus on text-based hate speech and overlook the prevalent multimodal patterns prevalent in contemporary social media.

2.2. Multimodal Hate Speech Detection

Multimodal hate speech detection represents an emerging classification task geared towards identifying negative content, encompassing hate speech, harmful rhetoric, offensive language, and sarcasm. The surge in studies focusing on multimodal hate speech detection can be attributed to the availability of datasets containing hateful memes released in recent years. Notably, Facebook introduced the Hateful Memes Challenge (Kiela et al., 2020), prompting researchers to discern harmful categories such as nationality and religion. Previous research endeavors have explored classical dual-stream models, amalgamating visual and textual features extracted from image and text encoders via attention-based mechanisms and other fusion techniques to classify hate speech (Suryawanshi et al., 2020; Kiela et al., 2020; Das et al., 2020; Kiela et al., 2020; Lippe et al., 2020). Recent studies have also ventured into leveraging data augmentation (Zhou et al., 2021; Zhu, 2020; Lee et al., 2021; Cao et al., 2022) and ensemble strategies (Yang et al., 2022, 2023) to improve the hate speech classification performance.

With the development of hate speech detection communities, Pramanick *et al*.(Pramanick et al., 2021a) have expanded the spectrum of hateful categories by introducing two new benchmarks related to COVID-19 and US politics. Concurrently, MOMENTA has been proposed, leveraging intra-modal attention to systematically analyze the local and global perspectives of input memes (Pramanick et al., 2021b). Suryawanshi *et al*.(Suryawanshi et al., 2020) have also curated an offensive dataset comprising abusive messages targeting individuals or minority groups. Building upon this dataset, Lee *et al*.(Lee et al., 2021) propose the DisMultiHate model to disentangle visual and textual representations of memes, facilitating better understanding. Furthermore, we have found that sarcasm and hate speech have similar expressions, tending to utilize race, gender, and other factors to attract attention. For sarcasm speech detection, Cai *et al*.(Cai et al., 2019) construct a dataset from image-text tweets and propose a hierarchical fusion model. Building upon this dataset, several models have been developed to uncover implicit associations between images and texts in sarcasm (Xu et al., 2020; Pan et al., 2020). Liang *et al*.(Liang et al., 2021, 2022) deploy a heterogeneous graph structure to learn the sarcastic features from both intra- and inter-modality perspectives. However, the above works overlook the misalignments between image and text caused by the persistent modality gap, as well as the inherent uncertainty arising from the varying contributions of each modality to hate sentiment in memes. Therefore, we propose an uncertainty-aware cross-modal alignment framework to model the image-text misalignments and the uncertainty between modalities, adaptively aggregating unimodal and multimodal feature representations to discriminate hate speech in memes.

3. Methodology

3.1. Task Definition

In the task of multimodal hate speech detection, each meme comprises an image I and a text segment T, represented as a sequence of words. Both the visual and textual modalities are associated with a class label y. Our objective is to devise a classification model capable of predicting the label of a given meme (hateful or non-hateful) by effectively integrating information from both the visual and textual modalities.

3.2. Model Overview

In this section, we describe the proposed uncertainty-aware cross-modal alignment (UCA) framework for hate speech detection in detail. As illustrated in Figure 2, the architecture of UCA contains five key components: 1) *Cross-modal feature encoder*, which captures the image and text features by the modal-specific encoder; 2) *Cross-modal alignment module*, which reduces semantic gaps between modalities by representing subspace alignment; 3) *Cross-modal fusion module*, which learns semantic interactions between modalities to capture cross-modal correlations and provide complementary features for memes; 4) *Cross-modal uncertainty learning module*, which estimates the uncertainty between modalities by leveraging the distributional divergence of unimodal features; 5) *Hate speech detector*, which concatenates unimodal and cross-modal features as inputs to identify whether memes are hateful.

