TOT: Topology-Aware Optimal Transport for Multimodal Hate Detection

Linhao Zhang[1,2,3], **Li Jin**[1,2*], **Xian Sun**[1,2], **Guangluan Xu**[1,2], **Zequn Zhang**[1,2], **Xiaoyu Li**[1,2], **Nayu Liu**[1,2,3], **Qing Liu**[1,2], **Shiyao Yan**[1,2,3]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences
[2]Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute
[3]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences
zhanglinhao20@mails.ucas.ac.cn, jinlimails@gmail.com

Abstract

Multimodal hate detection, which aims to identify the harmful content online such as memes, is crucial for building a wholesome internet environment. Previous work has made enlightening exploration in detecting explicit hate remarks. However, most of their approaches neglect the analysis of implicit harm, which is particularly challenging as explicit text markers and demographic visual cues are often twisted or missing. The leveraged cross-modal attention mechanisms also suffer from the distributional modality gap and lack logical interpretability. To address these semantic gap issues, we propose TOT: a topology-aware optimal transport framework to decipher the implicit harm in memes scenario, which formulates the cross-modal aligning problem as solutions for optimal transportation plans. Specifically, we leverage an optimal transport kernel method to capture complementary information from multiple modalities. The kernel embedding provides a non-linear transformation ability to reproduce a kernel Hilbert space (RKHS), which reflects significance for eliminating the distributional modality gap. Moreover, we perceive the topology information based on aligned representations to conduct bipartite graph path reasoning. The newly achieved state-of-the-art performance on two publicly available benchmark datasets, together with further visual analysis, demonstrate the superiority of TOT in capturing implicit cross-modal alignment.

Introduction

The flourishing expansion of social media has facilitated the exchange of opinions among individuals, different cultural or social communities. However, under the influence of major events such as the Russian-Ukrainian conflict and COVID-19, these platforms are flooded with hateful content, disseminating discriminatory statements toward social groups based on their races, religions or other characteristics. Such hateful content is sowing seeds of disunity to exacerbate violent and criminal behaviors in conflict areas. Therefore, the automatic identification of hateful content on the Internet is a social research issue with great urgency.

*Corresponding Author.
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

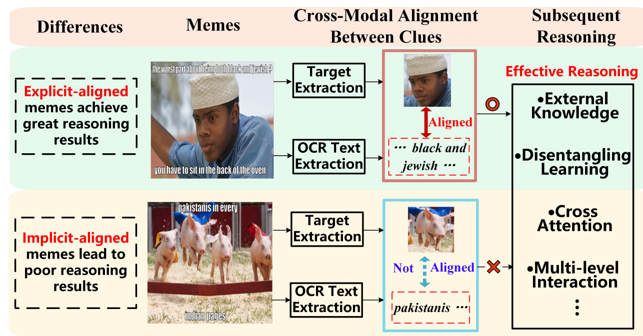

Figure 1: Differences between explicit-aligned and implicit-aligned memes. The aligning procedure is naturally completed in explicit harmful content, but is indispensable for implicit cases.

Memes, as a popular kind of social media user-generated content (UGC), have become a prevalent method for propagating harmful content due to their viral nature. Typically, a meme is an image embedded with a short piece of text. Although memes have a puckish sense of humor, some implicit hateful content is concealed. The critical factor in detecting harmful memes is combining well-aligned visual-linguistic clues and conducting multimodal reasoning.

Previous works have made remarkable progress in detecting explicit-aligned hate. (Lee et al. 2021; Pramanick et al. 2021). Their models achieved various kinds of practical reasoning through external knowledge enhancement (Zhu 2020; Zhou, Chen, and Yang 2021), sophisticated interaction mechanisms (Lippe et al. 2020) or an ensemble of multiple visual-linguistic models (Das, Wahi, and Li 2020; Velioglu, Rose, and Analytics 2020). However, these approaches neglect the nuance of the multimodal aligning process, which is necessary for subsequent reasoning procedures. Take the second meme in Figure 1 as an example, which dehumanizes Pakistanis as pigs. Although the word 'Pakistanis' has no semantic correspondence with the 'pigs' in the picture, a human can easily capture the implicit cross-modal alignment between 'Pakistanis' and 'pigs' to perceive the potent offense. Nevertheless, this implicit harm is challenging for efficient detection. Previous works neglected the aligning procedure and conducted the reasoning process directly

(Pan et al. 2020), resulting in suboptimal performance on implicit-aligned memes.

Recently, optimal transport (OT), as one of the research hotspots from optimization theory, has attracted extensive attention in computer vision, natural language processing and other fields due to its excellent performance on sequence alignment and domain adaption problems (Duan et al. 2022). OT has the ability to reduce distributional bias in a more interpretable way under low-resource scenarios, which solves the problem of lacking external labels for aligning multimodal clues in implicit harmful memes (Chen et al. 2020; Maretic et al. 2022).

