The Three Musketeers: Using Audio, Video, and Text for multimodal Classification

Anonymous ACL submission

Abstract

Human communication relies on signals across multiple modalities, yet natural language processing research has predominantly focused on text and, to a lesser extent, images. This work investigates a new multimodal classification sampling method and feature extraction process. While previous research has attempted to focus on fusion, we introduce a new sampling strategy for multimodal classification which takes advantage of the conversational structure of dialogue. We compare our method with standard samplers and feature extractors for state-of-the-art encoder models to detect emotion, sentiment, and sarcasm in English datasets. Using our our method, we obtain state-of-the-art results for the uni- and trimodal settings for two of our tasks. Our results indicate that encoding dialogue structure into sampling can provide performance benefits for multimodal modelling dialogues.[1]

1 Introduction

The internet has revolutionised global communication, affording easy and instantaneous communication across the globe. The primary format of online communication has traditionally been text. Despite the prevalence of text, it overlooks critical aspects of human communication such as tone, pitch, and facial expressions (Song et al., 2022; Liu et al., 2023; Lee and Lee, 2021). To compensate such absences, symbols like emojis and memes have been adopted to afford expressing intent and emotions (Seargeant, 2019).

In efforts to address the lack of physical and audio signals, prior work has sought to include additional information in the text modality. For instance, Felbo et al. (2017) have presented methods for meaningful representation of emojis, and a line of research has sought to combine image and text for modelling language, e.g., for hate speech

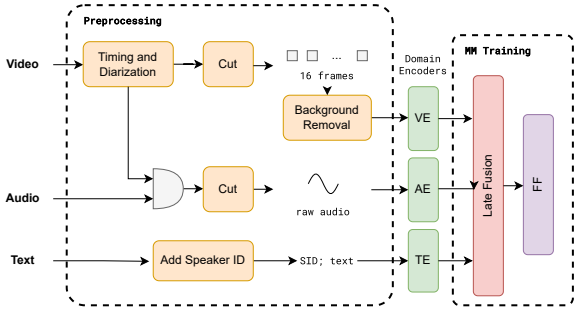

Figure 1: Method Overview: Unimodal preprocessing removes irrelevant video and audio segments, and augments text with speaker IDs. Domain-specific encoders (green) extract contextual features, which are fused (red) and projected (purple) to classification labels.

detection (Mathias et al., 2021; Attanasio et al., 2022; Kiela et al., 2020, *inter alia*). Contemporary multimodal approaches are typically characterized by different fusion methods pertaining to combining the different modalities or their changes in feature selection by their difference in noise removal within the audio and video modalities.

This paper proposes an alternative to multimodal classification methods through precise speaker diarization, noise removal in videos, and dialogue-aware data sampling. These techniques allow for extracting relevant audio segments (e.g., those containing actual dialogue), and, within those, remove non-relevant content (e.g., backgrounds in video frames – see Figure 2). Moreover, custom sampling handles class imbalance across dialogues, a long-standing challenge in multimodal dialogue learning. In our experiments on sarcasm detection, emotion detection, and sentiment analysis in English, our unique sampling strategy outperforms prior work by 4.32% on average for all datasets and modalities/methods. In addition, we introduce a novel feature extraction method for audio and video, improving the performance on average by 5.65% for unimodal models.

As a result, our fusion-based multimodal model takes advantage of better data representations, out-

[1] Research artefacts will be made available on acceptance.

performing the state-of-the-art models by 0.44% on our triple modalities.

Thus, our contributions include:

- A novel preprocessing method to extract the precise timing of dialogue in audio and video signals, thus removing irrelevant data signals (i.e., audio and video without any dialogue turns);
- A unique sampling strategy which combines an iterative sampler with gradient accumulation and weighted sampler, which affords exposing the model to utterances within the context of the dialogue and in isolation from the context; and
- we examine the application of our sampling and feature extraction strategy to two trimodal fusion methods.

2 Previous Work

Emotion recognition, sarcasm detection and sentiment analysis in conversations are complex and subjective tasks that require understanding the context of the conversation, including previous and future interactions; social cues, such as body language and facial expressions; and cultural values of how emotions are expressed (Stark, 2018). In recent years, researchers have explored various approaches to this problem, including the use of text, audio, and video data (e.g., Chudasama et al., 2022; Xie et al., 2021; Mao et al., 2020). While current state-of-the-art models often rely solely on text data (e.g., Song et al., 2022), we investigate methods for multimodal modelling to incorporate the full breadth of available data.

