Improving Multimodal Classification of Social Media Posts by Leveraging Image-Text Auxiliary Tasks

Danae Sánchez Villegas[1] **Daniel Preoţiuc-Pietro**[2] **Nikolaos Aletras**[1]
[1]University of Sheffield, [2]Bloomberg
{dsanchezvillegas1, n.aletras}@sheffield.ac.uk
dpreotiucpie@bloomberg.net

Abstract

Effectively leveraging multimodal information from social media posts is essential to various downstream tasks such as sentiment analysis, sarcasm detection or hate speech classification. Jointly modeling text and images is challenging because cross-modal semantics might be hidden or the relation between image and text is weak. However, prior work on multimodal classification of social media posts has not yet addressed these challenges. In this work, we present an extensive study on the effectiveness of using two auxiliary losses jointly with the main task during fine-tuning multimodal models. First, Image-Text Contrastive (ITC) is designed to minimize the distance between image-text representations within a post, thereby effectively bridging the gap between posts where the image plays an important role in conveying the post's meaning. Second, Image-Text Matching (ITM) enhances the model's ability to understand the semantic relationship between images and text, thus improving its capacity to handle ambiguous or loosely related modalities. We combine these objectives with five multimodal models across five diverse social media datasets, demonstrating consistent improvements of up to 2.6 F1 score. Our comprehensive analysis shows the specific scenarios where each auxiliary task is most effective.[1]

1 Introduction

Multimodal content including text and images is prevalent in social media platforms (Vempala and Preoţiuc-Pietro, 2019; Sánchez Villegas and Aletras, 2021). The content of both text and images has been widely used to improve upon single modality approaches in various downstream tasks such as sentiment analysis (Niu et al., 2016; Ju et al., 2021; Tian et al., 2023b), hate speech and rumor detection (Zhao et al., 2021; Hossain et al., 2022; Cao et al., 2022; Ocampo et al., 2023; Mu et al., 2023) and sarcasm detection (Xu et al., 2020; Liang et al., 2022; Ao et al., 2022; Tian et al., 2023a).

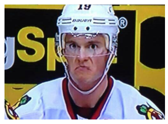
Post		
	When @USER gets more followers than you in 12 hours	My baby approves
Image-Text Relation	The image adds to the meaning	The image does not add to the meaning
Caption	A close up of a hockey player wearing a helmet	A gray and white chicken standing in the dirt

Figure 1: Image-text relations in social media posts from Vempala and Preoţiuc-Pietro (2019) and corresponding image captions generated with InstructBLIP. While image captions have a clear visual-language connection, image-text relationships in social media posts may no be apparent.

Multimodal classification methods for social media tasks often combine text and image representations obtained from pre-trained models. These are usually pre-trained on standard vision-language data such as image captions where strong image-text connections are assumed, i.e., captions that explicitly describe a corresponding image (Hessel and Lee, 2020; Xu and Li, 2022). Modeling text-image pairs from social media posts presents additional challenges. A notable difficulty lies in effectively capturing latent cross-modal semantics that may not be apparent. Figure 1 (left) shows an example where the text refers specifically to the mood of the person in the photo (i.e., "unhappy feeling" *when @USER gets more followers...*). Moreover, cases where the visuals are weakly related to the text are also prevalent (Xu et al., 2022). For instance, Figure 1 (right) shows an image of a hen accompanied by the text *My baby approves*. It is difficult to draw a direct relationship between the two without any

[1]Code is available here: https://github.com/danaesavi/SocialMedia-TextImage-Classification-AuxLosses.

Findings of the Association for Computational Linguistics: EACL 2024, pages 1126–1137
March 17-22, 2024 ©2024 Association for Computational Linguistics

additional context.

Multimodal models for social media classification can be divided into: (1) *single-stream* models where image and text features are concatenated together and fed into the same module such as Unicoder (Li et al., 2020), VisualBERT (Li et al., 2019), ViLT (Kim et al., 2021) and ALPRO (Li et al., 2022); and (2) *dual-stream* approaches where images and text are processed separately, e.g., ViLBert (Lu et al., 2019), LXMERT (Tan and Bansal, 2019), METER (Dou et al., 2022) and BLIP-2 (Li et al., 2023). Consequently, these models might still suffer from the aforementioned issues.

In this work, we examine the use of two tasks – Image-Text Contrastive (ITC) and Image-Text Matching (ITM) – as auxiliary losses during fine-tuning for improving social media post classification. By using the ITC contrastive loss (He et al., 2020; Li et al., 2021; Yu et al., 2022), we anticipate that when the image contributes to the post's meaning, as illustrated in Fig. 1 (left), the model will place them closer in the representation space. Conversely, ITM leverages binary classification loss for image-text alignment (Chen et al., 2020; Tan and Bansal, 2019; Wang et al., 2021). We expect that this will improve the model's ability to handle posts where associations may not be explicitly stated as shown in Fig. 1 (right). Although ITC and ITM have been used as pre-training objectives using generic images and their corresponding captions (Radford et al., 2021; Wang et al., 2021; Chen et al., 2022), their potential for enhancing fine-tuning in social media classification has yet to be explored.

Our main contributions are as follows: (1) we present an extensive study on comparing multimodal models jointly fine-tuned with ITC and ITM covering both *single-* and *dual-stream* approaches; (2) we show that models using ITC and ITM as auxiliary losses consistently improve their performance across five diverse multimodal social media datasets; (3) we offer a comprehensive analysis revealing the effectiveness of individual auxiliary tasks and their combination across various image-text relationship types in posts.

