11-777 MMML: Mid-Term Report
Analysis of Multimodal Baselines for Emotion Recognition in Conversations

Athiya Deviyani [*1] **Abuzar Khan** [*1] **Neil Skarphedinsson** [*1] **Prasoon Varshney** [*1]

Abstract

Multimodal Emotion Recognition in Conversation (mERC) is an important stepping stone towards intelligent systems interacting with humans where emotional awareness is crucial to the system. As a result of advancements in Multimodal Machine Learning as well as availability of datasets for the mERC task, research in the area has proliferated with novel ways for computer systems to classify emotions and sentiments based on multi-sensory input. In this work, we present results and detailed error analysis of two baseline models for mERC from literature and apply them to the Multimodal EmotionLines Dataset (MELD). The first is DialogueRNN, which was the first speaker-aware multimodal approach on the MELD dataset released as a baseline along with the dataset. Second, we extend the current state-of-the-art on the MELD dataset, Supervised Prototypical Contrastive Learning (SPCL) model, by implementing representation fusion for text and audio in order to leverage information in the acoustic modality. We name our extended version of SPCL as **M**ulti**M**odal-**SPCL** (MM-SPCL). In future, we will contextualize and incorporate the video modality too.

1. Introduction

Emotion Recognition in Conversations (ERC) is an important area of study in the pursuit of building machines that interact with humans in an empathetic and understanding manner. It is an area of research rooted in natural language processing (NLP) largely due to the ability to retrieve conversational data from social media platforms (Poria et al., 2018). As a result of this ease of access to textual corpora, most methods have taken a text-only unimodal approach.

However, in human dialogue, we decipher emotions by leveraging information beyond just text such as intonations in speech, facial expressions, gestures, and context. Intonations in voice can change the emotion conveyed by a sentence (Levis, 1999), facial expressions can convey sarcasm, and speaker context can help a listener recognize emotion based on the who the speaker is.

If we want to model things such as intonations, facial expressions, gestures, and context then we need multiple modalities beyond just text. In the last few years, multimodal ERC (mERC) has become a popular research topic with frequent state of the art advancements on popular multimodal datasets (Poria et al., 2018; Zadeh et al., 2018) for mERC.

The mERC task is quite difficult due to the multiple ways we can perform reasoning to incorporate these different types of information which are often obvious to humans. This includes (1) modelling speaker context, (2) extracting facial expressions from the visual modality, (3) understanding relationships between speakers in a dialogue, (4) interpreting an utterance by taking previous parts of the dialogue into context.

In fact, advancements in mERC have frequently been made with novel methods that model this kind of reasoning. DialogueRNN (Majumder et al., 2018) introduces speaker context as well as context from previous time-steps. M2F-Net (Chudasama et al., 2022) focuses on extracting features relevant to facial expressions, and (Saxena et al., 2022) modeled the speaker contexts as a graph using graph neural networks, leveraging the graphical structure at hand. Interestingly, most approaches tend to model one of these reasoning objectives. Yet, they are not mutually exclusive. As a result, there may be opportunities to combine these techniques.

In this report, we analyse two different multimodal baselines on MELD which take two distinct approaches. First, we look at DialogueRNN (Majumder et al., 2018) which uses multiple modalities in addition to modeling speaker context. Second, we look at SPCL (Song et al., 2022) which is the state-of-the-art on MELD. SPCL is unimodal in nature so we extend it by fusing the audio and text modality. We call

[*]Equal contribution [1]Language Technologies Institute, School of Computer Science, Carnegie Mellon University, PA, USA. Correspondence to: Prasoon Varshney <pvarshne@andrew.cmu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

this extended version of SPCL **M**ulti**M**odal-**SCPL** (MM-SCPL). To the best of our knowledge, this is the first attempt in the research literature to do so.

We perform a thorough error analysis on these two baselines. In section 2 we look at related work. In section 3 we formally describe the research problem. In section 4 we describe the multimodal baselines. In section 5 we describe the dataset and the experiments we performed. In section 6 we present the results our experiments. Finally, in section 7 we discuss future research ideas.

2. Related Work

The task of ERC extends further back to 1974, where Ekman (1974) conducted a psychology study demonstrating that emotions could be classified given enough data. These emotions, namely joy, fear, anger, sadness, disgust, and surprise have been referred to as Ekman's universal emotions. The EmotionLines dataset (Hsu et al., 2018) is a supervised ERC dataset based on text utterances where the goal is to label each utterance as one of these six emotions (in addition to neutral, signifying the lack thereof). The Multimodal EmotionLines Dataset (MELD) (Poria et al., 2018) is a multimodal extension of the EmotionLines dataset that adds an acoustic and visual modality to these utterances {`a, v, t`}.

It should be noted that MELD is not the first multimodal ERC (mERC) dataset. Earlier datasets include IEMOCAP (Busso et al., 2008) and SEMAINE (McKeown et al., 2011). However, the conversations are dyadic in nature, which precludes the difficulty in tracking individual speaker states and handling co-reference. There are other, more recent mERC datasets such as CMU-MOSEI (Zadeh et al., 2018), MOSI, (Zadeh et al., 2016), and MOUD (Pérez-Rosas et al., 2013). However, unlike MELD they are not conversational. For this reason it is not possible to develop methods which leverage useful information such as speaker context, dialogue context, etc.

The authors of MELD implemented a number of baseline models for their dataset. These include text-CNN (Kim, 2014), bcLSTM (Poria et al., 2017), and DialogueRNN (Majumder et al., 2018). They applied these models with different combinations of the three modalities. However, they did not include the video modality in any of these experiments. The best performing model was the DialogueRNN (text + audio) with the weighted average F1 score of 60.25.