In this paper, inspired by the OT theory, we propose a general framework TOT: topology-aware optimal transport to capture implicit cross-modal alignment in memes. Specifically, we leverage an optimal transport kernel method to reformulate the alignment problem across different modalities. The leveraged Gaussian kernel provides the transformation ability to reproduce kernel Hilbert space (RKHS), in which the distributional modality gap gets eliminated to calculate an informative cost matrix for generating transportation plans. Secondly, based on the sinkhorn algorithm, we acquire optimal transportation plans, which are used for assigning source values to target distribution at minimum total cost. In our implementation, the cost matrix is a divergence of alignment represented by the pair-wise dot product of image and text sequence features in RHKS. Thus the optimal transportation plan with minimum total cost represents the assigning weights with maximum total alignment. Then we initialize topology structures to capture inter-modal correspondence between aligned feature nodes. The initialized structure is dynamically updated through topology reasoning for several steps to allow comprehensive and representative information propagation. Finally, we leverage a residual connection between learned input representations and reasoning scores based on cross-modal multi-head attention to realize a complementation. Our contributions are summarized as follows:

- We notice the challenging problem of detecting implicit harmful memes and leverage optimal transport kernel method to model the cross-modal aligning task. To the best of our knowledge, this is the first work to incorporate such kernel methods with transportation theory in multimodal hate content detection.
- We implement topology-aware optimal transport to capture inter-modal correspondence, which establishes credible alignment based on transported feature nodes, and perform dynamic topology reasoning to allow comprehensive and representative information propagation.
- Without any external knowledge, our model achieves state-of-the-art performance by a large margin on two publicly available benchmarks. The subsequent visualization studies further confirm our model's superiority in capturing implicit cross-modal alignment.

Related Work

Hate Content Detection

With the prevalence of social media platforms, automated identification of hate content has received academic attention. Researchers from diverse communities have explored this challenging work (Fortuna et al. 2018), and produced a large number of benchmark datasets (Bretschneider and Peters 2017; Ross et al. 2017; Mandl et al. 2021). Previous feature-engineering methods (Malmasi and Zampieri 2018; Mehdad and Tetreault 2016) mainly extracted and organized lower-level features, like n-gram and sentiment features. Nowadays, DNN-based methods have garnered better results by aggregating latent semantic features (Zhang, Robinson, and Tepper 2018) or fine-tuning large pre-trained models (Tekiroglu, Chung, and Guerini 2020) like GPT-2. Although the existing methods of hateful content detection have yielded considerable experimental progress and commercial application, they just focused on text-based hateful content, neglecting the abundant multimedia UGC.

Multimodal Hate Content Detection

Frequent and repetitive exposure to multimodal harmful content will increase personal prejudice against external groups and further affect the status of socially disadvantaged groups. Enabling explorations were made for flourishing hateful meme detection studies by publishing benchmarks. For instance, Facebook had proposed the Hateful Memes Challenge (Kiela et al. 2020), which encouraged researchers to identify the targeted harmful categories (e.g., race and sex). (Zia, Castro, and Tyson 2021) proposed to classify memes beyond hateful and confirm the type of attack (e.g., contempt and slur). Recent MOMENTA (Pramanick et al. 2021) further extended the limited harmful categories and proposed two benchmarks related to COVID-19 and US politics. Approaches of multimodal hateful memes classification (Kiela et al. 2020; Suryawanshi et al. 2020) adopted early fusion techniques to incorporate linguistic and visual signals at input space and yield sub-optimal results. Advanced systems fine-tuned large scale pre-trained unimodal and multimodal models such as VisualBERT (Li et al. 2019), UNITER (Chen et al. 2019b), VILLA (Gan et al. 2020), and ensemble thereof. More recently, DisMultiHate (Lee et al. 2021) disentangled target entities to certain categories of hate (i.e., gender, race, etc.) in multimodal memes to improve classification and explainability. MOMENTA (Pramanick et al. 2021) systematically analyzed the local and the global perspective of the input meme and performed cross-modal attention in general level to detect harmful memes. However, aforementioned rivaling approaches relies on external demographic knowledge and have limited effectiveness in detecting harmful content beyond categories of gender or race, especially memes with implicit modal alignment. This paper aims to address this gap by proposing a general framework with optimal transported graph.

Optimal Transport

Optimal transport is a well-studied topic in optimization theory, the original aim of which is to find the best trans-

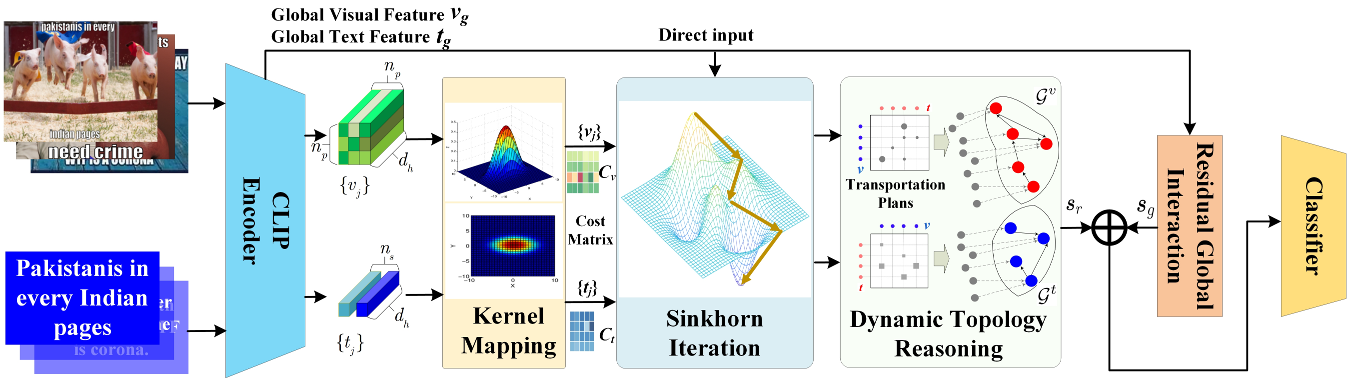

Figure 2: The model architecture of our TOT. Firstly, the representations of images and text can be learned through CLIP. Secondly, Kernel Mapping stage formulates the transportation cost matrix to optimize. Thirdly, we perform topology aligning and output two graphs with aligned representation. Finally, TOT conducts reasoning to calculate classification scores.