Emotion For emotion detection, Chudasama et al. (2022) propose a multimodal approach for emotion detection using audio, video, and text modalities. Their model uses a novel feature extractor with an adaptive margin-based triplet loss function to incorporate audio and video data with text data. Their method also makes use of a face-and-scene embedding to capture the scene's context and emotions expressed through facial expressions. These representations of features and embeddings are used in a multi-headed attention-based fusion mechanism to combine the latent representations of the three data modalities, obtaining state-of-the-art performances on their classification datasets.

Sentiment Sentiment analysis is a long-standing task in natural language processing (Kenyon-Dean et al., 2018) where the information provided in text has faced particularly difficult challenges due to sarcasm and irony. Recent work has attempted to attempted to use multiple modalities for sentiment analysis (e.g., Sun et al., 2019). For instance, Sun et al. (2019) propose a model that uses deep canonical correlation analysis to learn correlations between text, audio, and video features. the authors evaluate their approach on three datasets: CMU-MOSI and CMU-MOSEI (Zadeh et al., 2018), and IEMOCAP (Busso et al., 2008).

Sarcasm Ray et al. (2022) proposed a collaborative gating-based multimodal fusion model for detecting sarcasm. The model incorporates text, audio, and video modalities and is speaker-independent. The study uses BART for text, low-level audio descriptors, and ResNET video features as the best modality features. The authors also apply key frame extraction to reduce noise and improve video representation. The proposed approach outperforms state-of-the-art models and provides a benchmark for sarcasm detection.

3 Datasets

This study focuses on multimodal learning across sentiment, emotion, and sarcasm detection across three state-of-the-art dialog-based datasets. We report all dataset statistics in Appendix A.

MELD The Multimodal EmotionLines Dataset (MELD, Poria et al., 2018) is a multimodal dataset that contains video clips with audio and subtitles from the *Friends* TV series. The dataset contains over $13,000$ utterances distributed over $1,400$ conversations (see Figure 3 for an example dialogue), split into train, validation, and test sets. Each utterance has one of three sentiment labels: positive, negative, and neutral. We experiment with sentiment analysis to show the utility of our method across tasks and datasets, as we use IEMOCAP for emotion detection.

IEMOCAP The IEMOCAP dataset (Busso et al., 2008) is a comprehensive, multi-speaker, and multimodal dataset that includes video recordings of two-person interactions. This dataset spans 12 hours and is divided into individual conversations, which is further segmented into several utterances. Each utterance is labelled with one of six emotions: happiness, sadness, neutrality, anger, excitement, and frustration. The sample splits in this study follow those in the EmoBERTa (Kim and Vossen, 2021).

MUSTARD++ The MUStARD++ database (Ray et al., 2022) is an extensive, multimodal dataset that encompasses multiple speakers. Like MELD, it comprises video segments from television programs and YouTube content, specifically identifying sarcasm in interactions. The dataset includes approximately 1.5 hours of audio and video recordings. The annotations in this dataset are binary, categorising interactions as sarcastic or non-sarcastic.

This dataset comes with both context and target sarcasm utterances. Thus, we separately apply the methodology below to the context and target, then take a weighted average between them. We set the weights to be 75% of target and 25% for context.

4 Methodology

Our approach leverages curated preprocessing (§4.1), domain-adapted encoders, and ad-hoc sampling techniques (§4.2). While preprocessing retains only relevant segments and content of each dialog, custom encoders and fine-tuning strategies help solving the classification task effectively. See Figure 2 for a graphical overview.

4.1 Data Preprocessing

Given the large amount of information available for our models, we emphasise developing features and preprocessing steps that afford more robust models and minimising the amount of data the model is exposed to. Specifically, we follow Kim and Vossen (2021) in prepending the speaker names and the previous utterances to the text modality onto MELD and IEMOCAP. For MUStARD++, we prepend previous utterances but not the speaker, because some utterances were labelled with the speaker "Other."

For the audio and video modalities, we take additional steps. We first extract word timestamps in an utterance using Wav2Vec2 (Baevski et al., 2020) and filter out any audio segment that does not contain any person speaking, to focus on purely the spoken utterance. For video, we extract sixteen time equidistant frames from our filtered input and use Yolov4 (Bochkovskiy et al., 2020) to identify bounding boxes over humans in the frames. We use the positions of the bounding boxes to black out the remainder of the images.

4.2 Multimodal Training

Our approach to emotion detection, sarcasm detection and sentiment analysis is a two-step process. In the first step, we adapt modality-specific encoders to recognise task-specific cues. In the second, we fine-tune the entire architecture end-to-end (see Appendix B for full details).