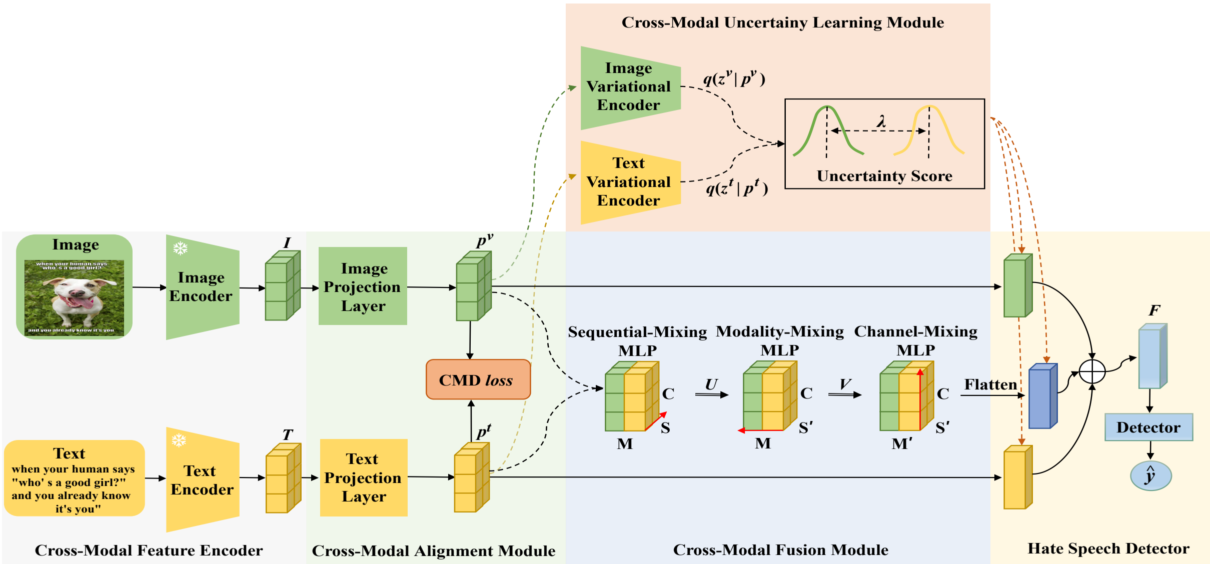

Figure 2: The illustration of the proposed UCA framework.

3.3. Cross-Modal Feature Encoder

CLIP (Radford et al., 2021) is a visual-linguistic model pretrained on a vast dataset of 400 million image-text pairs sourced from the Internet, leveraging contrastive learning. This pretraining equips CLIP with remarkable zero-shot capabilities, enabling it to effectively capture semantics for image-text inputs. Numerous studies (Gu et al., 2022; Li et al., 2022) have demonstrated CLIP's exceptional ability to generalize across various domains. Hence, our feature encoders are initialized from CLIP. The image feature is represented as $I = v_1, v_2, ..., v_i \in \mathbb{R}^{i\times 1024}$, while the text feature is represented as $T = t_1, t_2, ..., t_j \in \mathbb{R}^{j\times 768}$.

3.4. Cross-Modal Alignment Module

Multimodal hate speech exhibits metaphorical properties, necessitating a deeper semantic understanding where aligning features across different modalities serves as the cornerstone. CLIP facilitates the establishment of similarity between the feature spaces of images and their corresponding text captions. However, in the dataset used for pretraining, image and text pairs typically convey identical semantics, which may not hold true for hate speech. To enhance the learning of semantic relationships between image and text feature spaces in memes, we introduce a trainable projection layer at the output of CLIP's image and text encoders. The projection layer consists of a fully-connected feed-forward layer followed by a non-linear rectification linear unit (ReLU) as follows:

$$p_i^v = \mathrm{ReLU}(W_v v_i + b_v), \tag{1}$$

$$p_j^t = \mathrm{ReLU}(W_t t_j + b_t), \tag{2}$$

where $p_i^v \in \mathbb{R}^{256}$ and $p_j^t \in \mathbb{R}^{256}$ are the vectors projected from the image representation I and the text representation T, respectively. W_v, b_v, W_t, and b_t are learnable parameters of the projection layers.