portation plan between two data distributions with minimum cost. Recently, OT has indeed been widely used in alignment problems by researchers from various communities including computer vision and natural language processing (Grave, Joulin, and Berthet 2019). VOLT (Xu et al. 2021) formulated the quest of vocabularization as an optimal transport (OT) problem by finding the optimal transport matrix from the character distribution to the vocabulary token distribution. (Chen et al. 2019a) used the transport plan as an ad-hoc attention score in the context of network embedding to align data modalities. MuLOT (Pramanick, Roy, and Patel 2022) utilized optimal transport based domain adaptation to learn strong cross-modal dependencies for sarcasm and humor detection. (Ge et al. 2021) innovatively revisited the label assignment from a global perspective and proposed to formulate the assigning procedure as an optimal transport (OT) problem. However, none of these studies have exploited optimal transport to learn element-level alignment representation as graph nodes. This paper contributes to memes analysis by proposing a novel framework TOT, which performs multimodal aligning for implicit harmful memes and constructs topology alignment graphs to allow dynamic topology reasoning.

Methods

Input Representation Learning

Unlike previous works that leveraged isolated encoders for visual and linguistic signals (Yu et al. 2021), we choose CLIP (Radford et al. 2021) to learn the unified representations for input memes samples. CLIP is a visual-linguistic model pre-trained on 400M image–text pairs from the Internet through contrastive learning, which offers excellent zero-shot capabilities in capturing semantics for image-text inputs. Towards proposed TOT framework, we feed a meme's image and its OCR-extracted text to CLIP model initialized with pre-trained weights, then four embeddings $\boldsymbol{v}_g, \boldsymbol{t}_g, \boldsymbol{V}, \boldsymbol{T}$ from the multimodal encoder are extracted, where $\boldsymbol{v}_g, \boldsymbol{t}_g \in \mathbb{R}^{d_h}$ are global representations for multimodal input, and $\boldsymbol{V} \in \mathbb{R}^{n_p^2 \times d_h}$, $\boldsymbol{T} \in \mathbb{R}^{n_s \times d_h}$ are sequence vectors containing local image and text information respectively. Note that $n_p \times n_p$ is the patch numbers per image and n_s is the word numbers per sentence.

Transportation Problem Formulating

This section exploits the Gaussian kernel method (Mialon et al. 2021) to formulate the cross-modal alignment problem as solutions for optimal transportation plans.

Based on the CLIP local visual feature sequence $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{n_p^2}\}$ and local linguistic feature sequence $\boldsymbol{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_{n_s}\}$, we can acquire the cost matrix $\mathbf{C}$ towards transportation problem through a kernel method. Let κ be the Gaussian kernel with reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ and its associated kernel embedding $\varphi : \mathbb{R}^d \to \mathcal{H}$. Then the $n_s \times n_p^2$ cost matrix $\mathbf{C}$ can be obtained by calculating the comparisons $\kappa(\boldsymbol{v}_i, \boldsymbol{t}_j)$ before detailed alignment.

With the cost matrix $\mathbf{C}$, the transportation plan from image to text, denoted by the $n_s \times n_p^2$ matrix $\mathbf{P}(\boldsymbol{V}, \boldsymbol{T})$ is then defined as the unique solution of:

$$\underset{\mathbf{P} \in U}{min} \sum_{ij} \mathbf{C}_{ij} \mathbf{P}_{ij} - \varepsilon H(\mathbf{P}) \tag{1}$$

$$H(\mathbf{P}) = -\sum_{ij} \mathbf{P}_{ij} (log(\mathbf{P}_{ij}) - 1) \tag{2}$$

where $\mathbf{C}$ represents the pairwise costs to align the elements of $\boldsymbol{V}$ and $\boldsymbol{T}$, $H(\mathbf{P})$ refers to the entropy regularization with parameter ϵ to control the sparsity of $\mathbf{P}$, U is the space of admissible couplings.

Topology-Aware Optimal Transport

This section describes the key module of proposed TOT framework, which is designed to align multimodal representations based on transportation plans, and then establish topology structures to capture inter-modal correspondence as candidate edges among aligned feature nodes.

Based on transportation plan $\mathbf{P}(\boldsymbol{V}, \boldsymbol{T})$, the aligned image nodes $\mathcal{G}^v = \{\boldsymbol{g}_1^v, \boldsymbol{g}_2^v, \ldots \boldsymbol{g}_{n_p^2}^v\}$ can be acquired by:

$$\mathcal{G}^v = \sqrt{n_p^2} \times \left(\sum_{i=1}^{n_p^2} \mathbf{P}_{i1}\varphi(\boldsymbol{v}_i), \ldots, \mathbf{P}_{ip}\varphi(\boldsymbol{v}_{n_p^2}) \right) \\ = n_p \times \mathbf{P}(\mathbf{V}, \mathbf{T})^T \varphi(\mathbf{V}) \tag{3}$$