2 Multimodal Auxiliary Tasks

Image-Text Contrastive (ITC) Modeling text-image pairs in social media posts involves capturing hidden cross-modal semantics (Vempala and Preoţiuc-Pietro, 2019; Kruk et al., 2019). For instance, in Figure 1 (left) the visible mood of the person on the photo is related to the text of the post. Instead of directly matching images with textual descriptions (e.g., *a man wearing a helmet*), we aim to encourage the model to capture the dependencies between the image and text within the posts.

For this purpose, we use the ITC objective (He et al., 2020; Li et al., 2021; Yu et al., 2022) which pushes towards a feature space in which image and text representations of a post are brought closer together, while image and text representations that appear in different posts are pushed further apart. Let L_n and I_n be the n-th (normalized) representation of text and accompanying image of a post in a training batch. While the cosine similarity of the pair L_n and I_n is minimized, the cosine similarity of all other random pairs (e.g., L_n and I_m; I_m is an image from a different post in the current batch) is maximized. Given N posts within a training batch, ITC loss is defined as follows:

$$l_{ITC} = \frac{1}{2}(l_1 + l_2) \tag{1}$$

$$l_1 = -\frac{1}{N}\Sigma_{n=1}^{N} log \frac{exp(LI^T/e^{\tau})}{\Sigma_{j=1}^{N} exp(LI^T/e^{\tau})} \tag{2}$$

$$l_2 = -\frac{1}{N}\Sigma_{n=1}^{N} log \frac{exp(IL^T/e^{\tau})}{\Sigma_{j=1}^{N} exp(IL^T/e^{\tau})} \tag{3}$$

τ is a learnable temperature parameter to scale the logits (Jia et al., 2021).

Image-Text Matching (ITM) In social media posts, unrelated or weakly related text-image pairs are common (Hessel and Lee, 2020; Xu et al., 2022) such as the post depicted in Fig. 1 (right). To address this, we use the ITM objective (Chen et al., 2020; Tan and Bansal, 2019; Wang et al., 2021) during fine-tuning to understand the semantic correspondence between images and text. ITM involves a binary classification loss that penalizes the model when a given text and image do not appear together in a post. Let I_n and L_n be the image and text representation of the n-th post in a training batch, we randomly replace I_n with an image of another post from the current batch with a probability of 0.5 following (Wang et al., 2021; Kim et al., 2021). If I_n is replaced, then the image and text do not match, otherwise I_n and L_n match. Thus, the ITM loss corresponds to the cross-entropy loss for penalizing incorrect predictions, $l_{ITM} = -\Sigma_{i=1}^{2} t_i log(p_i)$ where t_i is the gold label (matched or mismatched) and p_i is the softmax probability for each label.

Joint Fine-tuning Objectives The joint fine-tuning loss function includes the cross-entropy classification loss (l_{CE}) and the two auxiliary training

Dataset	Classification Task	#	Train	Val	Test	All
TIR (Vempala and Preoţiuc-Pietro, 2019)	Text-Image Relation Classification	4	3,575	447	449	4,471
MVSA (Niu et al., 2016)	Sentiment Analysis	3	3,611	451	451	4,511
MHP (Gomez et al., 2020; Botelho et al., 2021)	Hate Speech Classification	4	3,998	500	502	5,000
MSD (Cai et al., 2019)	Sarcasm Detection	2	19,816	2,410	2,409	24,635
MICD (Sánchez Villegas et al., 2023)	Influencer Commercial Content Detection	2	11,377	1,572	1,435	14,384

Table 1: Description and statistics of each dataset. # refers to number of classes.

objectives defined as: $l_{C+M} = \lambda_1 l_{CE} + \lambda_2 l_{ITC} + \lambda_3 l_{ITM}$, where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters to control the influence of each loss.

3 Experimental Setup

3.1 Datasets

We experiment with five diverse multimodal public datasets in English: (1) **TIR** – text-image relationship categorization (Vempala and Preoţiuc-Pietro, 2019); (2) **MVSA** – multi-view sentiment analysis (Niu et al., 2016); (3) **MHP** – multimodal hate speech detection (Gomez et al., 2020; Botelho et al., 2021); (4) **MSD** – multimodal sarcasm detection (Cai et al., 2019): and (5) **MICD** – multimodal commercial influencer content detection (Sánchez Villegas et al., 2023). Table 1 presents dataset statistics.

3.2 Single Modality Methods

Text-only We fine-tune **BERT** (Devlin et al., 2019) and **Bernice** (DeLucia et al., 2022), a BERT based model pre-trained on a corpus of multilingual tweets. We also experiment with few-shot (FS) prompting using **Flan-T5** (Chung et al., 2022) and **GPT-3** (Brown et al., 2020). For each dataset, we construct a few-shot prompt and include two randomly selected training examples for each class.[2]

Image-only We fine-tune **ResNet**152 (He et al., 2016) and **ViT** (Dosovitskiy et al., 2020), both pretrained on ImageNet (Russakovsky et al., 2015). We experiment with few-shot prompting using **IDEFICS** (Laurençon et al., 2023) and zero-shot prompting using **InstructBLIP** (Dai et al., 2023). Prompts include two randomly chosen image-only training examples per class (see Appx. B).

3.3 Multimodal Models

Ber-ViT We use Bernice and ViT to obtain representations of the text (L) and image (I). **Ber-ViT-Conc** appends the text and image vectors from the corresponding L and I [CLS] tokens to obtain the multimodal representation h^{LI}; **Ber-ViT-Att** computes cross-attention between L and I. h^{LI} is obtained by appending the [CLS] token from L and the [CLS] token from the attention layer. We fine-tune each model by adding a classification layer.

