RethinkingTMSC: An Empirical Study for Target-Oriented Multimodal Sentiment Classification

Anonymous ACL submission

Abstract

Recently, Target-oriented Multimodal Sentiment Classification (TMSC) has gained significant attention among scholars. However, current multimodal models have reached a performance bottleneck. To investigate the causes of this problem, we perform extensive empirical evaluation and in-depth analysis of the datasets to answer the following questions: **Q1**: Are the modalities equally important for TMSC? **Q2**: Which multimodal fusion modules are more effective? **Q3**: Do existing datasets adequately support the research? Our experiments and analysis reveal that the current TMSC systems primarily rely on the textual modality, as most of targets' sentiments can be determined *solely* by text. Consequently, we point out several directions to work on for the TMSC task in terms of model design and dataset construction.

1 Introduction

Target-oriented sentiment classification, also known as aspect-based sentiment classification, is a fundamental task of sentiment analysis (Pontiki et al., 2014, 2015, 2016). It aims to judge the sentimental polarity (positive, negative, or neutral) of a specific target within text. To improve the performance by considering multimodal information, Target-oriented Multimodal Sentiment Classification (TMSC) is proposed to integrate both visual and textual information (Yu and Jiang, 2019).

Recently, the performance of the TMSC systems gradually reaches a plateau and the progress in tackling this task has slowed down. Using the F1-score metric on the popular datasets, Twitter15 and Twitter17 (Yu and Jiang, 2019), we observe that state-of-the-art baselines only achieve an F1-score of around 70. Therefore, in this paper, we aim to analyze the causes behind it at both model level and modality level. Roughly speaking, the modules in the model structures can be categorized into two types: 1) encoders to model the representations of different modalities; and 2) multimodal fusion modules to model the interactions between modalities. Moreover, we give a deep analysis of the characteristics of two widely-used datasets, aiming to answer the following three questions:

Q1: Are the modalities equally important for TMSC? To explore this issue, we compare and analyze the performance of unimodal models on this task. For the textual modality, we use BERT (Devlin et al., 2019) as the backbone, as it is a widely-used pre-trained language model outperforming earlier models like LSTM (Hochreiter and Schmidhuber, 1997), memory network (Weston et al., 2015), etc. For the visual modality, ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2021), and Faster R-CNN (Ren et al., 2015) are adopted (see Figure 1).

Q2: Which multimodal fusion modules are more effective? The current models use various fusion strategies to model the interactions between modalities, while obtaining little improvement. To explore the effectiveness of different fusion approaches, we summarize the fusion strategies into six categories: Concatenation, Tensor Fusion (Zadeh et al., 2017), Self Attention, Image2Text, Text2Image and Bi-direction. Then we perform a comparative study of them using a unified setup to eliminate potential bias from model size and structure (see Figure 2).

Q3: Do existing datasets adequately support the research? We analyze the existing datasets (i.e., Twitter15 and Twitter17) in depth and obtain the following findings: 1) The size of existing datasets is limited; 2) The multimodal sentiment is much more consistent with the textual sentiment than the visual sentiment; 3) A large number of targets do not exist in images; 4) There are only a small number of samples where the sentiment is decided by both text and image.

The main contributions of this work are as follows: 1) We investigate the effectiveness of different model structures for TMSC, including various

Model	Image Encoder			Fusion Module				
	ResNet	ViT	Faster R-CNN	Concat	Tensor Fusion	Self Attention	Image2Text	Text2Image
Res-BERT+BL	✓	✗	✗	✓	✗	✓	✗	✗
Res-BERT+BL-TFN	✓	✗	✗	✗	✓	✓	✗	✗
mBERT	✓	✗	✗	✓	✗	✓	✗	✗
TomBERT	✓	✗	✗	✓	✗	✓	✓	✗
EF-CapTrBERT	✓	✗	✗	✗	✗	✓	✗	✗
SMP	✗	✓	✗	✗	✗	✗	✓	✓
VLP	✗	✗	✓	✗	✗	✓	✗	✗

Table 1: The model structures of various baselines. All text encoders in the above models except for VLP are initialized with BERT.