where the global visual representation $\boldsymbol{v}_g$ is appended as the first node of the aligned image nodes. To establish a topology structure for comprehensive alignment reasoning, we compute the candidate edge from $\boldsymbol{g}_i^v \in \mathcal{G}^v$ to $\boldsymbol{g}_j^v \in \mathcal{G}^v$ as:

$$e(\boldsymbol{g}_i^v, \boldsymbol{g}_j^v, \mathbf{W}_Q, \mathbf{W}_K) = \frac{exp((\mathbf{W}_Q \boldsymbol{g}_i^v)(\mathbf{W}_K \boldsymbol{g}_j^v))}{\sum_j exp((\mathbf{W}_Q \boldsymbol{g}_i^v)(\mathbf{W}_K \boldsymbol{g}_j^v))} \tag{4}$$

where $\mathbf{W}_Q \in \mathbb{R}^{d_h \times h}$ and $\mathbf{W}_K \in \mathbb{R}^{d_h \times h}$ are linear transformations for generating query and key vector to compute candidate edges. Note that during edge calculation process, our topology image graph is directed and thus allows complex information propagation.

Similarly, the aligned text nodes $\mathcal{G}^t = \{\boldsymbol{g}_1^t, \boldsymbol{g}_2^t, \ldots \boldsymbol{g}_{n_s}^t\}$ can be acquired based on the transportation plan $\mathbf{P}(\mathbf{T}, \mathbf{V})$ from text to image:

$$\mathcal{G}^t = \sqrt{n_s} \times \left(\sum_{i=1}^{n_s} \mathbf{P}_{i1}\varphi(\boldsymbol{t}_i), \ldots, \mathbf{P}_{ip}\varphi(\boldsymbol{t}_{n_s}) \right) \\ = \sqrt{n_s} \times \mathbf{P}(\mathbf{T}, \mathbf{V})^T \varphi(\mathbf{T}) \tag{5}$$

where the edge calculation of text nodes is based on equation 4. With the constructed TOT graphs, we perform dynamically topology reasoning to capture the inner-modal correspondence by iteratively updating graph nodes and connected edges. The nodes $\boldsymbol{x}_i^n$ at n-th step are updated as:

$$\overline{\boldsymbol{x}}_i^{n+1} = \sum_j e(\boldsymbol{x}_i^n, \boldsymbol{x}_j^n, \mathbf{W}_Q^n, \mathbf{W}_K^n) \cdot \boldsymbol{x}_j^n \tag{6}$$

$$\boldsymbol{x}_i^{n+1} = \text{ReLU}(\mathbf{W}_a^n \overline{\boldsymbol{x}}_i^{n+1} + b_a^n) \tag{7}$$

where $\boldsymbol{x}_i^n$ are initialized by taking the i-th node from graph $\mathcal{G}_v$ or $\mathcal{G}_t$ at n = 0, and $\mathbf{W}_Q^n, \mathbf{W}_K^n, \mathbf{W}_a^n, b_a^n$ are learnable parameters for n-th graph layer. After each step of topology reasoning, the current node $\boldsymbol{x}_i^n$ is replaced with $\boldsymbol{x}_i^{n+1}$. When N iterations stop, the global node is fed to a fully-connected layer to infer the reasoning score $\boldsymbol{s}_r \in \mathbb{R}^{d_h}$.

Residual Global Interaction

To complement the topology reasoning results and infer harmfulness from global perspective, we add residual connection between CLIP global representation and classification scores based on multi-head cross model attention.

Specifically, for the i-th head attention, the multimodal input is interacted based on the dot-product attention mechanism:

$$\text{ATT}_i(\boldsymbol{v}_g, \boldsymbol{t}_g) = \sigma \left(\frac{[\mathbf{W}_{Q_i} \boldsymbol{v}_g]^T [\mathbf{W}_{K_i} \boldsymbol{t}_g]}{\sqrt{d_h/m}} \right) \mathbf{W}_{V_i} \boldsymbol{t}_g \tag{8}$$

where $\{\mathbf{W}_{Q_i}, \mathbf{W}_{K_i}, \mathbf{W}_{V_i}\} \in \mathbb{R}^{d_h/m \times d_h}$ are learning parameters corresponding to queries, keys and values respectively. Then the output of m heads attention is concatenated together, followed by linear transformation and residual connection to get the multimodal representation m_r:

$$\boldsymbol{m}_r = \boldsymbol{t}_g + \mathbf{W}_m[\text{ATT}_1, \text{ATT}_2, ..., \text{ATT}_m] \tag{9}$$

With the global multimodal representation, a MLP layer is calculated to acquire the residual global score $\boldsymbol{s}_g \in \mathbb{R}^{d_h}$.

Training Objectives

With the residual global score $\boldsymbol{s}_g$ and topology reasoning score $\boldsymbol{s}_r$, we get the final multimodal hidden states $\boldsymbol{s}_f \in \mathbb{R}^{d_h}$ by weighted sum of both $\boldsymbol{s}_g$ and $\boldsymbol{s}_r$. Then the final hidden states are fed into a linear function followed by a softmax function for the classification:

$$p(y|\boldsymbol{s}_f) = \text{softmax}(\mathbf{W_c}[(1-\gamma) \cdot s_g + \gamma \cdot s_r] + b_c) \tag{10}$$

where $\mathbf{W}_c \in \mathbb{R}^{c \times d_h}$, c is the category number of dataset.