Since MELD was released, there has been a considerable amount of different, novel methods proposed. A commonality amongst twelve highest scoring methods on MELD[1] is the use of RoBERTa (Liu et al., 2019). What differentiates these approaches is how they leverage the additional information present in MELD to outperform methods simply classifying text utterances based on mere sentence-level embeddings. EmoBERTa (Kim & Vossen, 2021) does this by making the model speaker-aware by prepending speaker names to the utterances. Just as DialogueRNN achieved state of the by introducing speaker-context, RoBERTa at the time achieved state-of-the-art on MELD. As these two examples have shown it can be highly beneficial to apply reasoning by building the structured nature of a dialogue into the inference. Saxena et al. (2022) used graph neural networks to model both the dialogue participants and speaker personality. M2F-Net (Chudasama et al., 2022), the current multimodal state-of-the-art does not leverage context information to the same extent as suggested by EmoBERTa. Rather, it uses novel feature extractors that leverage features such as facial expressions in the video. The authors of M2F-Net do not specify whether or not they tried prepending the speaker names in front of each sentence. This does demonstrate, however, the advantage of extracting relevant features such as facial expressions. The current state-of-the-art, SPCL (Song et al., 2022), is unimodal in nature. SPCL takes advantage of prototypical networks (Snell et al., 2017b) to address the class imbalance problem in MELD. To the best of our knowledge there exists no published method in the research literature which attempts to utilize multiple modalities to extend SPCL.

3. Problem Statement

The goal is to classify an utterance u_i as having emotion k_j where $i \in \{1, ..., N\}$ and $j \in \{1, ..., K\}$. N is the number of utterances and $K = 3$ when classifying sentiments and $K = 7$ when classifying emotions.

In the dataset, there are M speakers, where $M < N$. This follows from the fact that each utterance u_i has only one speaker and there are multiple speakers in the dataset. We furthermore use u_t to denote the utterance at a particular time-step. Thus given a dialogue n of length T, we can denote the sequence of all utterences in that dialogue in the following manner: $u_0^{(n)}, u_1^{(n)}, ..., u_T^{(n)}$. Since we never reason about more than one dialogue at any given time, we will drop the n and write the sequence of utterances instead as $u_0, u_1, ..., u_T$.

There is a one-to-one mapping between utterances and speakers. The function s maps an utterance (a vector in d dimensions) to the index of its speaker $s : R^d \mapsto \mathbb{Z}^+$.

[1] https://paperswithcode.com/sota/emotion-recognition-in-conversation-on-meld

4. Multimodal Baselines

4.1. DialogueRNN

When DialogueRNN (Majumder et al., 2018) was released in 2019 it significantly improved the state of the art on two mERC tasks, IEMOCAP (Busso et al., 2008) and AVEC (Schuller et al., 2012). DialogueRNN was shortly thereafter used as one of the baselines in the original MELD (Poria et al., 2018) paper where it achieved the best performance amongst multiple baselines.

One of DialogueRNN's fundamental contributions was that it not only modelled features encoded from three different modalities {a,v,t} but also incorporating speaker information. This has been adapted in some form by more recent research, such as something that has since then been incorporated into other successful methods such as EmoBERTa(Kim & Vossen, 2021), M2F-Net (Chudasama et al., 2022), and GNNs for Emotion Recognition (Saxena et al., 2022). By modelling the knowledge of previous speakers and their utterances they perform reasoning (Liang et al., 2022) by modelling the task in a manner that might be obvious to humans but is unlikely so for a neural network.

Underlying the DialogueRNN is an architecture mostly consisting on Gated Recurrent Units (GRUs) (Cho et al., 2014). In total, DialogueRNNs employs three GRUs to model the speaker, preceding utterances, as well as their emotions.

Even though DialogueRNN was published in 2019, 7 months after BERT (Devlin et al., 2018), Majumder et al. (2018) employed a relatively outdated feature extractor, CNNs, proposed 5 years earlier (Kim, 2014). Majumder et al. (2018) did this in order to do controlled the study and ensure that their performance gain is due to their novel modelling of speaker context. For audio and visual feature extractors, the authors used 3D-CNN and openSMILE (Eyben et al., 2010), respectively.

4.1.1. Different states

DialogueRNN maintains, what they aptly name, *party state* for each person (party) in the conversation. The party state for person i is effectively the speaker state when the speaker of u_t is party i. We can write this formally as $s(u_t) = i$. The party state is an input to one of the GRUs: $GRU_{\mathcal{P}}$. The party state for party i t time step t is denoted as $q_{i,t}$ for any participant in the dialogue. The party state for the speaker at time step t can also be written as $q_{s(u_t),t}$

In addition to the party state, DialogueRNN maintains a *global state* which is shared amongst parties. The global state is also an input to one of the three GRUs and is recursively defined in the following equation:

$$g_t = GRU_{\mathcal{G}}(g_{t-1}, (u_t \oplus q_{s(u_t),t-1}))) \tag{1}$$

That is, global state g_t at time t is given by the previous global state g_{t-1} and the concatenation ($\oplus$) of the utterance at that timestep, u_t and the party state of the speaker at timestep t before he uttered u_t. Intuitively, u_t transforms $q_{s(u_t),t-1}$ to $q_{s(u_t),t}$. We use the former as the input to the global state. If we assume that the emotion of u_t is in some way reactionary to the speakers party state prior to uttering u_t then we can think of $q_{s(u_t),t-1}$ as encoding context which may result in a particular emotion for the speaker.

4.1.2. Context through attention

DialogueRNN uses global context c at every time step to update the speaker states and does so by attending over the global states $g_1, g_2, ...g_t$. It maintains a learnable matrix W_α which is used to compute the attention scores α for each of the global states up until time step $t-1$

$$\alpha = softmax(u_t \mathbf{W}_\alpha [g_1, g_2, ..., g_{t-1}]) \tag{2}$$

$$c_t = \alpha[g_1, g_2, ..., g_{t-1}] \tag{3}$$

The context c_t is then used to update the party state of the speaker in the following way:

$$q_{s(u_t),t} = GRU_{\mathcal{P}}(q_{s(u_t),t-1}, (u_t \oplus c_t)) \tag{4}$$

On the other hand, listener state for any listener i will not updated until i becomes a speaker.