Domain adaptation (Long et al., 2015) has demonstrated remarkable proficiency in aligning feature distributions, primarily through two approaches. Firstly, there are statistic moment matching-based methods such as Maximum Mean Discrepancy (MMD) and Kullback-Leibler (KL) divergence (Long et al., 2017, 2018; Zhu et al., 2019). Secondly, there are adversarial learning-based methods, including domain adversarial adaptation (Ganin et al., 2016; Hoffman et al., 2018). In our framework, we leverage Central Moment Difference (CMD) (Zellinger et al., 2017) to align the feature distributions. CMD evaluates the discrepancy between the distributions of two representations by comparing the differences in their corresponding order-wise moments. As the CMD distance decreases, the two distributions become more similar. Compared to MMD or KL-divergence methods, CMD explicitly matches higher-order moments without requiring costly distance and kernel matrix computations. Additionally, compared to adversarial training methods, the CMD formulation is more straightforward, as it does not involve a discriminator with additional parameters.

Consider bounded random samples X and Y with probability distributions p and q respectively, defined on the interval $[a, b]^N$. The Central Moment Discrepancy regularizer, denoted as CMD$_K$, is defined as an empirical estimate of the CMD metric, expressed as:

$$\begin{aligned}\mathrm{CMD}_K(X,Y) &= \frac{1}{|b-a|}\|\mathbf{E}(X)-\mathbf{E}(Y)\|_2 \\ &+ \sum_{k=2}^{K}\frac{1}{|b-a|^k}\left\|C_k(X)-C_k(Y)\right\|_2,\end{aligned} \tag{3}$$

where $\mathbf{E}(X) = \frac{1}{|X|}\sum_{x\in X} x$ is the empirical expectation vector computed on the sample X and $C_k(X) = \mathbf{E}((x - \mathbf{E}(X))^k)$ is the vector of all the k^{th} order sample central moments of the coordinates of X. The CMD loss between each image and text is calculated as follows:

$$\mathcal{L}_{\text{align}} = \text{CMD}_K(p^v, p^t). \tag{4}$$

3.5. Cross-Modal Fusion Module

Cross-modal fusion plays a crucial role in capturing semantic interactions between different modalities, offering complementary features essential for hate speech detection. This becomes especially significant when image and text feature representations exhibit conflicting sentiment tendencies within the same memes. Therefore, we have devised the cross-modal fusion module to discern and learn the correlations between modalities. Newly proposed architectures for vision tasks leverage Multilayer Perceptron (MLP)-based models. These models, such as MLP-mixer (Tolstikhin et al., 2021) and ResMLP (Touvron et al., 2022), substitute MLPs for the traditional self-attention mechanism, resulting in significant reductions in computational costs while maintaining high performance. Typically, these models feature two independent MLPs—one processing the sequential length and the other handling the channel size. More recently, CubeMLP (Sun et al., 2022) has been introduced to effectively process multimodal features. Drawing inspiration from CubeMLP, we adopt three MLPs to comprehensively mix features along the sequential, modality, and channel axes.

Specifically, we concatenate the unimodal features to form a multimodal tensor $D \in \mathbb{R}^{S\times M\times C}$, where S represents the sequential length, M denotes the number of modalities, and C signifies the size of feature channels. Subsequently, the multimodal features are fed into stacked three MLP units for mixing. Each MLP unit comprises two fully-connected layers followed by a nonlinear activation GELU (Hendrycks and Gimpel, 2016), designed to mix the multimodal features along its respective axis. A residual connection is employed in the unit according to (Touvron et al., 2022). Taking the first sequential-mixing MLP as an example, the tensor $D \in \mathbb{R}^{S\times M\times C}$ can be conceptualized as a collection of vectors $D_{*,m,c} \in \mathbb{R}^{S\times 1\times 1}$, where $(m, c) \in \{(1, 1), (1, 2), ..., (2, 1), (2, 2), ..., (M, C)\}$. Here, $D_{*,m,c}$ represents the vector corresponding to the m^{th} modality and c^{th} channel. Each fully-connected layer within the sequential-mixing MLP unit can be expressed as:

$$\text{FC}_S(D_{*,m,c}) = W_S D_{*,m,c} + b_S, \tag{5}$$

where $W_S \in \mathbb{R}^{S\times S'}$ and $b_S \in \mathbb{R}^{S'}$ represent two matrix-represented learnable parameters. S' denotes the reduced dimensionality along the S-axis, which serves as a hyperparameter. All $D_{*,m,c}$ instances share the parameters W_S and B_S. Consequently, the entire sequential-mixing MLP can be delineated as:

$$\begin{aligned} U_{*,m,c} = {} & \text{LayerNorm}(\text{FC}_S(\text{GELU}(\text{FC}_S(D_{*,m,c}))) \\ & + D_{*,m,c}), \end{aligned} \tag{6}$$

where the output tensor $U \in \mathbb{R}^{S'\times M\times C}$ can be viewed as a collection of vectors $U_{*,m,c} \in \mathbb{R}^{S'\times 1\times 1}$.