The model parameters are optimized through back propagation with minimizing the standard cross-entropy loss function as the objective:

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|} \sum_{k=1}^{|\mathcal{D}|} \mathbf{log} p(y^k | \mathbf{s}_m^k) \tag{11}$$

Experiment

To verify the superiority of our proposed approach, in this section we firstly describe the evaluation settings. Then we display the model performance on two benchmarks of our TOT against other state-of-the-art unimodal and multimodal approaches. Finally, we conduct qualitative studies to analyze TOT's superiority in aligning clues from multiple modalities. We also display the error cases to discuss limitations of our model.

Datasets Settings

Datasets We evaluate our model on two publicly available harmful memes detection datasets, Harm-C and Harm-P, which consist of real-world memes relate to COVID-19 and US politics, respectively. The statistic information of the two benchmarks are displayed in Table 1. To make a fair comparison with the previous work, we report three metrics: Accuracy (Acc), Macro-F1 (F1), and Macro-Averaged Mean Absolute Error (MMAE).

Dataset	Harmfulness	Train	Test	Valid	Total
Harm-C	Very	182	21	10	213
	Partially	882	103	51	1036
	Non	1949	230	116	2295
	Total	3013	354	177	3544
Harm-P	Very	216	25	17	258
	Partially	1270	148	69	1487
	Non	1534	182	91	1807
	Total	3020	355	177	3552

Table 1: Statistics of Harm-C and Harm-P dataset.

Modality	Model	2-class Harmful Meme Detection						3-class Harmful Meme Detection					
		Harm-C			Harm-P			Harm-C			Harm-P		
		Acc	F1	MMAE	Acc	F1	MMAE	Acc	F1	MMAE	Acc	F1	MMAE
Text	BERT	70.17	66.25	0.2911	80.12	78.35	0.1660	68.93	48.72	0.5591	74.55	54.08	0.7742
	RoBERTa	70.77	66.40	0.2923	80.74	78.64	0.1644	70.11	48.94	0.5584	74.87	54.33	0.7771
	Bertweet	71.32	67.30	0.2842	81.88	79.34	0.1573	71.65	49.82	0.5512	75.22	54.87	0.7652
Image	DenseNet	68.42	62.54	0.3125	74.05	73.68	0.1845	65.21	42.15	0.6326	71.80	50.98	0.8388
	ResNet	68.74	62.97	0.3114	73.14	72.77	0.1800	65.29	43.02	0.6264	71.02	50.64	0.8900
Multimodal	MMBT	73.48	67.12	0.3258	82.54	80.23	0.1413	68.08	50.88	0.6474	78.14	58.03	0.7008
	ViLBERT	78.53	78.06	0.1881	87.25	86.03	**0.1276**	75.71	48.82	0.5329	84.66	64.70	0.6982
	Visual BERT	81.36	80.13	0.1857	86.80	86.07	0.1318	74.01	53.85	0.5303	84.02	63.68	0.7020
	CLIP	78.10	77.64	0.2010	84.02	83.85	0.1508	71.05	45.55	0.5887	80.75	60.23	0.7058
	MOMENTA	**83.82**	**82.80**	**0.1743**	**89.84**	**88.26**	0.1314	**77.10**	**54.74**	**0.5132**	**87.14**	**66.66**	**0.6805**
Ours	TOH	82.16	81.36	0.1877	88.56	86.93	0.1428	75.44	53.08	0.5276	84.71	64.21	0.6957
	TOA	83.55	82.14	0.1813	89.17	87.44	0.1367	76.25	53.74	0.5233	85.34	65.39	0.6864
	COT	86.48	85.28	0.1697	90.24	90.88	0.1278	81.56	54.21	0.5147	86.37	69.04	0.6742
	TOT	**87.01**	**85.93**	**0.1634**	**91.55**	**91.29**	**0.1245**	**82.76**	**55.38**	**0.5027**	**88.61**	**71.54**	**0.6697**
	$\triangle_{\text{TOT-best\_model}}$	3.19	3.13	0.0109	1.71	3.02	0.0031	5.66	0.64	0.0105	1.47	4.88	0.0108

Table 2: Overall performance of proposed TOT model against a set of baselines. For Acc (↑) and F1 (↑), the larger values are better , while for MMAE(↓), smaller values are better.

Baselines For both Harm-C and Harm-P, we compare the model performance with a set of strong baselines:

- **BERT, RoBERTa, Bertweet:** For the language models, we choose influential pre-trained models: BERT (Kenton and Toutanova 2019), RoBERTa (Liu et al. 2019) and Bertweet (Nguyen, Vu, and Nguyen 2020) .
- **DenseNet, ResNet:** For the visual models, we choose two famous models pre-trained on ImageNet: DenseNet (Huang et al. 2017) and ResNet (He et al. 2016).
- **MMBT:** This is a multimodal bitransformer (Kiela et al. 2019), which is able to capture the intra-modal and the inter-modal dynamics of the two modalities.
- **ViLBERT:** This is a robust model trained on an intermediate multimodal objective (conceptual captions) (Sharma et al. 2018) with task-agnostic representations.
- **Visual BERT:** This is a visual BERT (Li et al. 2019) pre-trained on the COCO dataset (Lin et al. 2014).
- **MOMENTA:** This is a novel multimodal deep neural network using background context to detect hate.

Implementation Details For each meme, we take the $d_h = 512$ to extract representations. We limit the max lengths of image and text feature sequence by setting $n_p^2 = 49$ and $n_s = 77$ respectively, which adopts the same configuration as the pre-training process. For the Kernel Mapping, we take Gaussian kernel with $\epsilon = 0.1$ and set the max numbers of sinkhorn iteration for as 3. For the topology reasoning, we take a configuration of reason step $N = 3$ and a dimension $h = 256$ for node vectors.