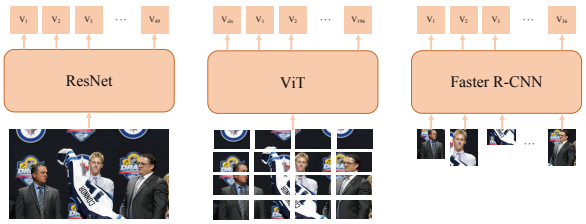


Figure 1: Different image encoders.

unimodal encoders and multimodal fusion modules; 2) We give an in-depth analysis of limitations of existing widely-used datasets; 3) We derive several valuable observations and point out promising directions for the future research of TMSC model design and dataset creation.

2 Empirical Study

We summarize the model structures and performance of the baselines for the TMSC task in Table 1. Their structural differences are mainly reflected in the different unimodal encoders and multimodal fusion modules used. Therefore, we carry out several experiments to analyze the impact of these two aspects on performance.

2.1 Unimodal Encoder

As previously mentioned in Section 1, we primarily focus on exploring the different image encoders, ResNet, ViT, and Faster R-CNN (see Figure 1), while using BERT as the text encoder.

ResNet. Following most of the baselines (e.g., mBERT (Yu and Jiang, 2019), TomBERT (Yu and Jiang, 2019) and EF-CapTrBERT (Khan and Fu, 2021)), we adopt ResNet-152 as one of the image encoders. Each image is resized into 224 by 224, and then passed through the model to obtain 49 regions, which are used as the image representation $I = [v_1, v_2, ..., v_{49}]$, where $v_i \in \mathbb{R}^{2048}$.

ViT. Following SMP (Ye et al., 2022), we adopt ViT to model the image by dividing it into 16 by 16 patches. A CLS token is added at the beginning and fed into the Transformer (Vaswani et al., 2017) encoder to obtain the image representation $I = [v_{cls}, v_1, v_2, ..., v_{196}]$, where $v_i \in \mathbb{R}^{768}$.

Faster R-CNN. Similar to VLP (Ling et al., 2022), we adopt Faster R-CNN that is retrained on the Visual Genome dataset (Krishna et al., 2017). We select the top 36 object proposals as the image representation $I = [v_1, v_2, ..., v_{36}]$, where $v_i \in \mathbb{R}^{2048}$ is obtained from the ROI pooling layer of the Region Proposal Network (Ren et al., 2015).

2.2 Multimodal Fusion

We categorize the current multimodal fusion modules into six groups as follows (see Figure 2).

Concatenate is the simplest form of fusion, where the pooled text representation $H_p^T \in \mathbb{R}^{768}$ is directly combined with the pooled image representation $H_p^I \in \mathbb{R}^{768}$ [1] to obtain the multimodal representation $H = H_p^I \bigoplus H_p^T$, where $\bigoplus$ is a concatenation operation and $H \in \mathbb{R}^{768+768}$.

Tensor Fusion is proposed for modeling interactions between modalities while preserving the characteristics of individual modalities. We obtain $H = H_p^I \bigotimes H_p^T$, where $\bigotimes$ is an outer product operation and $H \in \mathbb{R}^{768\times768}$.

Self Attention concatenates the image representation $H^I \in \mathbb{R}^{l_I\times768}$ and the text representation $H^T \in \mathbb{R}^{l_T\times768}$, where l_I and l_T are the lengths of image and text, respectively. Then it is passed through three self-attention layers and a pooling layer to obtain $H \in \mathbb{R}^{768}$.

Image2Text is one type of cross-attention mechanism, using H^I as the query and H^T as the key and value, through three attention layers to get $H \in \mathbb{R}^{768}$. **Text2Image** uses H^T as the query and H^I as the key and value instead. Furthermore, we concatenate these two as **Bi-direction** representation $H \in \mathbb{R}^{768+768}$.

[1] A linear mapping layer is added after the image encoder to map the image representation to 768 dimensions to ensure uniformity when using different image encoders.