4.1.3. Classifying emotions

The third and final GRU is used to model the emotion e at time step t:

$$e_t = GRU_{\mathcal{E}}(e_{t-1}, q_{s(u_t),t-1})) \tag{5}$$

Lastly, e_t is fed through a two-layer neural network to give a prediction $\hat{y}_t$ for u_t.

$$x_t = ReLU(W_A e_t + b_A) \tag{6}$$

$$\hat{y}_t = \arg\max_i (W_B x_t + b_B)_i \tag{7}$$

Available implementations of DialogueRNN[2] have achieved a weighted F1 score of 57.03. Using that same implementation we achieved a weighted F1 score of 57.08.

[2] https://github.com/declare-lab/conv-emotion/tree/master/DialogueRNN

4.2. MultiModal-SPCL (MM-SPCL)

4.2.1. SPCL

In their paper on Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation, Song et al. (2022) address the issue of class imbalance by leveraging prototypical networks (Snell et al., 2017a). Interestingly, the authors also address a problem that arises with MELD having being collected in a multimodal fashion which resulted in some utterances' text being misleading about the emotion that the utterance carries. To counter this, the authors leveraged curriculum learning (Bengio et al., 2009) to counter such extreme samples.

4.2.2. Encoding the context

We will extend the notation introduced in section 3 to effectively describe SPCL. Here, we use $\mathcal{D}$ to denote a dialogue. In contrast to DialogueRNN, where speakers were indices and utterances were d-dimensional vectors, here both are strings. The textual transcript for an utterance i is denoted as u_i (e.g. `"Oh, honey"`) and the name of its respective speaker is denoted as $s(u_t)$ (e.g. `"Rachel"`). Mathematically, we can describe a particular dialogue n with length T^n as:

$$\mathcal{D}^{(n)} = [s(u_1), u_1, s(u_2), u_2, ..., s(u_{T_n}), u_{T_n}] \tag{8}$$

In our problem setting, we consider the context of k utterances for a dialogue and denote it as as $D^{(n)}_{t-k:t}$. This context is then used as input to predict the label y_t for u_t. We call this the context for $\mathcal{D}^n$ at timestep t.

$$\mathcal{D}^{(n)}_{t-k:t} = [s(u_{t-k}), u_{t-k}, ..., s(u_t), u_t] \tag{9}$$

During training and inference, $D^{(n)}_{t-k:t}$ is converted into a string by concatenating all items in the context. We denote it as $\mathcal{S}^{(n)}_{t-k:t}$. Here is an example of such as string where $k = 3$.

```
"Rachel: What are you doing
here? Ross: hey, you know,
this building is on my paper
route. Rachel: Oh, honey."
```

Natural language can be encoded using BERT-like models by using the encoding of the `CLS` token as a representation of the input string. Instead, Song et al. (2022) use prompt-based learning (Liu et al., 2021) where they construct a prompt which extends the string $\mathcal{S}^{(n)}_{t-k:t}$ by appending `"for` u_t, s_t `feels <mask>"` at the end of it.

We denote the embedding for the `"<mask>"` token as z, which is then used for the downstream task of classifying the utterance u_t.

4.2.3. Prototypical Contrastive Learning

An important contribution from Song et al. (2022) is that of combining prototypical learning with contrastive learning to alleviate the class imbalance problem. To lay down terminology, we have a batch I wich is a set of `"<mask>"` token encodings $I = \{z_1...z_N\}$, a score function $\mathcal{G}$, and the vanilla supervised contrastive loss for z_i as:

$$\mathcal{F}(z_i, z_j) = exp(\mathcal{G}(z_i, z_j)/\tau) \tag{10}$$

$$\mathcal{P}_{sup}(i) = \sum_{z_p \in P(i)} \mathcal{F}(z_i, z_p) \tag{11}$$

$$\mathcal{N}_{sup}(i) = \sum_{z_j \in I_{-i}} \mathcal{F}(z_i, z_p) \tag{12}$$

$$\mathcal{L}^{sup}_i = -log \frac{1}{|P(i)|} \frac{\mathcal{P}_{sup}(i)}{\mathcal{N}_{sup}(i)} \tag{13}$$

Where $P(i)$ is the set of positive samples in I, and $I_{-i} = \{z_j \in I \mid i \neq j\}$. This setup, however, is impacted by class imbalance. To counter this, prototypical learning is employed.

A fixed-length queue for each emotion is maintained as $Q_j = [z^j_1...z^j_L]$ for the j^{th} emotion, a support set of size K is uniformly sampled from this. A prototype T_j is derived as the mean of the support set. With this, we augment the negative jscores of the i^{th} sample as $\mathcal{N}_{spcl}(i) = \mathcal{N}_{sup}(i) + \sum_{k \in \mathcal{E}_{-y_i}} \mathcal{F}(z_i, T_k)$ and the positive scores as $\mathcal{P}_{spcl}(i) = \mathcal{P}_{sup}(i) + \mathcal{F}(z_i, T_{y_i})$. $\mathcal{E}$ is the set of all emotion labels and $\mathcal{E}_{-y_i} = \{y_j \in \mathcal{E} \mid i \neq j\}$. Finally, this gives the SPCL loss:

$$\mathcal{L}^{spcl}_i = -log \frac{1}{|P(i)| + 1} \frac{\mathcal{P}_{spcl}(i)}{\mathcal{N}_{spcl}(i)} \tag{14}$$

4.2.4. MM-SPCL

As an initial attempt to extend SPCL into a multimodal method MM-SPCL, we use representation fusion between z_i, the unimodal representation of emotion extracted via the context encoder, and the audio features a_i provided in MELD for each utterance u_i. We derive a multimodal representation of emotion and audio, $\hat{z}_i$ as:

$$\hat{z}_i = \mathbf{W}(z_i \oplus a_i) \tag{15}$$

The dimensions are as follows: $a_i \in \mathbb{R}^{300}$, $z_i \in \mathbb{R}^{1024}$, $(z_i \oplus a_i) \in \mathbb{R}^{1324}$, $\mathbf{W} \in \mathbb{R}^{1024 \times 1324}$, and $\hat{z}_i \in \mathbb{R}^{1024}$.