Similar to the first MLP unit operating along the S-axis, the output $V \in \mathbb{R}^{S'\times M'\times C}$ of the second MLP unit along the M-axis can be interpreted as a collection of vectors $V_{s,*,c} \in \mathbb{R}^{1\times M'\times 1}$. Likewise, the output $G \in \mathbb{R}^{S'\times M'\times C'}$ of the third MLP unit along the C-axis can be seen as a set of vectors $G_{s,m,*} \in \mathbb{R}^{1\times 1\times C'}$. Here, M' and C' represent reduced dimensions along the M-axis and C-axis, respectively. Notably, $(s, c) \in \{(1, 1), (1, 2), ..., (2, 1), (2, 2), ..., (S', C)\}$ and $(s, m) \in \{(1, 1), (1, 2), ..., (2, 1), (2, 2), ..., (S', M')\}$. Finally, the modality-mixing MLP and the channel-mixing MLP can be represented as:

$$\begin{aligned} V_{s,*,c} = {} & \text{LayerNorm}(\text{FC}_M(\text{GELU}(\text{FC}_M(U_{*,m,c}))) \\ & + U_{*,m,c}), \end{aligned} \tag{7}$$

$$\begin{aligned} G_{s,m,*} = {} & \text{LayerNorm}(\text{FC}_C(\text{GELU}(\text{FC}_C(V_{s,*,c}))) \\ & + V_{s,*,c}), \end{aligned} \tag{8}$$

where $G \in \mathbb{R}^{S'\times M'\times C'}$ is the mixed cross-modal feature representation.

3.6. Cross-Modal Uncertainty Learning Module

The multimodal hate speech detection task aims to obtain a comprehensive feature set from the input data. One distinctive aspect of this task is the intrinsic uncertainty between modalities, stemming from the varying contribution degrees of each modality to hate sentiment. This uncertainty impacts the efficacy of cross-modal fusion representations. To address this challenge, we introduce the cross-modal uncertainty learning module.

We assess the Kullback-Leibler (KL) divergence between unimodal distributions approximated by two modality-specific variational encoders (Chen et al., 2022). The derived uncertainty score is then utilized to dynamically regulate the contribution of cross-modal and unimodal features in hate speech detection. Initially, we conceptualize the unimodal features (p^v and p^t) from a generative standpoint, where the features are extracted by sampling from a latent space with isotropic Gaussian priors. A fundamental assumption underlying our approach is that the disparity in the distributions of unimodal features reflects the information gap between different modalities. Consequently, the uncertainty can be estimated by divergences computed across the feature spaces. Formally, the corresponding variational posterior for a unimodal observation is denoted as $q(z|p) = \mathcal{N}[z|\mu(p), \sigma(p)]$, where the mean

μ and variance σ are obtained from the modality-specific variational encoder. Furthermore, for each data sample n comprising aligned image feature p_n^v and textual feature p_n^t, the variational posteriors for both modalities are expressed as follows:

$$q(z_n^v \mid p_n^v) = \mathcal{N}[z_n^v \mid \mu(p_n^v), \sigma(p_n^v)], \tag{9}$$
$$q(z_n^t \mid p_n^t) = \mathcal{N}[z_n^t \mid \mu(p_n^t), \sigma(p_n^t)]. \tag{10}$$