MMBT (Kiela et al., 2019). Image embeddings obtained from Resnet152 are concatenated with token embeddings and passed to a BERT-like transformer. The [CLS] token is used as the multimodal representation (h^{LI}) for classification.

LXMERT (Tan and Bansal, 2019) consists of three encoders and their corresponding outputs for vision I, language L, and a multimodal vector h^{LI}.

ViLT We fine-tune ViLT (Dosovitskiy et al., 2020) and extract the multimodal h^{LI} that corresponds to the first token from the last hidden state.

ITC and ITM Inputs The ITC auxiliary task inputs are the corresponding text and image vectors of each model. The ITM auxiliary task input is the respective multimodal representation h^{LI}.

3.4 Evaluation

Results are obtained over three runs using different random seeds reporting average and standard deviation. We use weighted F1 for model evaluation following standard practice on the TIR, MHP and MICD datasets to manage class imbalance.[3]

4 Results

4.1 Performance Comparison

Image-text auxiliary tasks improve multimodal classification. Table 2 shows that multimodal models surpass single-modality approaches across all datasets. We consistently find performance gains when using either ITC, ITM, or both auxiliary losses during fine-tuning, with improvements up to

[2]Appx. B shows the prompt templates.

[3]Implementation details are included in Appx. A.

Model	TIR	MVSA	MHP	MSD	MICD	▲
Majority Class	16.0	59.8	53.4	45.2	48.0	-
Text-only Models						
BERT	$37.2_{1.3}$	$70.1_{0.8}$	$73.3_{1.3}$	$83.9_{0.2}$	$74.3_{0.6}$	-
Bernice	$38.9_{1.1}$	$71.6_{0.6}$	$73.6_{0.6}$	$84.5_{0.8}$	$74.5_{2.2}$	-
Flan-T5*	$3.8_{0.0}$	$58.9_{0.0}$	$46.5_{1.3}$	$59.6_{2.2}$	$48.7_{1.6}$	-
GPT-3*	$16.3_{6.1}$	$55.9_{0.1}$	$58.2_{4.6}$	$69.6_{2.7}$	$69.6_{1.5}$	-
Image-only Models						
ResNet152	$48.2_{0.0}$	$63.8_{0.1}$	$51.8_{5.8}$	$46.9_{0.1}$	$59.6_{0.5}$	-
ViT	$51.4_{1.3}$	$68.2_{0.6}$	$57.2_{1.2}$	$71.5_{0.1}$	$60.8_{1.3}$	-
IDEFICS*	$12.4_{3.6}$	$34.7_{6.1}$	$34.9_{2.7}$	$58.9_{2.4}$	$35.6_{0.0}$	-
InstructBLIP*	$3.9_{0.0}$	$47.2_{0.0}$	$11.0_{0.0}$	$22.7_{0.0}$	$35.6_{0.0}$	-
Multimodal Models						
Ber-ViT-Conc	$43.6_{1.2}$	$70.4_{0.0}$	$76.6_{0.6}$	$88.8_{0.0}$	$75.5_{1.9}$	-
+C	$44.9_{0.7}$	$72.0^{\dagger}_{0.2}$	$77.3_{1.1}$	$\underline{\mathbf{89.7}}^{\dagger}_{0.0}$	$\underline{77.2}_{0.4}$	1.2
+M	$44.1_{0.2}$	$\underline{73.6}^{\dagger}_{0.9}$	$\underline{77.8}_{0.6}$	$89.2^{\dagger}_{0.1}$	$76.1_{0.8}$	1.2
+C+M	$\underline{45.8}_{0.8}$	$73.4^{\dagger}_{0.4}$	$77.7^{\dagger}_{0.6}$	$\underline{\mathbf{89.7}}^{\dagger}_{0.2}$	$76.3_{0.5}$	1.6
Ber-ViT-Att	$53.7_{1.0}$	$72.1_{0.7}$	$76.8_{0.5}$	$88.8_{0.3}$	$75.6_{0.8}$	-
+C	$54.8_{0.8}$	$72.8_{0.2}$	$77.5_{0.6}$	$89.5^{\dagger}_{0.5}$	$\underline{\mathbf{77.8}}^{\dagger}_{0.5}$	0.8
+M	$\underline{\mathbf{55.9}}^{\dagger}_{0.8}$	$73.5^{\dagger}_{0.2}$	$77.4_{0.6}$	$89.4_{0.5}$	$76.6_{0.5}$	1.2
+C+M	$54.6_{0.7}$	$\underline{\mathbf{74.6}}^{\dagger}_{0.3}$	$\underline{\mathbf{78.0}}^{\dagger}_{0.1}$	$\underline{\mathbf{89.7}}^{\dagger}_{0.3}$	$76.3_{0.2}$	1.7
MMBT	$53.2_{1.2}$	$72.4_{0.4}$	$74.5_{0.5}$	$83.2_{0.0}$	$73.6_{0.4}$	-
+C	$\underline{53.7}_{1.1}$	$73.2_{1.0}$	$75.7_{1.7}$	$\underline{84.4}^{\dagger}_{0.3}$	$74.1_{0.8}$	1.1
+M	$\underline{53.7}_{0.7}$	$73.4_{0.8}$	$75.4_{1.3}$	$84.3^{\dagger}_{0.3}$	$\underline{74.8}^{\dagger}_{0.6}$	0.9
+C+M	$53.6_{0.2}$	$\underline{73.5}^{\dagger}_{0.0}$	$\underline{75.7}_{1.2}$	$83.4_{0.2}$	$73.8_{0.5}$	0.6
LXMERT	$51.3_{0.5}$	$68.2_{1.1}$	$70.7_{0.8}$	$81.9_{0.5}$	$69.9_{1.0}$	-
+C	$51.9_{0.3}$	$\underline{70.4}^{\dagger}_{0.5}$	$\underline{72.1}^{\dagger}_{0.2}$	$\underline{82.7}_{0.1}$	$70.8_{0.5}$	1.2
+M	$51.8_{0.4}$	$69.5_{0.2}$	$71.8_{0.8}$	$82.3_{0.5}$	$\underline{70.9}_{0.2}$	0.9
+C+M	$\underline{52.3}_{1.4}$	$69.3_{0.9}$	$71.9_{1.7}$	$82.1_{0.4}$	$70.3_{0.3}$	0.8
ViLT	$53.1_{1.1}$	$70.5_{1.3}$	$71.8_{0.0}$	$83.0_{0.8}$	$67.8_{1.6}$	-
+C	$\underline{55.7}^{\dagger}_{0.2}$	$\underline{72.9}_{1.0}$	$72.5^{\dagger}_{0.4}$	$83.4_{0.4}$	$68.3_{0.2}$	1.3
+M	$\underline{55.7}^{\dagger}_{0.3}$	$72.1_{2.3}$	$72.0_{0.5}$	$\underline{83.5}_{0.2}$	$68.7_{1.1}$	1.1
+C+M	$55.3^{\dagger}_{0.3}$	$\underline{72.9}_{1.3}$	$\underline{73.4}_{1.4}$	$83.2_{0.4}$	$\underline{70.0}_{1.3}$	1.7