5. Experimental Methodology

5.1. Dataset

For this project, we focus on the mERC task on MELD (Poria et al., 2018). MELD is a multimodal extension of the EmotionLines Dataset introduced by Hsu et al. (2018), which contains data in the acoustic, visual, and text modalities {a, v, t}. MELD contains a total of 1432 dialogues where each dialogue contains a sequence of utterances. In total there are 13708 utterances. Each utterance consists of a video clip {a, v} and a textual transcript {t}, along with two sets of labels that describe the emotion and sentiment, respectively. There are 7 emotion classes in the dataset, joy, sadness, surprise, fear, disgust, anger, neutral, and three sentiment classes, positive, negative, neutral. From a preliminary exploratory data analysis on the dataset, we observe a class imbalance, where the neutral label presents itself as a large majority in both the emotion and sentiment classes.

The dataset has already been split into train, validation, and test sets. The train set contains 1038 dialogues (9989 utterances). The validation set contains 114 dialogues (1109 utterances). Finally, the test set contains 280 dialogues (2610 utterances).

5.2. Models

We trained the models on a g4dn.2xlarge AWS EC2 instance with a 16GB NVIDIA Tesla T4 GPU. Our code implementation of the baseline models can be found on GitHub[3].

5.2.1. DialogueRNN

We trained a DialogueRNN model using the implementation that is listed on the MELD benchmark[4]. Due to the speed of training, we were able to train the model on the full dataset. We kept the hyperparameters the same.

For the three different GRU cells, $GRU_{\mathcal{P}}, GRU_{\mathcal{G}}$, and $GRU_{\mathcal{E}}$ the hidden states are 150, 150, and 100 respectively. The size of the hidden layer given by W_A is 200. Dropout for the linear layers is 0.5 while the dropout for the recurrent cells is 0.3.

We train the model for 60 epochs with a batch size of 30, learning rate of 1e-4, using the Adam optimizer (Kingma & Ba, 2014), and with the L2-loss coefficient set to 3e-5.

[3] https://github.com/abuzar08/11777-F22-Project

[4] https://github.com/declare-lab/conv-emotion

5.2.2. MM-SPCL

We trained the model for 10 epochs, with the temperature parameter set to 0.05, a batch size of 8, a support set size of 64, and a seed of 2333 for reproducibility. We have kept all other hyperparameters to the default values as specified by Song et al. (2022).

5.3. Evaluation

We will evaluate the performance of the models using their weighted F1-score, which is the weighted mean of F1-measure with weights equal to class probability. This is because the class distribution in MELD is imbalanced and using metrics such as accuracy will be extremely biased towards the classification performance on the majority class.

6. Results and Discussion

6.1. Quantitative analysis

Table 1 shows the results of the three baseline models on the MELD dataset. While the authors of SPCL reported an F1-score 67.25%. By running the same, publicly available model, on the same data, our implementation achieves an F1-score of 66.49%. This is likely due to different random seeds and different batch sizes. As reported by the authors of SPCL, its performance increases with a larger batch size. This can be attributed to larger batches leading to more positive samples in a batch for minority classes, and negative samples in a batch for the majority class (neutral emotion). Although Song et al. (2022) employ prototypical learning to combat this, having a higher batch size would regardless likely result in better performance and lower variance over different random seeds.

Model	F1 Score	Accuracy (%)
SPCL	0.6649	66.4
MM-SPCL	0.6614	65.6
DialogueRNN	0.5708	59.4

Table 1. Weighted average F1 score and accuracy metrics for the three baseline models.

In table 1 we can see that both SPCL and MM-SPCL perform much better than DialogueRNN in terms of F1 scores, which is expected. Therefore going forward, we focus our analyses on comparing MM-SPCL and SPCL. This is also motivated by the fact that one of our research ideas focuses on attempting to imporve the state-of-the-art for mERC on MELD, as explained in Section 7.

Table 2 presents summary statics for agreements between SPCL and MM-SPCL. We notice that both models agree on the emotions for 87.39% of test utterances. When they do disagree, SPCL is correct more often (6.74% of test

Model	# Correct	(%) of Test Set
SPCL	1734	66.44
MM-SPCL	1711	65.56
Either Correct	1887	72.30
Both Correct	1558	59.69
Both Wrong	723	27.70
Only SPCL Correct	176	6.74
Only MM-SPCL Correct	153	5.86

Table 2. Distribution of prediction correctness and errors between SPCL and MM-SPCL

utterances) than MM-SPCL (5.86%). Another interesting observation is that at least one of the models is correct 72.30% of the time.

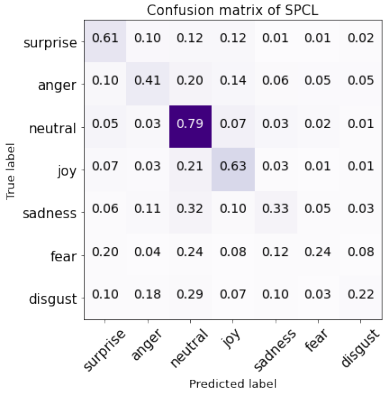

Figure 1. Normalized confusion matrix of SPCL

Figures 1 and 2 show confusion matrices for the predictions of SPCL and MM-SPCL, respectively, on the test set. It is evident that both models are equally good predictors of `neutral` and `anger`, however there are slight differences over other emotions. Along the diagonal, we observe that MM-SPCL is slightly better at identifying `joy` and `disgust`, while SPCL is better at `surprise`, `sadness`, and `fear`.