Then, we obtain the variational posterior distributions for both modalities by averaging the variational posteriors for each data sample. This allows us to capture the overall distribution of both modalities across the entire dataset for the purpose of modeling the uncertainty. For the visual modality, we have:

$$q(z^v) = \mathbb{E}_{p^v}[q(z^v \mid p^v)] = \frac{1}{N}\sum_{n=1}^{N} q(z_n^v \mid p_n^v), \tag{11}$$

where z^v is the latent variable for the visual modality, and N is the total number of data samples. Similarly, for the textual modality, we have:

$$q(z^t) = \mathbb{E}_{p^t}[q(z^t \mid p^t)] = \frac{1}{N}\sum_{n=1}^{N} q(z_n^t \mid p_n^t), \tag{12}$$

where z^t is the latent variable for the textual modality. The uncertainty of different modalities in data sample n can be quantified by the average KL divergence between unimodal distributions, given by:

$$\lambda_n^1 = \left(\frac{\mathcal{D}_{KL}\left[q(z_n^v \,||\, p_n^v) \,||\, q(z_n^t \,||\, p_n^t)\right]}{\mathcal{D}_{KL}\left[q(z^v) \,||\, q(z^t)\right]}\right), \tag{13}$$
$$\lambda_n^2 = \left(\frac{\mathcal{D}_{KL}\left[q(z_n^t \,||\, p_n^t) \,||\, q(z_n^v \,||\, p_n^v)\right]}{\mathcal{D}_{KL}\left[q(z^t) \,||\, q(z^v)\right]}\right), \tag{14}$$
$$\lambda_n = \mathrm{Sigmoid}\left(\frac{\lambda_n^1 + \lambda_n^2}{2}\right), \tag{15}$$

where the uncertainty score λ_n is computed as the symmetrized KL divergence obtained by averaging the normalized values of $\mathcal{D}_{KL}[q(z_n^v \,||\, p_n^v) \,||\, q(z_n^t \,||\, p_n^t)]$ and $\mathcal{D}_{KL}[q(z_n^t \,||\, p_n^t) \,||\, q(z_n^v \,||\, p_n^v)]$. The sigmoid function is used as the activation function to map the uncertainty scores to the range $[0, 1]$. The uncertainty score λ_n serves as the weight controlling the fusion of unimodal and cross-modal features during both training and inference. Specifically, in the process of cross-modal uncertainty learning, cross-modal features are adaptively utilized while unimodal features are dropped out when uncertainty is high, and vice versa.

3.7. Hate Speech Detector

We flatten the mixed multimodal features and adaptively concatenate two unimodal feature embeddings. Specifically, we utilize the uncertainty score λ_n to guide the fusion of features. The cross-modal feature is multiplied by λ_n and each unimodal feature is multiplied by $1 - \lambda_n$.

$$F_n = \lambda_n G \oplus (1-\lambda_n)p^v \oplus (1-\lambda_n)p^t, \tag{16}$$

where $\oplus$ denotes the concatenation operation. Subsequently, the fused feature F_n is passed to the hate speech detector for classification. The detector comprises a two-layer fully connected feed-forward network with intermediate ReLU non-linearity, along with a softmax layer utilized to estimate the probability of hatefulness.

$$H_n = \mathrm{ReLU}(W_1 F_n + b_1), \tag{17}$$
$$\hat{y}_n = \mathrm{Softmax}(W_2 H_n + b_2), \tag{18}$$

where W_1, W_2, b_1, and b_2 are learnable parameters. $\hat{y}_n$ is the estimated probability. The cross-entropy loss $\mathcal{L}_{\text{task}}$ is employed for hate speech detection task:

$$\mathcal{L}_{\text{task}} = -\frac{1}{N}\sum_{n=1}^{N} y_n \log(\hat{y}_n), \tag{19}$$

where y_n is the ground-truth one-hot label. We combine classification loss and modal alignment loss to obtain the optimization objective of UCA framework.

$$\mathcal{L}_{\text{Loss}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{align}}. \tag{20}$$

4. EXPERIMENTS

4.1. Datasets

The experiment is conducted on five publicly available datasets, described briefly as follows:

Hate dataset: This dataset is part of the Hateful Memes Challenge 2020 for multimodal hate speech detection, published in (Kiela et al., 2020). It comprises $10K$ memes with binary labels indicating whether they are hateful or non-hateful.

Harm-C dataset: This dataset, related to COVID-19, is published in (Pramanick et al., 2021a) for multimodal harmful detection. It contains nearly $3.5K$ memes with binary labels indicating whether they are harmful or non-harmful.