Table 2: Results in weighted F1 for all datasets. Best results for each base multimodal model are underlined and best results for each dataset are in bold. † indicates statistically significant improvement (t-test, $p < 0.05$) over the corresponding base model. Subscripts denote standard deviation over three runs. ▲ refers to the average relative improvement over each base model across datasets.* denotes prompting. +C,+M, C+M refer to +ITC, +ITM and +ITC+ITM.

2.6 F1 over each base model. Therefore, we can improve performance without costly pre-training on social media text-image tasks. These findings are especially valuable in multimodal computational social science studies, where grasping the interplay between text and images is vital (Sánchez Villegas et al., 2021; Xu et al., 2022).

Dual-stream methods are effective in leveraging information from the auxiliary tasks. Across MVSA, MHP and MSD datasets, the Ber-ViT-Att+C+M model achieves the best performance (74.6, 78.0, and 89.7 F1 respectively). Generally, we observe that both ITC and ITM contribute to the performance improvements of Ber-ViT-Att. Overall, Ber-ViT-Att+C and Ber-ViT-Att+M models average improvements over the base model across datasets are 0.8 and 1.2 respectively, while Ber-ViT-Att+C+Mimprovement is 1.7. The performance gap between *dual-* and *single-stream* models is narrower in TIR. ViLT+M achieves 55.7 F1 while Ber-ViT-Att+M obtains 55.9. This is likely due to the importance of visual information for this task (i.e., predicting the semiotic relationship between images and text), which is better aligned with ViLT as a visual-based model.

4.2 Training with different number of samples

To test the generalizability and data efficiency of our models, we conduct experiments using our best performing model, Ber-ViT-Att, across different training data sizes, thus simulating low resource scenarios. We assessed the weighted F1 scores of Ber-ViT-Att both independently and with the incorporation of each auxiliary loss, as well as a combination of both. The results of these experiments are presented in Figure 2. While Table 2, highlights that the highest performance is generally achieved using both auxiliary losses, in Figure 2 we observe the best performing models are predominantly distributed between Ber-ViT-Att+C and Ber-ViT-Att+C+M.

We find that the difference between training with 20% of random examples and using the entire dataset is modest in some cases, particularly when fine-tuning with both ITC and ITM losses on MVSA, MSD, and MICD. Specifically, for MSD the difference is 6.8 F1 points, while for MVSA and MICD, it is less than 5 F1 points. These results suggest that our models exhibit robust generalization. However, MHP exhibits a more substantial difference, with a gap of 21.6 F1 points when Ber-ViT-Att is trained with 20% of the training examples, narrowing to 14.1 F1 points with Ber-ViT-Att+C. This suggests the viability of employing ITC as an auxiliary loss during fine-tuning for hate speech classification in low-resource scenarios.

5 Analysis

We analyze Ber-ViT-Att's predictions on TIR to understand when each auxiliary task benefits different image-text relations as categorized by Vempala and Preoţiuc-Pietro (2019) based on image contribution and text representation (Figure 3 and 4).

When the text is represented in the image using both auxiliary tasks (models denoted with +C+M), the model achieves the best performance,