6.2. Qualitative error analysis

6.2.1. Analysing disagreements

In table 3, we have present three dialogues which describe three kinds of disagreements. A more extensive version of the table can be found in the appendix A.

Dialogue 1 This dialogue exemplifies the lack of context in the audio modality, and how this results in erroneous predictions when its effect dominates. Upon listening to the audio of each utterance in isolation, the tone of the speaker does indeed sound assertive (almost angry), however, from context it is clear that the scene depicts a joyous moment.

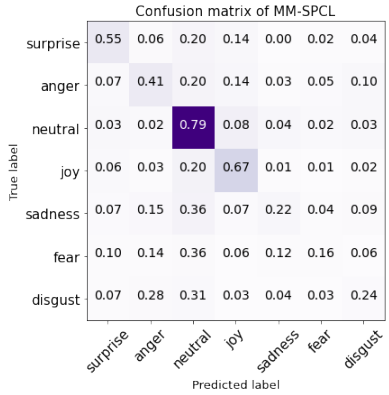

Figure 2. Normalized confusion matrix of MM-SPCL

Dialogue 85 This dialogue is an example of where we (as human annotators) agree more often with the predictions of MM-SPCL than either SPCL or the ground truth labels. A particularly interesting case is utterance 1, where we find that the emotion is more of a mix of `fear` and `surprise`, than either in isolation, and very far from being `anger`.

Dialogue 111 This is an example of MM-SPCL being correct over SPCL, and upon listening to the audio, we attribute this to the fact that in this particular case, most of the information about the emotion of the speaker was contained in the audio.

6.2.2. Agreement

Table 4 shows an example of SPCL and MM-SPCL agreeing on the predicted emotion, but neither getting it right. A longer list of examples can be found in table 4 in Appendix A.

Dialogue 17, 88 In both of these examples we observed that the video modality contained the information which was required to correctly predict the emotion. We also observe that in some cases, though the audio modality has the information it is unable to correctly predict the emotion, and we attribute this to the features not being learned for the downstream task of emotion recognition.

6.3. Bias analysis

We have manually labeled the gender of the speakers in the test data by examining the relevant video clips associated to each utterance. We used this information to evaluate the emotion classification performance of the SPCL models with respect to the different genders.

6.3.1. Gender bias in SPCL (unimodal)

The weighted F1-score for emotion classification for the female and male speakers are 0.62 and 0.64 respectively. This

Dia,Utt ID	Utterance	True Label	SPCL	MM-SPCL	Remarks
1, 2	Joey: Push 'em out, push 'em out, harder, harder.	joy	joy	anger	Assertive, almost angry tone
1, 3	Joey: Push 'em out, push 'em out, way out!	joy	anger	anger	Assertive, almost angry tone
85, 0	Joey: But um, I don't think it's anything serious.	neutral	neutral	neutral	Chandler and Joey are worried
85, 1	Chandler: This sounds like a hernia. You have to—you-you—Go to the doctor!	surprise	anger	fear	since Joey is in pain, and
85, 2	Joey: No way!	anger	anger	fear	fear explains the emotion better than anger.
85, 3	Joey: 'Kay look, if I have to go to the doctor for anything it's gonna be for this thing sticking out of my stomach!	anger	anger	fear	Questionable ground truth
111, 6	Mrs. Bing: Really stupid.	sadness	disgust	sadness	MM-SPCL gets audio context right
279, 12	Ross: They're not listening too me?	surprise	anger	surprise	MM-SPCL gets audio context right

Table 3. Disagreements between predicitons of MM-SPCL and SPCL on the test set. The values under SPCL and MM-SPCL denote the predictions of the corresponding model.

Dia,Utt ID	Utterance	True Label	SPCL	MM-SPCL	Remarks
17, 4	Don't look honey. Change the channel! Change the channel!	disgust	anger	anger	Disgust portrayed in audio and video, not text
17, 6	What a wank!	anger	disgust	disgust	Angry voice and facial expressions
88, 8	Joey: Can we please turn this off?	sadness	anger	anger	Sadness clearly conveyed in audio and video

Table 4. An example of an erroneous agreement between predictions of SPCL and MM-SPCL . The values under SPCL and MM-SPCL denote the predictions of the corresponding model.

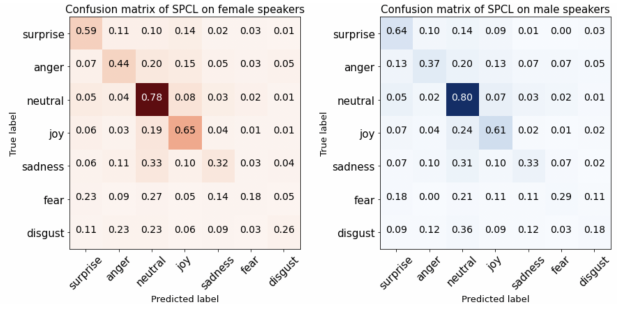

Figure 3. Normalized confusion matrix of SPCL on female speakers (left) and male speakers (right)

signifies that the unimodal SPCL model performs uniformly overall on classifying the emotions of male and female speakers.

To examine the classification performance disparities more closely, we can observe the confusion matrix on figure 3. We can see that for the female speakers, the model is able to classify the emotions `anger`, `joy`, and `disgust` better than the male speakers. Conversely for the male speakers, the model is able to classify the emotions `surprise`, `neutral`, `sadness`, and `fear` better than the female speakers.

It is also important to note that for most of the misclassified emotions, the unimodal SPCL model has the tendency to classify them as `neutral` for both genders. However, this tendency to assign a `neutral` label is more often seen for the male speakers, while the misclassified emotion labels for the female speakers are more evenly distributed. We can't draw any significant conclusions from the results of the unimodal SPCL model alone, as the performance is relatively uniform across the various emotions for the different speakers with respect to gender.