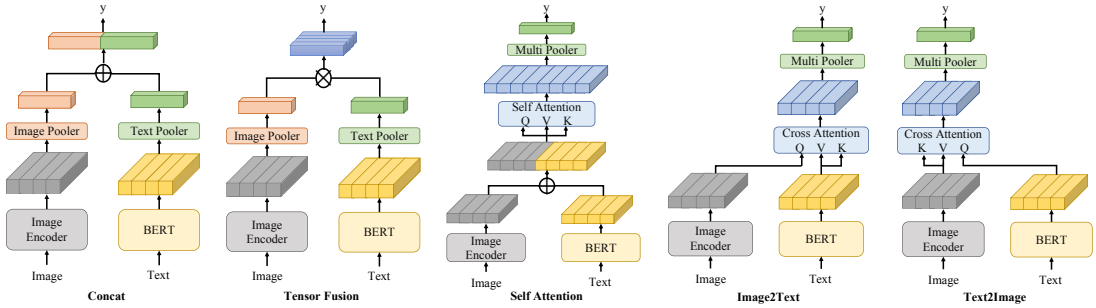


Figure 2: Various multimodal fusion modules.

Modality	Model	Twitter15 ACC	Twitter15 F1	Twitter17 ACC	Twitter17 F1
Text	BERT	76.72±1.16	71.19±2.19	68.04±0.40	65.66±0.35
Image	ResNet	57.65±1.00	32.52±2.66	57.79±0.99	51.98±1.23
	ViT	59.65±1.13	31.25±2.71	59.53±0.95	54.08±0.78
	Faster R-CNN	55.97±1.10	35.72±5.43	56.18±0.85	49.88±1.70
Multimodal	ResNet				
	Concatenate	75.29±0.45	68.71±1.34	67.92±0.56	65.32±0.53
	Tensor Fusion	74.19±0.94	68.93±0.57	66.66±1.21	63.99±1.61
	Self Attention	76.03±0.96	70.57±2.39	68.01±0.96	65.41±1.60
	Image2Text	77.13±1.33	71.48±1.90	69.37±0.36	66.85±0.79
	Text2Image	75.18±1.66	67.77±4.81	68.07±0.58	65.18±1.48
	Bi-direction	77.32±0.63	**72.06±0.81**	68.41±1.01	66.39±1.39
	ViT				
	Concatenate	76.22±0.90	70.37±1.45	67.94±0.70	66.17±0.78
	Tensor Fusion	73.44±0.78	67.46±1.45	65.46±1.67	62.02±1.40
	Self Attention	75.08±0.41	68.94±0.83	67.52±0.58	65.56±0.35
	Image2Text	77.11±0.44	71.91±0.42	69.14±0.52	66.96±0.68
	Text2Image	75.12±1.01	69.40±1.38	67.52±1.06	64.49±1.46
	Bi-direction	76.70±0.75	71.67±1.45	69.16±0.17	67.25±0.56
	Faster R-CNN				
	Concatenate	75.45±0.73	69.77±1.23	67.60±1.15	64.74±1.69
	Tensor Fusion	72.09±0.66	66.77±1.04	66.34±1.45	62.96±2.09
	Self Attention	76.09±0.89	70.08±1.37	68.09±1.10	66.12±1.23
	Image2Text	**77.36±0.37**	71.69±0.37	68.43±0.65	66.44±1.10
	Text2Image	70.82±2.99	57.94±5.81	60.31±6.43	54.50±7.06
	Bi-direction	76.57±0.46	70.88±0.89	**69.51±0.62**	**67.50±0.37**

Table 2: Results on Twitter15 and Twitter17. The overall best results and those within each corresponding block are marked with **bold** and underline, respectively.

2.3 Results Analysis

We perform experiments of different unimodal encoders and fusion modules over Twitter15 and Twitter17. In Table 2, we show the results and we have the following observations [2]:

First, the text-only model (i.e., BERT) performs well, while the visual-only models (i.e., ResNet, ViT, and Faster R-CNN) perform relatively poorly, revealing that the reliance on text is much greater than that on images for the TMSC task on these two datasets. In comparison, this phenomenon is more pronounced in Twitter15.

Second, the performance of the model is affected by the use of different fusion methods. Specifically, fusion modules that primarily focus on acquiring the textual information (e.g., Image2Text) perform better than those focused on acquiring the visual information (e.g., Text2Image). This again reveals the inconsistent importance of text and images.

Third, compared with the text-only model, the various fusion modules do not have significant gains in performance and some are even worse. This is due to the fact that some images do not provide related information, but rather distracting information instead [3].