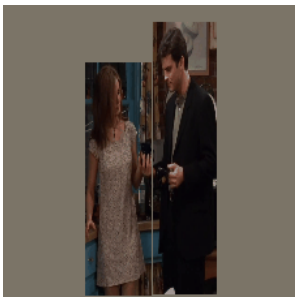

Figure 2: Remove all non-humans within video frames.

Encoder Adaptation Our model used base checkpoints for RoBERTa (Liu et al., 2019) for text, Whisper (Radford et al., 2022) for audio, and VideoMae (Tong et al., 2022) for video modalities. We run domain-adaptation of each encoder separately for each dataset and modality, following standard fine-tuning procedures. This ensures that extracted features are attuned to the dataset.

Fine-tuning Next, we perform end-to-end fine-tuning of the full architecture for each experimental setup. First, we extract 512 feature embeddings for each modality. While Whisper and VideoMAE can produce longer sequences, using the same length for each modality ensures that all modalities are equally contributing information. For Whisper, a sequence length of 512 tokens corresponds to 10.24 seconds of audio, due to its sampling rate of 20 milliseconds.[2] VideoMAE requires a high masking rate to learn effectively. We therefore mask approximately 66% of its features and retrieve sixteen frames evenly-spaced across time.[3]

Due to class imbalances in the datasets (see Table 4), we initially experimented with a weighted sampler with standard cross-entropy loss, and random sampler with weighted cross-entropy loss. In both cases, though, models prediction were skewed toward majority classes. We therefore developed our novel two-staged sampler. For the first N training epoch, we use a *weighted sampler* to improve the model's ability to learn the minority classes.

[2]For MELD and IEMOCAP, only 2% of the instances exceed this length, while >37% exceeded it for MUStARD++. Of the latter, 94% belong to context – i.e., preceding the target utterance – we take the last 10.24 seconds as our features.

[3]Despite this rate of masking, the extracted features remain relevant as the feature extractor is only provided pixels corresponding to a body or face, ensuring reliability of the features during the fusion.

Modality	Combining Samplers			One Sampler			ΔModality
	MELD	IEMOCAP	MUStARD++	MELD	IEMOCAP	MUStARD++	–
Text	72.24 (σ0.99)	60.58 (σ1.81)	63.36 (σ2.05)	72.49 (σ0.37)	56.6 (σ4.23)	61.45 (σ1.02)	1.88
Audio	59.06 (σ0.89)	55.4 (σ0.81)	68.49 (σ2.02)	46.23 (σ0.8)	28.73 (σ1.41)	66.89 (σ2.46)	13.7
Video	34.69 (σ2.34)	26.18 (σ1.85)	58.12 (σ2.05)	38.94 (σ0.73)	23.34 (σ0.4)	59.02 (σ1.36)	-0.77
T+A+V	72.41 (σ0.36)	64.48 (σ0.37)	72.62 (σ2.52)	72.04 (σ0.7)	59.19 (σ1.48)	70.82 (σ2.22)	2.49
ΔDataset	2.18	9.70	1.10				
Triple vs. Single	0.17	3.9	4.13	-0.45	2.59	3.93	

Table 1: The Δs report differences from combining samplers. *One Sampler* reports the best among Weighted and Iterative samplers. The difference using three modalities over the best unimodal model is given in *Triple vs. Single*.

Models	MELD (Sentiment)				IEMOCAP (Emotion)				MUStARD++ (Sarcasm)			
	T	A	V	T+A+V	T	A	V	T+A+V	T	A	V	T+A+V
Poria et al. (2018)	66.10	49.61	–	67.56	–	–	–	–	–	–	–	–
Chudasama et al. (2022)	–	–	–	–	66.20	21.79	13.10	69.86	–	–	–	–
Li et al. (2023)	–	–	–	–	**66.56**	53.45	**45.01**	**70.78**	–	–	–	–
Ray et al. (2022)	–	–	–	–	–	–	–	–	**69.2**	64.1	**60.0**	70.2
Ours	**72.24**	**59.06**	**34.69**	**72.41**	60.58	**55.4**	26.18	64.48	63.36	**68.49**	58.12	**72.62**

Table 2: Comparison of our best model using the combined sampling strategy with prior work.

Then, we switch to an *iterative sampler* with a weighted cross-entropy loss, where the weights are the reciprocal of the square root of the size of each class. Every N epochs, we switch the sampler. In the iterative sampler, we apply gradient accumulation to the length of each dialogue, performing the gradient descent step after the model has been exposed to the entire dialogue. This gives the model more situational context within the dialogue. Overall, our two-staged sampling and gradient accumulation has two main benefits: It allows the model to balance learning across majority and minority classes, and exposes the model to the conversational context, benefitting minority classes.