Harm-P dataset: This dataset, related to United States politics, is published in (Pramanick et al., 2021b) for multimodal harmful detection. It consists of nearly $3.5K$ memes with binary labels indicating whether they are harmful or non-harmful.

Offense dataset: Related to the 2016 United States presidential election, this dataset is published in (Suryawanshi et al., 2020) for multimodal offensive detection. It comprises nearly $1K$ memes with binary labels indicating whether they are offensive or non-offensive.

Sarcasm dataset: This dataset consists of image-text tweets collected in (Cai et al., 2019) for multimodal sarcasm detection. It contains nearly $25K$ memes with binary labels indicating whether they are sarcastic or non-sarcastic.

Models	Acc. ↑	AUROC ↑
Late Fusion	63.20	69.30
Concat BERT	61.53	67.77
MMBT-Region	67.66	73.82
ViLBERT	65.27	73.32
Visual BERT	66.67	74.42
DisMultiHate	71.26	79.89
CDKT	**76.50**	83.74
PromptHate	72.98	81.45
CLIP	59.00	68.30
UCA (Ours)	76.10	**84.32**

Table 1: Performance comparison on the Hate.

Models	Harm-C		
	Acc. ↑	F1 ↑	MMAE ↓
ViLBERT	78.53	78.06	0.1881
Visual BERT	81.36	80.13	0.1857
MOMENTA	83.82	82.80	0.1743
TOT	87.01	85.93	0.1634
CLIP	73.45	72.61	0.2508
UCA (Ours)	**88.98**	**88.31**	**0.1015**

Table 2: Performance comparison on the Harm-C.

4.2. Implementation Details

In the cross-modal feature encoder, we employ the CLIP-Large (Radford et al., 2021) model to initialize the image and text encoders. The output vector dimension are 1024 for the image encoder and 768 for the text encoder. In the cross-modal alignment module, the output dimension of the cross-modal projection layer is set to 256. For the cross-modal fusion module, we set S to 100, which entails zero-padding shorter sequences and truncating longer sequences to match the sequence size. M is fixed to 2 since we only have two involved modalities, while C is set to 256, consistent with the output dimension of the projection layer. Additionally, S', M', and C' are set to 10, 2 and 32, respectively. In the hate speech detector, the intermediate feature dimension of the detector is 64, and the dropout rate is 0.4. We utilize weighted Adam as the optimizer, employing a cosine annealing and warm-up strategy to regulate the variation of the learning rate. The initial learning rate is set to 0.001. The size of the minibatch is fixed at 64. The training epochs for each dataset is 20.

4.3. Evaluation Metrics

For the Hate dataset, we follow the evaluation method adopted by (Kiela et al., 2020), utilizing Area Under the Receiver Operating Characteristic curve (AUROC) and accuracy (Acc.) as evaluation metrics. The AUROC is the primary metric. For the Harm-C and Harm-P datasets, we adopt the evaluation method adopted by (Pramanick et al., 2021b), utilizing Acc., Macro-F1 (F1), and Macro-Averaged Mean Absolute Error (MMAE) as evaluation metrics. For the Offense dataset, we follow the evaluation strategy presented in (Suryawanshi et al., 2020), employing F1, precision (Pre.) and recall (Rec.) as evaluation metrics. For the Sarcasm dataset, we employ the evaluation method described in (Cai et al., 2019), using F1, Pre., Rec. and Acc. as evaluation metrics.

Models	Harm-P		
	Acc. ↑	F1 ↑	MMAE ↓
ViLBERT	87.25	86.03	0.1276
Visual BERT	86.80	86.07	0.1318
MOMENTA	89.84	88.26	0.1314
TOT	91.55	91.29	0.1245
CLIP	83.02	82.83	0.1604
UCA (Ours)	**92.68**	**92.66**	**0.0739**

Table 3: Performance comparison on the Harm-P.