Fourth, the impact of various image encoders is not clear, as evidenced by low performance and high standard deviation on the two datasets (see the "Image" part of Table 2). Moreover, differences in performance among various image encoders are small in the multimodal fusion settings (see the "Multimodal" part of Table 2). This is due to the characteristics of visual data in existing datasets, which is analyzed in depth in the following section.

Based on the comprehensive experimental analyses conducted above, we identify several key points to be considered when designing models for the TMSC task in the future: 1) leveraging text information to exploit the advantages of textual data fully; 2) devising more effective image encoding methods to extract semantic information from images better; and 3) enhancing the noise immunity of the fusion module to enable more flexible selection and utilization of informative features from both textual and visual modalities.

3 Data Analysis

To gain a deeper understanding of the performance issues mentioned above, we conduct detailed analyses of the two datasets, taking into account *quantity*, *diversity*, and *annotation*. Following the annotation procedure employed by Yu and Jiang (2019), we enlist the participation of three domain experts to annotate 400 randomly sampled test data (200 from

[2]The experimental setup is illustrated in Appendix B.

[3]We give a detailed analysis of the performance comparison for the multimodal model versus the text-only model in Appendix C.

Dataset	Twitter15					Twitter17				
	#Negative	#Neutral	#Positive	#Total	#Avg Targets	#Negative	#Neutral	#Positive	#Total	#Avg Targets
Train	368	1883	928	3179	1.348	416	1638	1508	3562	1.410
Dev	149	670	303	1122	1.336	144	517	515	1176	1.439
Test	113	607	317	1037	1.354	168	573	493	1234	1.450

Table 3: Statistics of the datasets. #Avg Targets means the average number of targets for each sample.

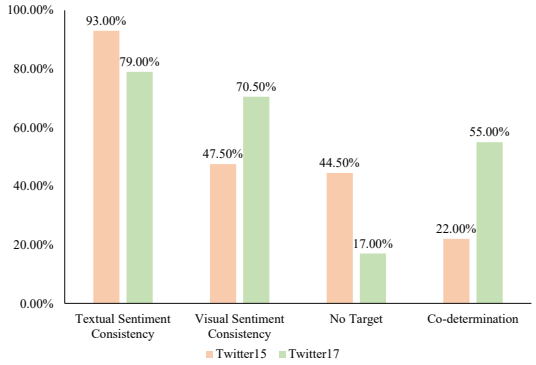


Figure 3: Annotation analysis. Textual/Visual Sentiment Consistency: the consistency of the target's sentiment in text/image with the sentiment in multimodal information. No Target: the percentage of images that are missing the target for sentiment analysis. Co-determination: the percentage of targets that sentiment is jointly determined by text and image.

Twitter15 and 200 from Twitter17) across four aspects, with the majority vote being considered as the final annotation result (Figure 3) [4]. We have the following observations:

First, as shown in Table 3, the sample size is relatively small, with an average of less than 1.5 targets per sample. Additionally, the distributions of the sentimental labels are unbalanced in both datasets, with neutral sentiment accounting for approximately 50% and negative sentiment accounting for less than 15%. The reason behind this is that Twitter15 and Twitter17 were originally constructed by Zhang et al. (2018) and Lu et al. (2018) respectively for the named entity recognition task, rather than specifically for TMSC.

Second, the multimodal sentiment has high consistency with the textual sentiment but low consistency with the visual sentiment. In Twitter15, 93% of the targets have the same textual sentiment as the multimodal sentiment, while only 47.5% have a visual sentiment that matches. This indicates the biased distribution existing in the dataset, i.e., the textual information is more useful for determining the multimodal sentiment. Although this phenomenon is mitigated in Twitter17, the textual information is still more consistent with the multimodal sentiment than the visual information.

Third, a large number of targets do not exist in images, which is also not suitable for the *target-oriented* multimodal sentiment classification task. This phenomenon may stem from the construction of the two datasets, where the targets are selected directly from the text, without taking into account the corresponding images (Yu and Jiang, 2019).

Fourth, due to the facts of irrelevant images and non-existence of targets in images, there is only a small portion of the data where the sentiment is determined by both text and images. Specifically, only 22% of Twitter15 and 55% of Twitter17 data require both text and images for the sentiment classification. As for the multimodal task, these two datasets may not be the best-suited in this aspect.