Finally, fusion solutions afford merging information from modality-specific features. We use two late fusion architectures. In the first, we reimplement M2FNet (Chudasama et al., 2022), which runs encoder layers over each modality and fuses them with their multiheaded attention. In the second, we concatenate over the hidden dimension.

5 Results

Following prior work (Poria et al., 2018; Busso et al., 2008; Ray et al., 2022), we report the class-weighted F1 score for each classification task.

Table 1 reports our main results averaged over five random initialisation seeds. We find that alternating between weighted and iterative is better than sticking with one for all modalities and datasets, expect for video. We believe that the performance drop for video is due to the complexity of the modality.

In comparing our method to prior work (see Table 2), our models outperform the state-of-the-art models in 7 of 12 cases. In particular, the performance for the audio outperforms prior work due to the use of stronger pretrained Whisper and our noise-removal preprocessing based on time-stamps identification. Our text modality, however, underperforms existing work. We believe that this is caused by only pre-appending one previous utterance to the target utterance. As a result, our fusion strategy underperforms, in particular for IEMOCAP, where the performances in prior work has heavily relied on high performance text models.

6 Conclusion

This paper presented a multimodal approach to emotion detection, sarcasm detection, and sentiment analysis in conversations that integrates text, audio, and video data. Our approach introduced a novel feature selection and training regime, resulting in new state-of-the-art performances on the MUStARD++ and MELD datasets.

Our results demonstrate the potential of better feature selection and sampling toward multimodal approaches and provide a foundation for future research in this area. In future work, we aim to enhance our model's performance by improving the text modality and exploring alternative fusion strategies to better our model.

Limitations

One limitation of our study is that our video modality is currently trained without explicitly attending to facial features, which are, therefore, not significantly captured by the model in spite of the rich information that facial expressions may convey.

In our study, we use datasets consisting of young white people from the television show "Friends" which has been criticized for its lack of representation of people of color (Mitchell, 2021; Cobb, 2018), as well as "Big Bang Theory", and "Silicon Valley". Furthermore, these television shows are all situational comedies, which likely influences how emotions, sarcasm, and sentiment are conveyed towards achieving comedic effects.

Ethical Considerations

Concerns have also been raised around machine learning for emotion classification, sentiment analysis, and sarcasm detection (e.g., Podoletz, 2023; Stark, 2018). For instance, machine learning for visual emotion detection has been criticized on the one hand for issues around its use in public settings, e.g., policing and hiring, due to issues of social biases, e.g., racism (Podoletz, 2023). It has been criticized within the text domain due to the scant justification for many resources and tools for detecting emotions (Stark, 2018). In part, the display of emotions and different emotional cues, such as smiling, is grounded within the local context, e.g., it would be inappropriate to smile when someone is seriously injured but appropriate when meeting a puppy. Moreover, the effects of a facial expression, such as a smile, are culturally contingent. Therefore, the communicative meaning of any facial expression may not travel across geographic and cultural contexts.

This work is based on datasets consisting of television shows, which means that the findings will not map fully to real-life situations, and this work is therefore *exclusively* appropriate for the academic domain of developing multimodal classification techniques for audio, video, and text modalities.

References

Giuseppe Attanasio, Debora Nozza, Federico Bianchi, et al. 2022. Milanlp at semeval-2022 task 5: using perceiver io for detecting misogynous memes with text and image modalities. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4652–4661.

Shelley Cobb. 2018. "I'd Like Y'all to Get a Black Friend": The Politics of Race in *Friends*. *Television & New Media*, 19(8):708–723.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Joosung Lee and Wooin Lee. 2021. Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. *arXiv preprint arXiv:2108.11626*.

Jiang Li, Yingjian Liu, Xiaoping Wang, and Zhigang Zeng. 2023. Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition.

Xiao Liu, Jian Zhang, Heng Zhang, Fuzhao Xue, and Yang You. 2023. Hierarchical dialogue understanding with special tokens and turn-level attention. *arXiv preprint arXiv:2305.00262*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Yuzhao Mao, Qi Sun, Guang Liu, Xiaojie Wang, Weiguo Gao, Xuan Li, and Jianping Shen. 2020. Dialoguetrm: Exploring the intra-and inter-modal emotional behaviors in the conversation. *arXiv preprint arXiv:2010.07637*.

Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206, Online. Association for Computational Linguistics.

Amanda Mitchell. 2021. Julie, Charlie, and the Women of Color 'Friends' Failed. *Marie Claire Magazine*.

Lena Podoletz. 2023. We have to talk about emotional AI and crime. *AI & SOCIETY*, 38(3):1067–1082.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. 2022. A multimodal corpus for emotion recognition in sarcasm. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6992–7003.

Philip Seargeant. 2019. *The Emoji Revolution: How technology is shaping the future of communication*. Cambridge University Press.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv preprint arXiv:2210.08713*.

Luke Stark. 2018. Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, 48(2):204–231.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2019. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. *arXiv preprint arXiv:2203.12602*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Baijun Xie, Mariia Sidulova, and Chung Hyuk Park. 2021. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion. *Sensors*, 21(14):4913.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

A Dataset Statistics

In this section, we briefly report dataset statistics which we recover from the papers introducing the datasets, i.e., Poria et al. (2018); Busso et al. (2008) and Ray et al. (2022).

Table 4 outlines the the sizes of the splits reported by the authors of each dataset, by the number of dialogues and the number of utterances, when the dialogue is split into individual utterances. Figure 3 shows an example dialogue, its dialogue turns, i.e., utterances, and how labels are assigned to each utterance in Poria et al. (2018).

Finally, Table 3 details the dataset statistics for meld, outlining the number of speakers in the dataset, the number of dialogues and dialogue turns, and the average length of an utterance.

B Experimental Setup

B.1 Pretrained Encoders

We used code and pretrained encoders available on the Hugging Face Hub and `transformers`

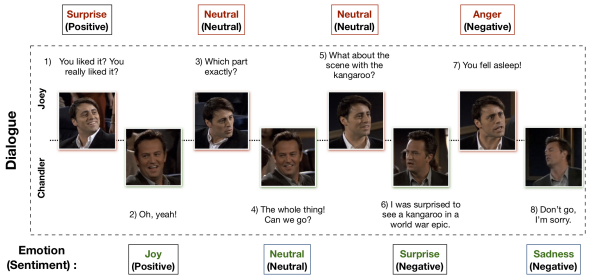

Figure 3: Example dialogue in Poria et al. (2018)

	Train	Dev	Test
# of modality	$\{a, v, t\}$	$\{a, v, t\}$	$\{a, v, t\}$
# of dialogues	1039	114	280
# of utterances	9989	1109	2610
# of speakers	260	47	100
Avg. duration of an utterance	$3.59s$	$3.59s$	$3.58s$

Table 3: Statistics for MELD (Poria et al., 2018)

(Wolf et al., 2020). In the following, we report the Hub's model IDs. We used roberta-large for text, MCG-NJU/videomae-base for video, and openai/whisper-large for audio.

B.2 Hyper-parameter Search

We conducted an extensive bayesian hyper-parameter search for both encoder adaptation and multimodal learning.

Table 5 reports the list of parameters and values we searched for unimodal models. They are shared among all models and tasks, whereas we adapt a model batch size to fit hardware constraints. Table 6 reports the same list for the multimodal training setup. For all models, we monitored the validation loss every 50 training steps and halted training at 14 consecutive non-improving losses. Then, we used the model with the best validation loss for testing.

B.3 Hardware and Computational Costs

We ran all experiments on 40GB Nvidia V100 GPUs.

Datasets	Dialogues			Utterances			classes	type	avg_utt	Evaluation
	Train	Val	Test	Train	Val	Test				
IEMOCAP	108	12	31	5163	647	1623	6	two-person	47	Weighted-F1
MELD	1038	114	280	9989	1109	2610	3	multi-party	9	Weighted-F1
MUStARD++	4264	598	1179	841	120	241	2	multi-party	5	Weighted-F1

Table 4: The statistics of each dataset. Avg_utt denotes the average number of utterances in a conversation.

Parameter	**Values**
Learning rate	1e-6, 1e-5, 1e-4, 5e-5, 5e-6
Weight decay	1e-5, 1e-4, 1e-3, 1e-2
Dropout	0.2, 0.3, 0.4, 0.5
Epoch	6, 7, 8
Epoch switch	2, 3

Table 5: Hyper-parameter search space for unimodal models.

Parameter	**Values**
Learning rate	[1e-4, 1e-6], uniform sampling
Weight decay	[1e-3, 1e-6], uniform sampling
Dropout	0.3, 0.4, 0.5
Epoch	6
Epoch switch	2, 3

Table 6: Hyper-parameter search space for the multimodal model.