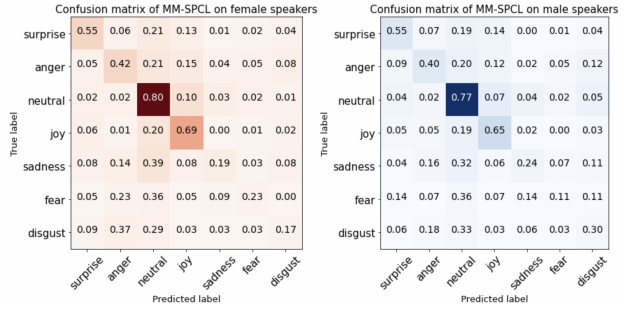

Figure 4. Normalized confusion matrix of MM-SPCL on female speakers (left) and male speakers (right)

6.3.2. Gender bias in MM-SPCL (bimodal)

The weighted F1-score for emotion classification for the female and male speakers are 0.61 and 0.63 respectively. Similarly, the bimodal MM-SPCL model performs uniformly overall on classifying male and female speaker emotion.

Paying close attention to the confusion matrix on figure 4, we can observe that the bimodal MM-SPCL model is better at classifying the `anger`, `neutral`, `joy`, and `fear` for female speakers, while it is better at classifying `sadness` and `disgust` for male speakers.

Since the bimodal MM-SPCL model takes into account the audio features as well as the text features, by comparing the values in the male and female confusion matrices in figure 4, we can extrapolate that the way female characters convey their utterances are more telling of their emotion than the male characters. This may be an indicative sign of bias that we need to be aware of, where female characters are portrayed as more emotional with respect to the tone in which they deliver their lines in.

7. New Research Ideas

MELD is designed for multimodal tasks, given that the utterances have been labeled with all three modalities {`a, v, t`} in mind. The current state-of the-art (SPCL) however, utilizes just the `text` modality. This causes two issues; (1) The model fails to incorporate information from other modalities, and (2) using this dataset for unimodal tasks, as in SPCL, requires some way to mitigate effects of extreme samples wherein the information present in the modality in question (for SPCL, `text`) is almost misleading, because the right information needed to predict the emotion correctly lies in other modalities. For the first issue, we propose to build a multimodal extension of SPCL (MM-SPCL) in a more sophisticated manner as motivated by our findings from the error analysis. The second issue, in SPCL is taken care of by utilizing Curriculum Learning. For MM-SPCL, it might also be worth looking into either removing curriculum learning or updating it.

With this in mind, in this section we will talk about our primary research ideas towards building MM-SPCL grounded in conclusions drawn from our error analysis. It should be noted, however, that some of our previous ideas such as quantification of modality impact, feature importance, and bias analyses will be conducted as post-hoc analyses pertinent to the research.

7.1. Generating better audio representations

For the analysis in this paper, we directly fused audio features present in MELD genereated through openSMILE (Eyben et al., 2010) with the textual representation of emotions in SPCL. It is important to note that these audio features are not learned for emotion prediction or to have explicit character awareness. The audio features are also devoid of any temporal context within the dialogues. Further, we know from our previous analysis (scatter plot in figure 5 in Appendix B) that audio features are not separable by emotions using t-SNE unlike glove-based text features. This points to a need for contrastive-learning based audio representations that help hard examples belonging to different emotions to be pushed farther and similar ones to come closer.

Further, there are 153 (5.86%) cases where MM-SPCL utilizes the audio modality well, but 176 (6.74%) other instances where SPCL performs better and MM-SPCL gets confounded. We hypothesize that with our current simple fusion strategy of concatenating non-contextualized audio and SPCL's contextualized text representations, the audio features can sometimes overpower the inferences made on the text modality, therefore, we wish explore better fusion strategies.

We propose to learn audio features via multi-task contrastive learning. The framework will follow a feature extractor $\mathcal{F}_a$ followed by two classifiers for the tasks in question, i.e. emotion recognition and character prediction, to allow for the features learned by $\mathcal{F}_a$ to have information about emotion in the utterance as well as the speaker. The features from $\mathcal{F}_a$ will also be used for contrastive learning to allow for these features to be separable in the emotion space. Next, considering each dialogue as a sequence of audios, we will use a recurrent network (Cho et al., 2014; Hochreiter & Schmidhuber, 1997) that takes in the learned features from $\mathcal{F}_a$ for the audio corresponding to each turn, and train it for ERC (in conversation now, since we have context). The final encodings will be residually added to the input representations to preserve information. Finally, we will fuse the features learned by this recurrent network with the textual representations in SPCL towards MM-SPCL.

This brings us to the fusion itself. We propose to use an attention-based bottlenecked fusion mechanism as described in (Nagrani et al., 2021), where restricting the flow between modalities in the fusion layer into a few attention nodes localizes information from both modalities and encourages maximization of sharing of salient information between the two modalities to build a better combined representation.

7.2. Incorporating video representations

As presented in Tables 4 and 6, we noticed multiple clips where the ground truth label was clearly portrayed by facial gestures, and as human judges, we didn't find enough information in either the text or audio to label it with the correct emotion. This goes to affirm the need for incorporating the video modality to augment our MM-SPCL model with information about facial expressions and other artifacts in

the video modality.

For this, we plan to use the method described in M2FNet (Chudasama et al., 2022), the previous state-of-the-art, to build video representations. This would entail using an MTCNN (Zhang et al., 2016) facial detector to come up with candidate bounding boxes for faces and passing them through a facial feature extractor trained on emotion classification to obtain emotion-relevant features for each face in a video. These can further be concatenated with an image-frame embedding to provide scene context in the video that will then be fed into a bottlenecked fusion mechanism as described in Section 7.1.

Further, we propose to test the 2D CNN approach in CLIP-BERT (Lei et al., 2021) towards leveraging widely available image-text pretraining for generating contextualized encodings for sampled frames from the video (with text being the text of the utterance), combine the frame-level representations using a simple weighted sum, and training these representations for MERC. These representations will then be fused with the text representations from SPCL.