Based on our analyses of existing datasets, we propose that high-quality TMSC datasets should possess the following characteristics: 1) accurately reflecting the real-world data distribution, including factors such as unbalanced label distribution, while also providing sufficient data samples for different cases; 2) large data diversity, i.e., various data types and domains, to facilitate valid testing for models' generalization capability and robustness; and 3) multi-dimensional annotation information, including both multimodal and unimodal sentiment, to enable thorough analysis of the model's ability to handle different data sources.

4 Conclusion and Future Work

In this paper, we conduct a series of in-depth experiments for TMSC and data analysis of existing datasets. Our findings reveal that current multimodal models do not exhibit significant performance gains compared to text-only models on the TMSC task. This is largely attributed to the over-reliance on textual modality in existing datasets, while visual modality playing a comparatively less significant role. Based on our experimental analyses, we propose future directions for designing models for the TMSC task and for constructing more suitable datasets which better capture the multimodal nature of social media sentiments.

[4] Illustrative examples with annotations are in Appendix D.

Limitations

Although we have conducted a series of experiments and data analysis for the TMSC task to the best of our ability, there are at least the following limitations to our work. First, our data analysis was performed mainly for the currently publicly available English datasets Twitter15 and Twitter17, neglecting the Chinese dataset Multi-ZOL, which has not been widely studied. Second, although our analysis indicated some problems in using the currently dataset to measure the TMSC task, we did not construct a new and better dataset for use in academic studies. We have included this task as one of our future works to be investigated. Third, in our experiments, we did not specifically compare the impact of different text encoding methods on the model performance. While we acknowledge that different text encoding methods may indeed have an impact, it is worth noting that BERT, being a well-established text encoding method, already performs adequately. And most existing models use BERT as the text encoder. Therefore, we focused our study on investigating image encoding methods and fusion modules, as we believe there is more room for improvement in these parts.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Zaid Khan and Yun Fu. 2021. Exploiting BERT for multimodal target sentiment classification through input space translation. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 3034–3042. ACM.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2149–2159. Association for Computational Linguistics.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1990–1999. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 19–30. The Association for Computer Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 486–495. The Association for Computer Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh

Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 371–378. AAAI Press.

Junjie Ye, Jie Zhou, Junfeng Tian, Rui Wang, Jingyi Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Sentiment-aware multimodal pre-training for multimodal sentiment analysis. *Knowl. Based Syst.*, 258:110021.

Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5408–5414. ijcai.org.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1103–1114. Association for Computational Linguistics.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5674–5681. AAAI Press.

A Related Work

As one of the tasks of sentiment analysis, TMSC has gained great attention from scholars in recent years (Yu and Jiang, 2019). Xu et al. (2019) constructed a Chinese dataset named Multi-ZOL and proposed a multi-hop memory network for handling modal interactions. Subsequently, Yu and Jiang (2019) constructed two English datasets, Twitter15 and Twitter17, and applied BERT to this task. The following research on the TMSC task can be divided into two directions. On the one hand, there is the continuous exploration of how to enhance the interactions between modalities (Khan and Fu, 2021), and on the other hand, there is the application of pre-trained models to this task (Ye et al., 2022; Ling et al., 2022). Despite these efforts, the current models have not yet achieved significant performance gains relative to the text-only models. We have conducted a series of experiments and data analysis, hoping to provide some insights for the future research of TMSC.

B Experimental Setup

For each set of experiments, we conduct tests using five different random seeds (i.e., 0, 42, 199, 2022, and 11122). We initialize the parameters of the BERT text encoder with bert-base-uncased. The image encoder parameters are frozen, and we use resnet-152, vit-base, and faster-rcnn retrained on the Visual Genome dataset as image encoders, respectively. For both the self-attention and cross-attention modules, we use the last three initialization parameters of bert-base-uncased. We utilize the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 2e-5 and run each experiment on a 3090 GPU for 8 epochs. We select the best epoch's parameters on the validation set for testing and calculate the mean and standard deviation as the final result.

To ensure a fair comparison, we set the models in Table 2 uniformly and without continuing pre-training, which may make it challenging to compare them with existing papers due to differences in overall structure and training details compared to Table 1, even if they may use the same unimodal encoder and multimodal fusion modules.