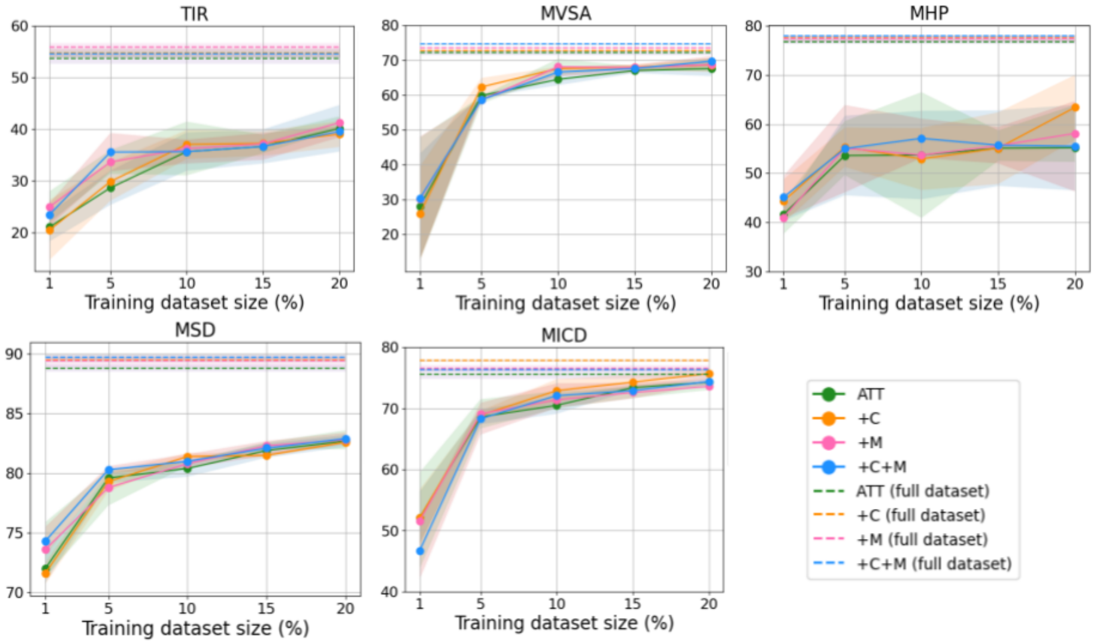


Figure 2: Results in weighted F1 using Ber-ViT-Att (ATT) for all datasets when training with different percentages of training data. We plot the mean and standard deviation across three runs.

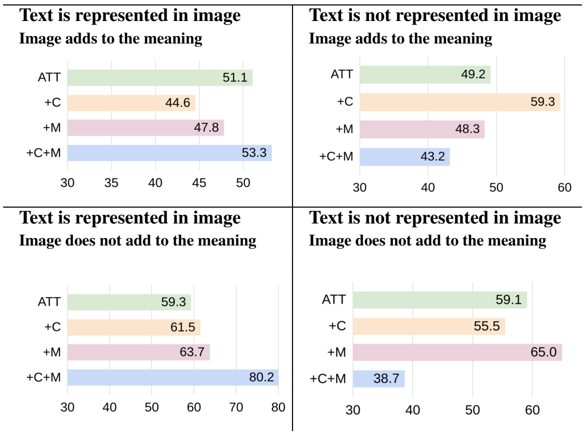


Figure 3: Accuracy per label using Ber-ViT-Att (ATT) across different image-text relation types based on image contribution to the post's meaning and text representation on the image.

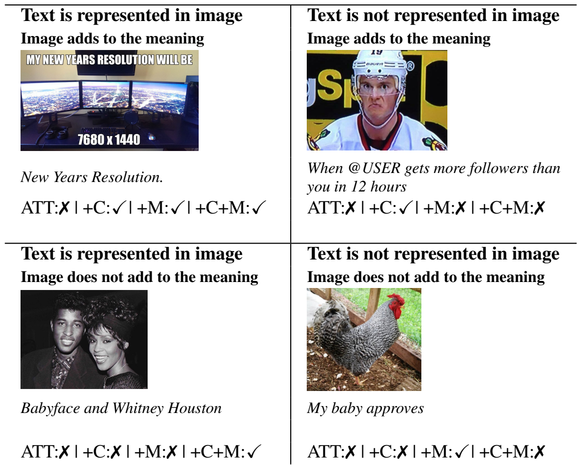


Figure 4: Bert-ViT-Att (ATT) predictions on randomly selected examples with varying image-text relations.

especially when the visual content is not semantically relevant to the post. We observe that 80.2% of the tweets are correctly classified achieving a substantial improvement over the Ber-ViT-Att baseline where only 59.3% of the posts are correctly classified.

When text is not represented on the image, we find that including ITC performs best when the visual content is relevant, with 59.3% of the tweets correctly classified compared to 49.2% using Ber-ViT-Att. Finally, in cases where the image does not enhance the semantic meaning, Ber-ViT-Att+M exhibits the highest performance, correctly classifying 65% of the posts. This validates our hypothesis that incorporating ITM helps models to effectively identify posts with weaker image-text relationships.

6 Conclusion

We presented an extensive study on the effectiveness of using two auxiliary tasks, Image-Text Contrastive and Image-Text Matching when fine-tuning multimodal models for social media posts classification. This approach addresses the challenges of hidden cross-modal semantics and weak image-text relationships in social media content.

Limitations

First, the datasets used in our experiments are solely in English. This choice allows for consistency and comparability across the datasets, but it does not test the generalizability of our findings to other languages. In future work, we plan to extend our research to a multilingual setting to address this limitation. The effectiveness of the models incorporating auxiliary tasks depends on the underlying base model, however, our approach can easily be adapted to new models. Finally, the inclusion of auxiliary tasks in our models introduces an increase in training time. For instance, the training time for Ber-ViT-Att on the TIR dataset is approximately 1.5 hours on an Nvidia A100 GPU. When incorporating the auxiliary tasks (Ber-ViT-Att+C+M), the training time extends to around 2.5 hours, a 66% relative increase in training time. However, the additional time is a one-time occurrence and relatively minor when compared to the pre-training times of large language models (LLMs).

Experiments on TIR dataset. We align with previous work on the TIR dataset by employing text-only and image-only models for classification (Vempala and Preoţiuc-Pietro, 2019), with the expectation that specific textual cues or image content can indicate relationships, even without considering the image content. For instance, (a) tweets concluding with an ellipsis or brief comments may serve as predictive indicators that the text is not represented in the accompanying image, and (b) images featuring people may be more likely to contain text corresponding to the names of those individuals. While unimodal models may not be ideal choices in real-world scenarios for this task, they serve as valuable performance baseline.