Quatitative Analysis

In this section, we report the 2-class and 3-class detection results on Harm-C and Harm-P datasets to demonstrate the superiority of proposed approach from multiple perspectives. Note that we implement three variations (TOH, TOA, COT) from TOT model to demonstrate superiority. TOH is a variation by replacing the topology aligning module with heuristic aligning (Diao et al. 2021), and TOA replaces the topology aligning module with attentional aligning. COT directly concatenates the multimodal features.

2-class Detection Results The left part of Table 2 reports the experimental results on Harm-C and Harm-P as the binary case. Firstly, it can be observed Bertweet achieves relative outstanding performance in the unimodal approaches. A possible reason for the phenomenon could be due to the excellent social media capability of Bertweet, which is pretrained on Twitter corpus. Secondly, the multimodal approaches achieve overall superior performance against unimodel ones. And approaches which finetunes the large-scale pre-trained visual-linguistic models are better compared Intermidiate Fusion, as the pre-trained models are able to capture stronger inter-modal dynamics of the two modalities. Thirdly, we can see that our proposed TOT model outperforms the existing methods by a larger margin, with 3.13% absolute points of M-F1 improvement on Harm-C and 3.02% points on Harm-P. The other two metrics of Acc and MMAE also get improved by a competitive margin, demonstrating the effectiveness of our method.

3-class Detection Results To further validate models' ability in fine-grained harmful contents detection, we also report the 3-class classification results in Table 2. Similar phenomena can be observed compared to the binary case. Our proposed TOT outperforms other methods by a larger margin, with 5.66% absolute Accuracy points on Harm-C and 4.88% M-F1 points on Harm-P. It is worth noted that competitive performance can be also achieved by COT, which directly concatenates the alignment representations instead of building transported graphs, further validating the superiority of our models.

Ablation Studies

In this section, we have conducted extensive ablation studies to verify the effectiveness of the three modules implemented in our method, including optimal transported image/text graphs (OTI/OTT), dynamic topology reasoning (DTOR) and Residual Global Interaction (REG). All the comparative experiments are conducted in the binary classification task.

Datasets	OTT	OTI	DTOR	Acc	M-F1
Harm-C	×			86.41	85.33
		×		86.17	85.07
			×	85.78	84.46
	×		×	84.93	83.76
		×	×	84.82	84.64
	×	×	×	83.74	82.46
Harm-P	×			90.34	90.17
		×		90.07	89.88
			×	89.71	89.43
	×		×	87.46	87.27
		×	×	87.39	87.14
	×	×	×	86.73	86.24

Table 3: Ablation results of transported graphs.

Optimal Transported Graphs Ablation Table 3 shows the binary classification results of different settings for optimal transported graphs and topology graph reasoning module (DTOR). For the models without DTOR, we adopt a MLP layer instead to output the topology reasoning scores. Based on the results we can make observations that the two transported graphs OTT and OTI are both helpful for classification. OTT and OTI contribute to the Accuracy score at 1.08% and 1.09% absolute points in Harm-C. And when the two graphs are exploited together, they make further contributions by performing additional complementary effect. And the implementation of DTOR brings performance improvement at 1.2% average absolute points in accuracy based on OT graphs.

Residual Global Interaction Ablation This subsection explores the complementary effect of classification scores by ablating topology reasoning score and residual global score. We also report different training strategies for these two modules: joint learning and independent learning, which aims to further verify their complementary effect during learning. The ablation results are shown in Table 4. We can observe that the scores from these two modules together achieve better performance compared to any single score (1.57% points in Acc and 1.45% points in M-F1 for Harm-C dataset), which well confirm the effectiveness of REG module. Besides, the joint training process contributes more compared to separately training by an improvement of 0.59% in Acc, as the two modules perform better complementary effect when training jointly.

Datasets	DTOR	REG	Split	Joint	Acc	M-F1
Harm-C	✓		✓		85.44	84.48
		✓	✓		84.37	83.27
	✓	✓	✓		86.42	85.67
	✓	✓		✓	87.01	85.93
Harm-P	✓		✓		89.83	89.44
		✓	✓		87.27	87.03
	✓	✓	✓		91.15	90.87
	✓	✓		✓	91.55	91.29

Table 4: Abaltion results of reasoning modules.

Qualitative Results

The ability to formulate and capture cross-modal alignment for deciphering implicit hate is a core contribution in our TOT model. Toward this research goal, we design qualitative analysis including case study and visualizations.

Visualizations We visualize the edge values after N step iterations of topology reasoning between global text nodes and other candidate nodes. Figure 3 displays what information the ToT module will aggregate given different visual features as input queries. The representations of visual regions are acquired by taking the average values of all the image patches that contains the target. For the offensive memes, TOT module captures implicit alignment between the meaningful instances ('mascot of the democratic party') and the slovenly donkey in the picture. Moreover, negative samples are exploited as queries to make comparison. It is observed that when exploiting negative samples, the final global node randomly aggregates meaningless information, as the misaligned representations by transportation plans influence the performance of topology reasoning.