Models	F1 ↑	Pre. ↑	Rec. ↑
StackedLSTM+VGG16	46.30	37.30	61.10
BiLSTM+VGG16	48.00	48.60	58.40
CNNText+VGG16	46.30	37.30	61.10
ERNIE-VIL	53.10	54.30	63.70
DisMultiHate	64.60	64.50	65.10
CLIP	58.94	60.98	59.07
UCA (Ours)	**65.89**	**66.09**	**66.90**

Table 4: Performance comparison on the Offense.

4.4. Experimental Results

As shown on Table 1-5, UCA significantly outperforms all the compared methods in all metrics for each dataset, which demonstrates the effectiveness of the proposed UCA framework. Specifically, UCA obtains a new state-of-the-art result with an AUROC of 84.32% on the Hate dataset, producing a significant improvement of approximately $+3\%$. For the Harm dataset, UCA could model the inherent uncertainty between modalities compared to TOT (Zhang et al., 2023), providing a more robust result. For the Offense dataset, UCA could produce a higher performance than DisMultiHate (Lee et al., 2021) without extracting additional features such as entities and demographic information. UCA also produces an ACC. of 87.8%, creating a new state-of-the-art result on the Sarcasm dataset. The

Models	F1 ↑	Pre. ↑	Rec. ↑	Acc. ↑
HFM	80.90	79.40	82.45	83.44
D&R Net	80.60	77.97	83.42	84.02
Res-Bert	81.57	78.87	84.46	84.80
MIII-MMSD	82.92	80.87	85.08	86.05
InCrossMGs	85.60	85.39	85.80	86.10
CDKT	83.89	79.37	**88.96**	85.60
CMGCN	87.00	87.02	86.97	87.55
CLIP	81.94	82.21	83.61	82.40
UCA (Ours)	**87.36**	**87.13**	87.64	**87.80**

Table 5: Performance comparison on the Sarcasm.

Models	Hate		Harm-C			Harm-P			Offense			Sarcasm			
	Acc. ↑	AUROC ↑	Acc. ↑	F1 ↑	MMAE ↓	Acc. ↑	F1 ↑	MMAE ↓	F1 ↑	Pre. ↑	Rec. ↑	F1 ↑	Pre. ↑	Rec. ↑	Acc. ↑
UCA (Ours)	**76.10**	**84.32**	**88.98**	**88.31**	**0.1015**	**92.68**	**92.66**	**0.0739**	**65.89**	**66.09**	**66.90**	**87.36**	**87.13**	**87.64**	**87.80**
UCA w/o P	75.20	83.41	88.21	87.45	0.1142	91.98	91.74	0.0814	63.48	63.45	64.00	87.27	86.93	87.61	87.63
UCA w/o A	74.60	81.82	86.87	86.36	0.1295	90.44	90.05	0.1007	62.71	62.66	63.14	86.59	86.36	86.89	87.05
UCA w/o F	74.80	82.13	88.06	87.24	0.1176	90.63	90.55	0.0921	63.13	63.03	63.38	86.62	86.41	86.89	87.09
UCA w/o U	72.10	80.26	86.47	85.63	0.1337	90.11	89.65	0.1138	61.74	61.66	61.96	86.42	86.18	86.74	86.88

Table 6: Ablation study evaluated on the Hate, Harm-C, Harm-P, Offense and Sarcasm datasets.

above stable improvement demonstrates the effectiveness of learning image-text alignment and inter-modal uncertainty. We also use CLIP to fine-tune each dataset and directly concatenate the output features into the classifier as a baseline. Compared to CLIP, UCA could produce a significant improvement, especially on the Hate dataset, with an improvement of over $+15\%$. Besides, UCA requires a lower computational complexity than CLIP.

4.5. Ablation Study

To evaluate the effectiveness of each component in UCA, we conduct a series of ablation studies on each dataset as shown on Table 6.

w/o P: After removing the projection layer of image and text, the performance decreases slightly, indicating that the projection before alignment could improve the semantic relationship between the image and text feature spaces of memes.

w/o A: After removing the cross-modal alignment loss, the performance decreases greatly, illustrating that reducing the gap between modalities to align image and text is particularly significant for identifying hateful memes.

w/o F: After removing the cross-modal fusion module and using the attention mechanism to capture dependencies between modalities, performance decreases to some extent, verifying that MLP-based cross-modal fusion could maintain high performance while reducing computational costs.

w/o U: After removing the cross-modal uncertainty learning module, performance decreases the most, demonstrating that considering the contribution of each modality to hate sentiment is the most critical factor for multimodal hate speech detection.