Acknowledgments

DSV and NA are supported by the Leverhulme Trust under Grant Number: RPG#2020#148. DSV is also supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by the UK Research and Innovation grant EP/S023062/1. We would like to thank Katerina Margatina, Mali Jin, Constantinos Karouzos, and all reviewers for their valuable feedback.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Xiao Ao, Danae Sanchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2022. Combining humor and sarcasm for improving political parody detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1800–1807, Seattle, United States. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Austin Botelho, Scott Hale, and Bertie Vidgen. 2021. Deciphering implicit hate: Evaluating automated detection algorithms for multimodal hate. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1896–1907, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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, Florence, Italy. Association for Computational Linguistics.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Learning universal image-text representations. *European Conference on Computer Vision*.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al.

2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176.

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*, pages 1470–1478.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4395–4405, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in Instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632, Hong Kong, China. Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. 2022. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

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*.

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, Dublin, Ireland. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Yida Mu, Kalina Bontcheva, and Nikolaos Aletras. 2023. It's about time: Rethinking evaluation on rumor detection benchmarks using chronological splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 736–743, Dubrovnik, Croatia. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*, pages 15–27. Springer.

Nicolas Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772, Toronto, Canada. Association for Computational Linguistics.

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. PMLR.

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):1137–1149.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Danae Sánchez Villegas and Nikolaos Aletras. 2021. Point-of-interest type prediction using text and images. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7785–7797, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danae Sánchez Villegas, Catalina Goanta, and Nikolaos Aletras. 2023. A multimodal analysis of influencer content on twitter. *arXiv preprint arXiv:2309.03064*.

Danae Sánchez Villegas, Saeid Mokaram, and Nikolaos Aletras. 2021. Analyzing online political advertisements. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3669–3680, Online. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023a. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2468–2480, Toronto, Canada. Association for Computational Linguistics.

Yuanhe Tian, Weidong Chen, Bo Hu, Yan Song, and Fei Xia. 2023b. End-to-end aspect-based sentiment analysis with Combinatory Categorial Grammar. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13597–13609, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Alakananda Vempala and Daniel Preoţiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of Twitter posts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2830–2840, Florence, Italy. Association for Computational Linguistics.

Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Chunpu Xu and Jing Li. 2022. Borrowing human senses: Comment-aware self-training for social media multimodal classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5644–5656, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chunpu Xu, Hanzhuo Tan, Jing Li, and Piji Li. 2022. Understanding social media cross-modality discourse in linguistic space. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2459–2471, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2021. A comparative study of using pre-trained language models for toxic comment classification. In *Companion Proceedings of the Web Conference 2021*, pages 500–507.

A Implementation details

A.1 Data Processing

Text For each tweet, we lowercase and tokenize text using the NLTK Twitter tokenizer (Bird and Loper, 2004). We also replace URLs and user @-mentions with placeholder tokens. Emojis are replaced with their corresponding text string, e.g thumbs_up following Nguyen et al. (2020).

Image Images are resized to (224×224) pixels representing a value for the red, green and blue color in $[0, 255]$. The pixel values are normalized to $[0-1]$. For LXMERT (Tan and Bansal, 2019) in Section 3.3, we extract *object-level* features using Faster-RCNN (Ren et al., 2016) as in Anderson et al. (2018) and keep 36 objects for each image as in Tan and Bansal (2019).

A.2 Data Splits

We use the same data splits for MVSA, MHP, MSD, and MICD as in the original papers. For TIR, instead of a 10-fold cross-validation, we randomly split the data in 80%, 10%, and 10% for training, validation, and testing for consistency with the other tasks.

A.3 Hyperparameters

We select the hyperparameters for all models using early stopping by monitoring the validation loss. We use the Adam optimizer (Kingma and Ba, 2014). We estimate the class weights using the 'balanced' heuristic (King and Zeng, 2001). All experiments are performed using an Nvidia A100 GPU with a batch size of 8 for TIR and MHP and 16 for MVSA and MSD datasets. For prompting implementation details see Appx. B.

Image-only For ResNet152 (He et al., 2016), we fine-tune for 1, 5, 8, 6 and 1 epochs for TIR, MVSA, MHP, MSD and MICD datasets respectively, with learning rate $\eta = 1e^{-5}$ and dropout $\delta = 0.05$ before passing the image representation through the classification layer. We fine-tune ViT (Dosovitskiy et al., 2020) for 3 epochs for TIR, MSD and MICD and 10 epochs for MVSA and MHP datasets with learning rate $\eta = 1e^{-5}$ and dropout $\delta = 0.05$. $\eta \in \{1e^{-3}, 1e^{-4}, 1e^{-5}\}$ and δ in $[0, 0.5]$, random search.

Text-only Transformers We fine-tune BERT and Bernice for 20 epochs and choose the epoch with the lowest validation loss. We use the pre-trained base-uncased model for BERT (Vaswani et al., 2017; Devlin et al., 2019) from the Hugging Face library (12-layer, 768-dimensional) (Wolf et al., 2019), and the base model for Bernice (DeLucia et al., 2022) with a maximal sequence length of 128. We fine-tune BERT for 3, 9, 5, 2 and 1 epochs for TIR, MVSA, MHP, MSD and MICD with learning rate $\eta = 1e^{-5}$ and dropout $\delta = 0.05$; and Bernice for 3, 4, 7, 3 and 3 epochs for TIR, MVSA, MHP, MSD and MICD datasets, $\eta = 1e^{-5}$ and $\delta = 0.05$. For all models $\eta \in \{2e^{-5}, 1e^{-4}, 1e^{-5}\}$ and $\delta \in [0, 0.5]$, random search.