7.3. Debiasing methods

There exists a disparity in how the speakers in MELD (characters in the Friends TV show) express emotions in different ways and modalities. It is evident from our bias analysis that the error distribution between male and female characters changes between the unimodal SPCL and bimodal MM-SPCL. There lies an opportunity to test for emotion intensity contained in different modalities across character demographics, as well as its effects on the final multimodal model emotion recognition performance.

Most of the current debiasing efforts focus on unimodal representations, particularly in the text modality. Recently, Wang et al. (2022) extended their previous work on double-hard debiasing of word embeddings (Wang et al., 2020) to the multimodal space by proposing a projection debias method to mitigate gender and age bias in visual representation. In their work, they introduced the Multibias-Mitigated and sentiment Knowledge Enriched Transformer (MMKET) architecture, which is able to leverage the debiased contextual and multimodal signals to predict emotions of the target utterance for the mERC task, due to its ability to capture the context and fast computation. Although they were able to mitigate biases across the age and gender axes in both the representations of the text and visual modalities, the debiased model saw a reduction in overall emotion recognition performance. The authors credit this to the fact that human emotions inherently contain prejudice, and thus removing the bias will reduce the ability of the model to classify emotion.

It would be interesting to investigate how their bias mitigation technique perform on aligned representations, or explore other debiasing techniques that reduces the disparity between the multimodal emotion recognition performance of the male and female speakers without reducing the overall model performance.

References

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359, 2008.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., and Onoe, N. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4652–4661, 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ekman, Friesen, O. C. D.-T. H. K. L. P. R.-B. Universals and cultural differences in the judgments of facial expressions of emotion, 1974. URL https://pubmed.ncbi.nlm.nih.gov/3681648/.

Eyben, F., Wöllmer, M., and Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

Hsu, C.-C., Chen, S.-Y., Kuo, C.-C., Huang, T.-H., and Ku, L.-W. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1252.

Kim, T. and Vossen, P. Emoberta: Speaker-aware emotion recognition in conversation with roberta, 2021. URL https://arxiv.org/abs/2108.12009.

Kim, Y. Convolutional neural networks for sentence classification, 2014. URL https://arxiv.org/abs/1408.5882.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T. L., Bansal, M., and Liu, J. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7331–7341, 2021.

Levis, J. M. Intonation in theory and practice, revisited. *TESOL quarterly*, 33(1):37–63, 1999.

Liang, P. P., Zadeh, A., and Morency, L.-P. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions, 2022. URL https://arxiv.org/abs/2209.03430.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021. URL https://arxiv.org/abs/2107.13586.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. Dialoguernn: An attentive rnn for emotion detection in conversations, 2018. URL https://arxiv.org/abs/1811.00405.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 2011.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34: 14200–14213, 2021. URL https://proceedings.neurips.cc/paper/2021/file/76ba9f564ebbc35b1014ac498fafadd0-Paper.pdf.

Pérez-Rosas, V., Mihalcea, R., and Morency, L.-P. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 973–982, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1096.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 873–883, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1081. URL https://aclanthology.org/P17-1081.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2018. URL https://arxiv.org/abs/1810.02508.

Saxena, P., Huang, Y. J., and Kurohashi, S. Static and dynamic speaker modeling based on graph neural network for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pp. 247–253, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-srw.31. URL https://aclanthology.org/2022.naacl-srw.31.

Schuller, B., Valstar, M., Cowie, R., and Pantic, M. Avec 2012: The continuous audio/visual emotion challenge - an introduction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pp. 361–362, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314671. doi: 10.1145/2388676.2388758. URL https://doi.org/10.1145/2388676.2388758.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017a.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning, 2017b. URL https://arxiv.org/abs/1703.05175.

Song, X., Huang, L., Xue, H., and Hu, S. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv preprint arXiv:2210.08713*, 2022.

Wang, J., Ma, F., Zhang, Y., and Song, D. A multibias-mitigated and sentiment knowledge enriched transformer for debiasing in multimodal conversational emotion recognition. *arXiv preprint arXiv:2207.08104*, 2022.

Wang, T., Lin, X. V., Rajani, N. F., McCann, B., Ordonez, V., and Xiong, C. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5443–5453, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.484. URL https://aclanthology.org/2020.acl-main.484.

Zadeh, A., Zellers, R., Pincus, E., and Morency, L.-P. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503, 2016. doi: 10.1109/LSP.2016.2603342.

A. Appendix: Detailed Dialogue Examples

Dia,Utt ID	Utterance	True Label	SPCL	MM-SPCL	Remarks
1, 0	Joey: Come on, Lydia, you can do it.	neutral	neutral	neutral	
1, 1	Joey: Push!	joy	joy	anger	
1, 2	Joey: Push 'em out, push 'em out, harder, harder.	joy	joy	anger	Assertive, almost angry tone
1, 3	Joey: Push 'em out, push 'em out, way out!	joy	anger	anger	Assertive, almost angry tone
1, 4	Joey: Let's get that ball and really move, hey, hey, ho, ho.	joy	joy	joy	Switches to playful joyous tone
85, 0	Joey: But um, I don't think it's anything serious.	neutral	neutral	neutral	Chandler and Joey are scared
85, 1	Chandler: This sounds like a hernia. You have to—you-you—Go to the doctor!	surprise	anger	fear	since Joey is in pain, and
85, 2	Joey: No way!	anger	anger	fear	fear explains the emotion better than anger.
85, 3	Joey: 'Kay look, if I have to go to the doctor for anything it's gonna be for this thing sticking out of my stomach!	anger	anger	fear	Therefore, the ground truth is questionable
111, 3	Chandler: You kissed my best Ross!	anger	joy	anger	
111, 4	Mrs. Bing: O-kay. Look, it, it was stupid.	sadness	sadness	sadness	
111, 5	Chandler: Really stupid.	anger	disgust	disgust	
111, 6	Mrs. Bing: Really stupid.	sadness	disgust	sadness	MM-SPCL gets audio context right
279, 11	Rachel: Yeah, I mean, come on Ross, no one will even notice...	neutral	anger	neutral	
279, 12	Ross: They're not listening too me?	surprise	anger	surprise	MM-SPCL gets audio context right
279, 13	Rachel: Of course they're listening to you! Everybody listens to you.	neutral	anger	neutral	

Table 5. (Full) Disagreements between MM-SPCL and SPCL on the test set, and our remarks after listening to the raw audio files for the corresponding dialogue and utterance files. The values under SPCL and MM-SPCL denote the predictions of the corresponding model.