Case Study As shown in Figure 4, we display several memes from the leveraged datasets, together with the heat map of edge values between global visual nodes and other candidate nodes. The text information in first meme is an innocent description, while its picture shows offensive content, which satirizes 'The Democrat Party' in the sentence. The second meme disseminates extreme multimodal harm reflected by the 'CHINA VIRUS' in the text and the word 'JOKE' with double meaning, which is implicitly aligned with the vicious painted picture of Trump. The third picture has a discrimination harm. The two sentences seem to

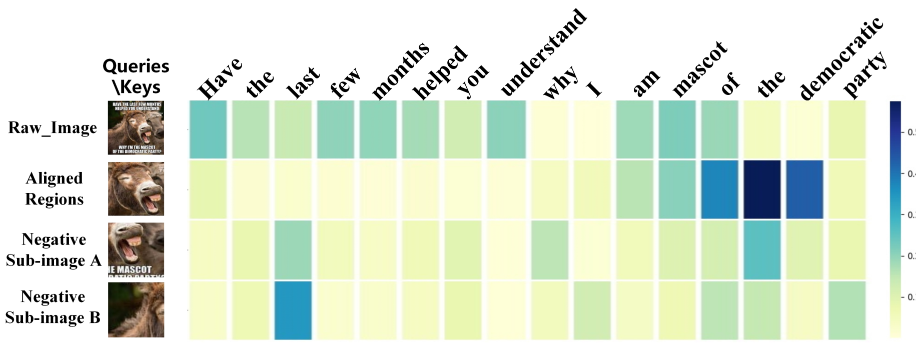

Figure 3: The visualizations of edge values in TOT. Raw_Image denotes the input meme image, Aligned Regions denotes the manually selected image regions containing alignments with text information for disseminating harm. The two negative samples are randomly selected sub-image incorporated for comparison.

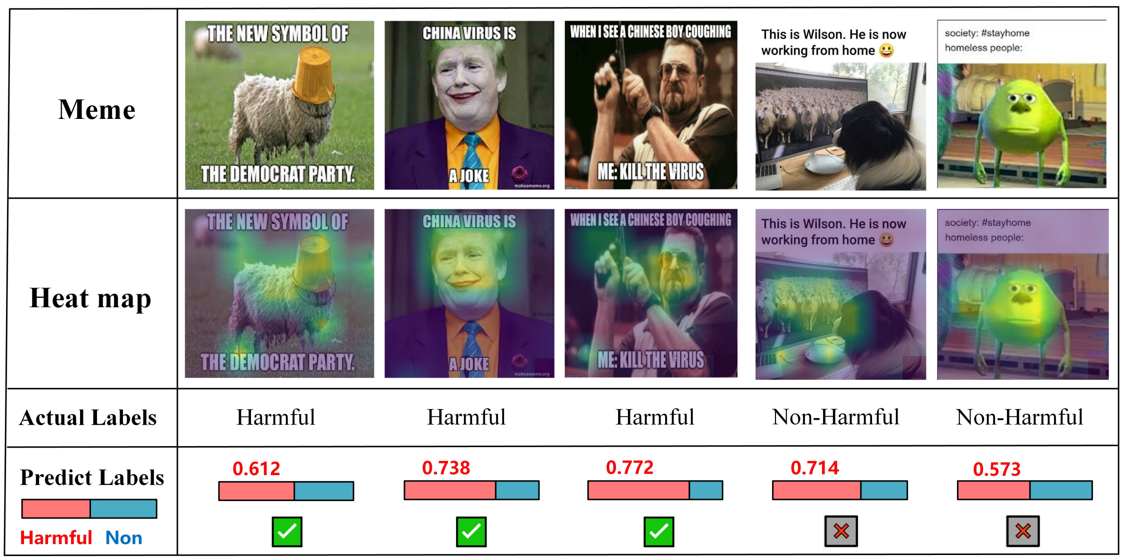

Figure 4: Case studies of memes from Harm-C and Harm-P datasets.

describe irrelevant things, but when they are put together with the image, in which a man is firing a gun, the alignment between 'KILL THE VIRUS' and 'A CHINESE BOY COUGHING' is established, and results in the extreme harm on race discrimination. These three cases demonstrate the necessity of aligning multimodal features.

Besides, we are also interested in exploring the limitations of our TOT. The wrongly classified memes are displayed in Figure 4. The fourth example is a funny meme, which shows a cute sheepdog watching the sheep in the screen to do his herding. Note that Wilson is the name of the dog. However, the anthropomorphized expression in the sentence, together with the dog pictures misleads our TOT to make wrong predictions, which is quite similar to the dehumanizing hateful content. The similar situation appears in the fifth meme, which may indicate that well cross-modal alignment is necessary but not adequate for detecting harm.

Conclusion

In this paper, we put forward a novel research problem of detecting implicit-aligned harmful memes, where this is the first work to incorporate such kernel methods with transportation theory in multimodal hate content detection. Specifically, topology-aware optimal transport (TOT) is conducted to formulate the cross-modal aligning problem as solutions for optimal transportation plans. With the dynamic topology reasoning module (DTOR), our TOT allows comprehensive information propagation based on the aligned information to make accurate reasoning. TOT is evaluated on two publicly available datasets, and our extensive experiments have shown that TOT outperforms the state-of-the-art baselines. We have also conducted visualizations and case studies to empirically demonstrated TOT's ability to capture implicit cross-modal alignment. While the error case shows that well cross-modal alignment is necessary but not adequate for detecting harm. We will incorporate a more advanced reasoning module for future improvements.

Acknowledgements

This research was funded by the National Natural Science Foundation of China (62206267), and the Strategic Priority Research Program of the Chinese Academy of Sciences grant (Y835120378).