Multimodal Predictive Models We train MMBT (Kiela et al., 2019), ViLT (Kim et al., 2021), LXMERT (Tan and Bansal, 2019) and Bernice-ViT models with λ_1, λ_2, λ_3; λ_2 and $\lambda_3 \in [0, 1.5]$ (as explained in Section 2), and number of fine-tuning epochs (E) for each model as shown in Table 4. For ViLT models we keep the vision layers frozen and we use a learning rate of $\eta = 1e^{-4}$, dropout $\delta = 0.05$ and weight decay of 0.0002. For all other multimodal models we use a learning rate of $\eta = 1e^{-5}$, dropout $\delta = 0.05$ and weight decay of 0.00025.

B Prompting

For each dataset, we construct a prompt to include two randomly selected training examples for each class (GPT-3, FLAN-T5, IDEFICS) as follows:

- TIR (GPT-3 & FLAN-T5)

  Label the next text as 'image adds and text is represented', 'image adds and text is not represented', 'image does not add and text is represented', 'image does not add and text is not represented'. Text: <TWEET-TRAIN> // <LABEL-TRAIN> $\times 8$
  Label the next text as 'image adds and text is represented', 'image adds and text is not represented', 'image does not add and text is represented', 'image does not add and text is not represented'. Text: <TWEET> //

- TIR (IDEFICS)

  User: <IMAGE-TRAIN> Label the image as 'image adds and text is represented', 'image adds and text is not represented', 'image does not add and text is represented', 'image does not add and text is not represented'. Assistant:<LABEL-TRAIN> $\times 8$
  User: <IMAGE-TEST> Label the image as 'image adds and text is represented', 'image adds and text is not represented', 'image does not add and text is represented', 'image does not add and text is not represented'. Assistant:

- TIR (InstructBLIP)
  - Prompt: *Label the image as 'image adds and text is represented', 'image adds and text is not represented', 'image does not add and text is represented', 'image does not add and text is not represented'*
  - Image: *<IMAGE-TEST>*

- MVSA (GPT-3 & FLAN-T5)

  Label the next text as 'positive' or 'negative' or 'neutral'. Text: <TWEET-TRAIN> // <LABEL-TRAIN> $\times 6$
  Label the next text as 'positive' or 'negative' or 'neutral'. Text: <TWEET> //

- MVSA (IDEFICS)

  User: <IMAGE-TRAIN> Is the sentiment of the image 'positive' or 'negative' or 'neutral'?. Assistant:<LABEL-TRAIN> $\times 6$
  User: <IMAGE-TEST> Is the sentiment of the image 'positive' or 'negative' or 'neutral'?. Assistant:

- MVSA (InstructBLIP)
  - Prompt: *Is the sentiment of the image 'positive' or 'negative' or 'neutral'?*
  - Image: *<IMAGE-TEST>*

- MHP

  Label the next text as 'hateful', 'counterspeech', 'reclaimed' or 'none'. Text: <TWEET-TRAIN> // <LABEL-TRAIN> $\times 8$
  Label the next text as 'hateful', 'counterspeech', 'reclaimed' or 'none'. Text: <TWEET> //

- MHP (IDEFICS)

  User: <IMAGE-TRAIN> Is the image 'hateful', 'counterspeech', 'reclaimed' or 'none'?. Assistant:<LABEL-TRAIN> $\times 8$
  User: <IMAGE-TEST> Is the image 'hateful', 'counterspeech', 'reclaimed' or 'none'?. Assistant:

- MHP (InstructBLIP)
  - Prompt: *Is the image 'hateful', 'counterspeech', 'reclaimed' or 'none'?*
  - Image: *<IMAGE-TEST>*

- MSD (GPT-3 & FLAN-T5)

  Label the next text as 'sarcastic' or 'not sarcastic'. Text: <TWEET-TRAIN> // <LABEL-TRAIN> $\times 4$
  Label the next text as 'sarcastic' or 'not sarcastic'. Text: <TWEET> //

- MSD (IDEFICS)

  User: <IMAGE-TRAIN> Is the image 'sarcastic' or 'not sarcastic'? Assistant:<LABEL-TRAIN> $\times 4$
  User: <IMAGE-TEST> Is the image 'sarcastic' or 'not sarcastic'? Assistant:

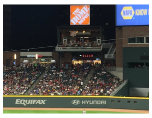
Dataset	Text	Image	Label	Outputs
MVSA	So proud of these kids! Not only talented, ENERGETIC and hardworking, but respectful and kind-hearted!		positive	GPT-3:positive Flan-T5: positive IDEFICS: positive InstructBLIP: positive
MSD	Text: it's the insensitive strikeouts at suntrust park. #braves #chopchop		sarcastic	GPT-3: sarcastic Flan-T5: sarcastic IDEFICS: not sarcastic InstructBLIP: not sarcastic

Table 3: Text-Image examples and corresponding labels assigned by each LLM model for MVSA (sentiment analysis) and MSD (sarcasm detection) datasets. For each model we use the prompt templates included in Appendix B.