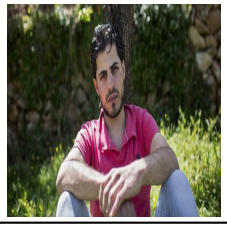
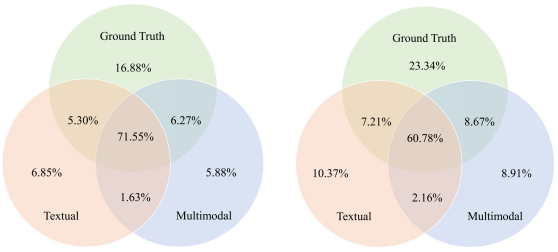
Text	Image	Target	Sentiment		
			Multimodal	Textual	Visual
Congratulations to our second draw winner - **Bulaire Leber** of ADSS Global , Haiti . Thanks for participating , Bulaire		Bulaire Leber	Positive	Positive ✓	Positive ✓
RT @ BeschlossDC : Coretta Scott King with Robert amp **Ethel Kennedy** after husband ' s assassination , which occurred tonight 1968		Ethel Kennedy	Negative	Negative ✓	Neutral ✗
Pres **Obama** takes the stage at @ RutgersU Commencement in school football stadium in Piscataway , NJ .		Obama	Positive	Neutral ✗	Positive ✓
RT @ Refugees : Today , 18 - year - old Yehya became the 1 millionth Syrian to register as a refugee in **Lebanon**		Lebanon	Neutral	Neutral ✓	No Target ✗

Table 4: The annotation examples.



(a) Twitter15 (b) Twitter17

Figure 4: Venn diagram for model performance visualization.

C Model Performance Visualization

We select Image2Text (Faster R-CNN) as the representative of the multimodal models and compare its performance with that of BERT with a random seed of 11122 to obtain Figure 4. The intersection of every two circles in the figure represents the part where the prediction results are consistent. Based on this comparison, we have the following observations: **First**, in terms of prediction accuracy, the multimodal model does not achieve a significant gain over the text-only model. **Second**, a portion of the data is predicted correctly by the multimodal model but incorrectly by the text-only model, and vice versa. The proportions of these two parts are similar. This suggests that when images do contribute valuable information to the multimodal model, they also introduce noise. In order to improve the performance, further investigation is required for how to properly incorporate the visual information. **Third**, over 16% of the data has sentiments that neither the text-only model nor the multimodal model predicts correctly. This indicates the weakness of the current models and we need further explorations.

D Annotation Examples

To clearly and visually illustrate the various scenarios that arise during the dataset annotation process, four samples are presented in Table 4.

The **first** example demonstrates a scenario where the textual sentiment and the visual sentiment matches, resulting in a multimodal sentiment determined by both modalities. In the example, the sentiment in the text is determined to be positive through the use of words such as "Congratulations" and "winner". Similarly, the sentiment in the image can be inferred as positive by identifying the target (i.e., the first person on the right) and noticing his smiling face.

The **second** example shows a scenario where the textual sentiment aligns with the multimodal sentiment but not with the visual sentiment, leading to a multimodal sentiment determined by the textual modality only. Specifically, the sentiment conveyed by the text is negative due to the phrase "after husband's assassination" and the sentiment conveyed by the image is neutral as it does not show an obvious facial expression on the person referred to in the text (i.e., the first person on the left). Therefore, the multimodal sentiment conveyed by both modalities is negative.

Corresponding to the second example, the **third** example illustrates a scenario where the visual sentiment aligns with the multimodal sentiment but not with the textual sentiment. In particular, the text simply states a fact with a neutral sentiment, while the image shows the target (i.e., the person waving his hand in front of the podium) with a positive facial expression and posture, resulting in a positive multimodal sentiment overall.

The **fourth** example presents a scenario where there is no target in the image, resulting in a multimodal sentiment determined solely by the textual modality. Here, the target is "Lebanon", but since there is only one person in the image and no information about "Lebanon", we can only conclude that the multimodal sentiment is neutral based on the text. It is worth mentioning that such a sample is not ideal for the TMSC task as the image does not convey any sentimental information towards the target.