References

Bretschneider, U.; and Peters, R. 2017. Detecting Offensive Statements towards Foreigners in Social Media. In *HICSS*, 1–10.

Chen, L.; Gan, Z.; Cheng, Y.; Li, L.; Carin, L.; and Liu, J. 2020. Graph Optimal Transport for Cross-Domain Alignment. In *ICML*, volume 119, 1542–1553.

Chen, L.; Wang, G.; Tao, C.; Shen, D.; Cheng, P.; Zhang, X.; Wang, W.; Zhang, Y.; and Carin, L. 2019a. Improving Textual Network Embedding with Global Attention via Optimal Transport. In *ACL*, 5193–5202.

Chen, Y.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2019b. UNITER: Learning UNiversal Image-TExt Representations. *CoRR*, abs/1909.11740.

Das, A.; Wahi, J. S.; and Li, S. 2020. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891*.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *AAAI*, 1218–1226.

Duan, J.; Chen, L.; Tran, S.; Yang, J.; Xu, Y.; Zeng, B.; and Chilimbi, T. 2022. Multi-modal Alignment using Representation Codebook. In *CVPR*, 15630–15639. IEEE.

Fortuna, P.; Ferreira, J.; Pires, L.; Routar, G.; and Nunes, S. 2018. Merging Datasets for Aggressive Text Identification. In *COLING*, 128–139.

Gan, Z.; Chen, Y.-C.; Li, L.; Zhu, C.; Cheng, Y.; and Liu, J. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33: 6616–6628.

Ge, Z.; Liu, S.; Li, Z.; Yoshie, O.; and Sun, J. 2021. OTA: Optimal Transport Assignment for Object Detection. In *CVPR*, 303–312.

Grave, E.; Joulin, A.; and Berthet, Q. 2019. Unsupervised Alignment of Embeddings with Wasserstein Procrustes. In *AISTATS*, volume 89, 1880–1890.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, 4700–4708.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Kiela, D.; Bhooshan, S.; Firooz, H.; and Testuggine, D. 2019. Supervised Multimodal Bitransformers for Classifying Images and Text. In *NeurIPS*.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *NeurIPS*, 2611–2624.

Lee, R. K.; Cao, R.; Fan, Z.; Jiang, J.; and Chong, W. 2021. Disentangling Hate in Online Memes. In *MM*, 5138–5147.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.; and Chang, K. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *CoRR*, abs/1908.03557.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, volume 8693, 740–755.

Lippe, P.; Holla, N.; Chandra, S.; Rajamanickam, S.; Antoniou, G.; Shutova, E.; and Yannakoudakis, H. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Malmasi, S.; and Zampieri, M. 2018. Challenges in discriminating profanity from hate speech. *J. Exp. Theor. Artif. Intell.*, 30(2): 187–202.

Mandl, T.; Modha, S.; Shahi, G. K.; Jaiswal, A. K.; Nandini, D.; Patel, D.; Majumder, P.; and Schäfer, J. 2021. Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages. *CoRR*, abs/2108.05927.

Maretic, H. P.; Gheche, M. E.; Chierchia, G.; and Frossard, P. 2022. fGOT: Graph Distances Based on Filters and Optimal Transport. In *AAAI*, 7710–7718.

Mehdad, Y.; and Tetreault, J. R. 2016. Do Characters Abuse More Than Words? In *SIGDIAL*, 299–303.

Mialon, G.; Chen, D.; d'Aspremont, A.; and Mairal, J. 2021. A Trainable Optimal Transport Embedding for Feature Aggregation and its Relationship to Attention. In *ICLR*.

Nguyen, D. Q.; Vu, T.; and Nguyen, A.-T. 2020. BERTweet: A pre-trained language model for English Tweets. In *EMNLP*, 9–14.

Pan, H.; Lin, Z.; Fu, P.; Qi, Y.; and Wang, W. 2020. Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection. In *Findings of EMNLP*, 1383–1392.

Pramanick, S.; Roy, A.; and Patel, V. M. 2022. Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection. In *WACV*, 546–556.

Pramanick, S.; Sharma, S.; Dimitrov, D.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *EMNLP*, 4439–4455.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Ross, B.; Rist, M.; Carbonell, G.; Cabrera, B.; Kurowsky, N.; and Wojatzki, M. 2017. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. *CoRR*, abs/1701.08118.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2556–2565.

Suryawanshi, S.; Chakravarthi, B. R.; Arcan, M.; and Buitelaar, P. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *TRAC@LREC*, 32–41.

Tekiroglu, S. S.; Chung, Y.; and Guerini, M. 2020. Generating Counter Narratives against Online Hate Speech: Data and Strategies. In *ACL*, 1177–1190.

Velioglu, R.; Rose, J.; and Analytics, S. D. 2020. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge. *arXiv preprint arXiv:2012.12975*.

Xu, J.; Zhou, H.; Gan, C.; Zheng, Z.; and Li, L. 2021. Vocabulary Learning via Optimal Transport for Neural Machine Translation. In *ACL*, 7361–7373.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI*, 10790–10797.

Zhang, Z.; Robinson, D.; and Tepper, J. A. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *ESWC*, 745–760.

Zhou, Y.; Chen, Z.; and Yang, H. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops*, 1–6.

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

Zia, H. B.; Castro, I.; and Tyson, G. 2021. Racist or sexist meme? classifying memes beyond hateful. In *WOAH*, 215–219.