Dataset	**TIR**		**MVSA**		**MHP**		**MSD**		**MICD**	
	$\lambda_1, \lambda_2, \lambda_3$	E	$\lambda_1, \lambda_2, \lambda_3$	E	$\lambda_1, \lambda_2, \lambda_3$	E	$\lambda_1, \lambda_2, \lambda_3$	E	$\lambda_1, \lambda_2, \lambda_3$	E
Ber-ViT-Conc	-	3	-	7	-	7	-	1	-	2
Ber-ViT-Conc+C	0.9, 0.1, 0	3	0.9, 0.1, 0	5	0.9, 0.1, 0	7	0.9, 0.1,0	6	0.9,0.1,0	2
Ber-ViT-Conc+M	0.9, 0, 0.1	4	0.9, 0, 0.1	6	0.9, 0, 0.1	9	0.9, 0, 0.1	3	0.9,0,0.1	1
Ber-ViT-Conc+C+M	0.8, 0.1, 0.1	6	0.8, 0.1, 0.1	4	0.8, 0.1, 0.1	6	0.8, 0.1, 0.1	3	0.8,0.1,0.1	2
Ber-ViT-Att	-	2	-	8	-	7	-	1	-	3
Ber-ViT-Att+C	0.9, 0.1,0	2	0.9, 0.1, 0	8	0.9,0.1,0	7	0.9, 0.1, 0	3	0.9,0.1,0	2
Ber-ViT-Att+M	0.92, 0, 0.08	3	0.9, 0, 0.1	6	0.9,0,0.1	6	0.9, 0, 0.1	3	0.9,0,0.1	1
Ber-ViT-Att+C+M	0.8, 0.1, 0.1	4	0.8, 0.1, 0.1	15	0.8,0.1,0.1	13	0.8, 0.1, 0.1	5	0.8,0.1,0.1	2
MMBT	-	2	-	9	-	5	-	1	-	1
MMBT+C	0.9, 0.1, 0	4	0.9, 0.1, 0	5	0.9, 0.1, 0	9	0.9,0.1,0	3	0.9,0.1,0	2
MMBT+M	0.9, 0, 0.1	4	0.7, 0 ,0.3	6	0.9, 0, 0.1	9	0.82, 0, 0.08	4	0.9,0,0.1	2
MMBT+C+M	0.84, 0.08, 0.08	3	0.85, 0.1, 0.05	11	0.8, 0.1, 0.1	10	0.85,0.1,0.05	3	0.6,0.2,0.2	4
LXMERT	-	2	-	5	-	5	-	2	-	3
LXMERT+C	0.9,0.1,0	2	0.9,0.1,0	8	0.9, 0.1, 0	5	0.9,0.1,0	2	0.9,0.1,0	2
LXMERT+M	0.85,0,0.15	1	0.9,0,0.1	6	0.8, 0, 0.1	12	0.85,0,0.15	2	0.9,0,0.1	3
LXMERT+C+M	0.9, 0.08, 0.02	2	0.83,0.02,0.15	7	0.8, 0.1, 0.1	11	0.85, 0.1, 0.05	2	0.8,0.1,0.1	3
ViLT	-	6	-	5	-	4	-	1	-	4
ViLT+C	0.9, 0.1, 0	6	0.9, 0.1, 0	11	0.9, 0.1, 0	4	0.9, 0.1, 0	1	0.95,0.05,0	2
ViLT+M	0.85, 0, 0.15	5	0.9,0,0.1	3	0.9, 0, 0.1	7	0.9, 0, 0.1	2	0.92,0,0.08	2
ViLT+C+M	0.8, 0.1, 0.1	2	0.8, 0.1, 0.1	13	0.8, 0.1, 0.1	9	0.8, 0.1, 0.1	2	0.87,0.05,0.08	1

Table 4: Hyperaprameter values for λ_1, λ_2, λ_3 as explained in Section 2, and number of fine-tuning epochs (E) for each model.

- MSD (InstructBLIP)
  - Prompt: *Is the image 'sarcastic' or 'not sarcastic'?*
  - Image: *<IMAGE-TEST>*

- MICD (GPT-3 & FLAN-T5)

  Label the next text as 'commercial' or 'not commercial'. Text: <TWEET-TRAIN> // <LABEL-TRAIN> ×4
  Label the next text as 'commercial' or 'not commercial'. Text: <TWEET> //

- MICD (IDEFICS)

  User: <IMAGE-TRAIN> Is the image 'commercial' or 'not commercial'? Assistant:<LABEL-TRAIN> ×4
  User: <IMAGE-TEST> Is the image 'commercial' or 'not commercial'? Assistant:

- MICD (InstructBLIP)
  - Prompt: *Is the image 'commercial' or 'not commercial'?*
  - Image: *<IMAGE-TEST>*

<Label-TRAIN> corresponds to the true label of the <TWEET-TRAIN> training example, <TWEET> refers to a testing example. We remove punctuation and spaces and map the output of each model (FLAN-T5 or GPT-3) to the corresponding label. Table 3 shows examples of outputs for each LLM model for MVSA and MSD datasets.

B.1 Implementation Details

FLAN-T5 & IDEFICS We use one GPU T4 to obtain the inference results from Flan-T5 (Chung et al., 2022) and IDEFICS (Laurençon et al., 2023) models. For Flan-T5 we use the large version from the Hugging Face library (780M parameters) (Wolf et al., 2019). For IDEFICS, we use the 9B parameters instruct version of the model (*idefics-9b-instruct*) via Hugging Face library.

InstructBLIP We use one A100 GPU to obtain inference results from InstructBLIP (Dai et al., 2023). We use the 7B-parameters version (*instructblip-vicuna-7b*) from the Hugging Face library.

GPT-3 For GPT-3 (Brown et al., 2020), we use the *text-davinci-003* model via the OpenAI[4] Library.

Note on GPT-4 For this work, we opted not to include GPT-4 due to (1) its nature as a black-box model accessible only through a paid API; (2) the lack of information regarding the pre-training data, raising concerns about potential exposure to the test sets and thus, information leakage.

[4]https://platform.openai.com/docs/api-reference