4.6. Case Study

The purpose of UCA is to model the misalignment and uncertainty between modalities for multimodal hate speech detection. To further understand UCA intuitively, we show some cases in Figure 3. Specifically, in the first meme, the image and text represent completely opposite sentiment tendencies, with the image expressing hate sentiment. Compared to CLIP, UCA focuses more on the alignment of background information and is more in line with the *watching* state. Uncertainty learning can bridge the information gap between modalities and provide a more complementary feature representation for memes. On the contrary, in the second meme, both the image and the text express the same sentiment tendencies. However, establishing the correlation between the *fun* in the text and the *crazy movements* in the image can lead to sentiment leaning towards hate. UCA could identify the presence of less uncertainty between modalities, thereby adaptively aggregating more unimodal features. The above cases demonstrate that UCA could promote alignment between modalities and determine when unimodal information is sufficient and when cross-modal fusion information is crucial.

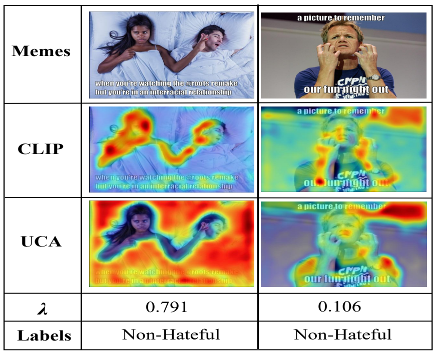

Figure 3: Case study of memes on the Hate.

5. Conclusion

In this paper, an uncertainty-aware cross-modal alignment framework is proposed by modeling the misalignment and uncertainty between modalities for multimodal hate speech detection. UCA consists of two crucial components: cross-modal alignment and uncertainty learning modules. The cross-modal alignment module enhances the semantic relationship between the image and text feature spaces of memes. On the other hand, the cross-modal uncertainty learning module plays a crucial role in determining the adequacy of unimodal information versus the necessity of cross-modal fusion information, offering a complementary perspective for memes. Experimental results on five publicly available datasets demonstrate that UCA produces a competitive performance compared with previous methods. The ablation and case studies provide additional insights into the effectiveness of each component in UCA.

6. References

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2506–2515.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332. Association for Computational Linguistics.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, pages 2897–2905.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Paula Fortuna, José Ferreira, Luiz Pires, Guilherme Routar, and Sérgio Nunes. 2018. Merging datasets for aggressive text identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 128–139.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, pages 2096–2030.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2022. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 2611–2624.

Myung Jong Kim, Younggwan Kim, JaeDeok Lim, and Hoirin Kim. 2010. Automatic detection of malicious sound using segmental two-dimensional mel-frequency cepstral coefficients and histograms of oriented gradients. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 887–890.

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*, pages 5138–5147.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. 2022. Language-driven semantic segmentation. In *International Conference on Learning Representations*.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4707–4715.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777.

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*.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1647–1657.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

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*, pages 299–303.

Niklas Muennighoff. 2020. Vilio: state-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 3722–3729.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Peter Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. 2022. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference* on Web and Social Media, volume 14, pages 683–693.

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*.

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*, pages 88–93.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3777–3786.

Chuanpeng Yang, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2023. Invariant meets specific: A scalable harmful memes detection framework. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4788–4797.

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4505–4514.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. In *International Conference on Learning Representations*.

Linhao Zhang, Li Jin, Xian Sun, Guangluan Xu, Zequn Zhang, Xiaoyu Li, Nayu Liu, Shiyao Yan, and Qing Liu. 2023. Tot: Topology-aware optimal transport for multimodal hate detection. *arXiv preprint arXiv:2303.09314*.

Weibo Zhang, Guihua Liu, Zhuohua Li, and Fuqing Zhu. 2020. Hateful memes detection via complementary visual and linguistic networks. *arXiv preprint arXiv:2012.04977*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 745–760.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290*.

Yongchun Zhu, Fuzhen Zhuang, and Deqing Wang. 2019. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5989–5996.