Dia,Utt ID	Utterance	True Label	SPCL	MM-SPCL	Remarks
17, 3	Ewww! Oh! It's the Mattress King!	disgust	surprise	disgust	
17, 4	Don't look honey. Change the channel! Change the channel!	disgust	anger	anger	Disgust portrayed in audio and video, not text
17, 5	Wait! Wait! I wanna see this. After I divorce him, half of that kingdom is gonna be mine.	joy	surprise	surprise	
17, 6	What a wank!	anger	disgust	disgust	Angry voice and facial expressions
88, 8	Joey: Can we please turn this off?	sadness	anger	anger	Sadness clearly conveyed in audio and video
88, 9	Rachel: Noo way, Kevin.	joy	disgust	disgust	Teasing, sarcastic, but joyous not disgust

Table 6. Examples of erroneous agreements between SPCL and MM-SPCL, and our remarks after listening to the raw video files for the corresponding dialogue and utterance files. The values under SPCL and MM-SPCL denote the predictions of the corresponding model.

B. t-SNE Plots

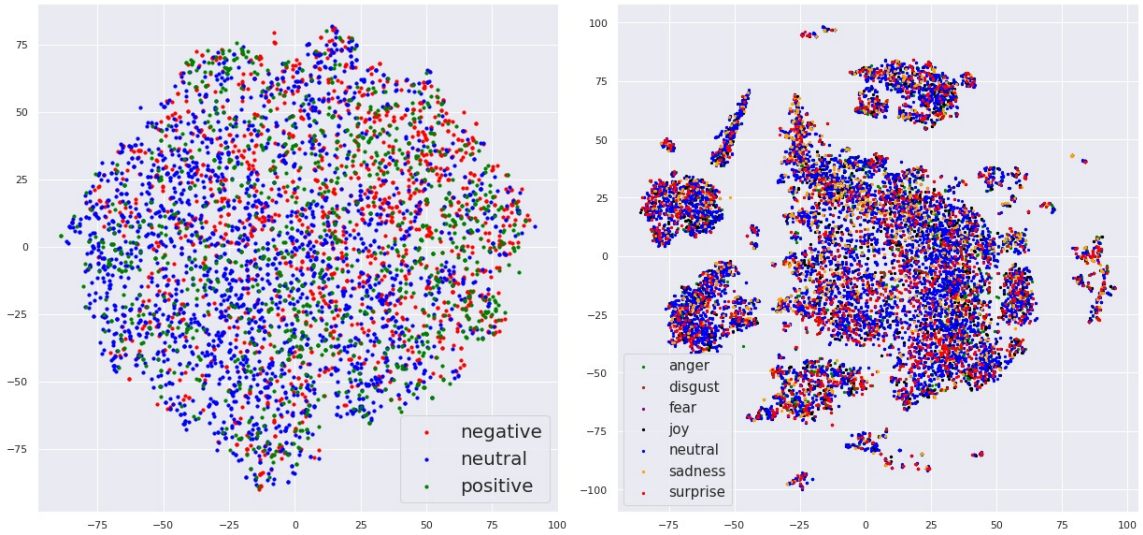

Figure 5. 2D t-SNE plot for audio features, sentiment (left) and emotion (right).

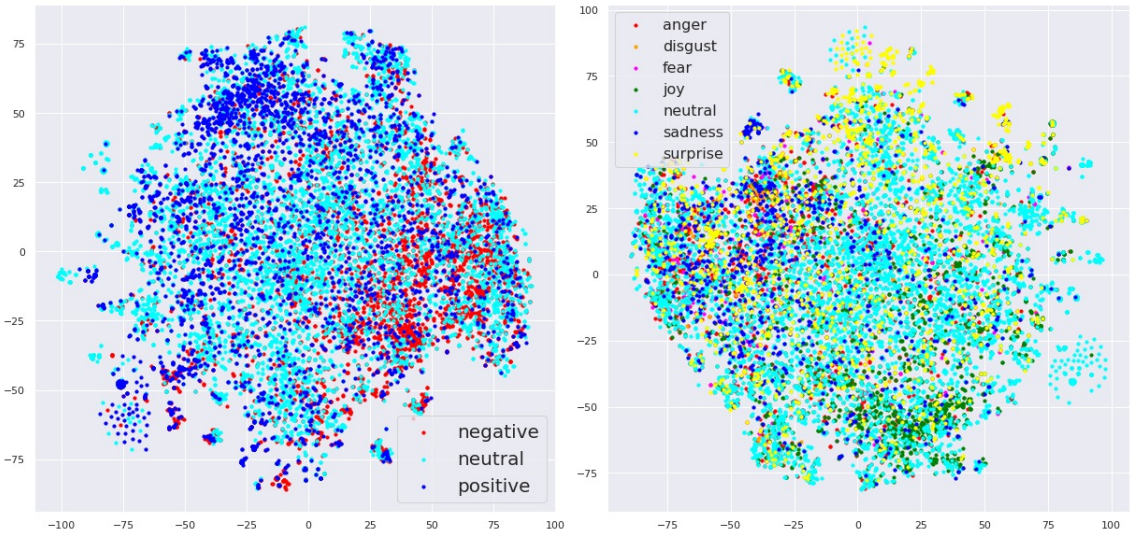

Figure 6. 2D t-SNE plots of BERT embeddings of utterances with neutral labels included, sentiment (left) and emotion (right).