Multimodal Emotion Prediction in Group Conversation

TOMA ALLARY, THIERRY BÉDARD-CORTEY, PHILIPPE BERGERON, YUAN LI

Abstract

Emotion prediction in group conversations is a challenging task due to the complexity of human interactions and the multimodal nature of emotional expressions.This project aims to develop a multimodal model that integrates textual and audio features for emotion prediction in group conversations. It combines RoBERTa for text embeddings, WavLM for audio embeddings, and a transformer-based fusion layer to enhance emotion recognition. The project is based on the study ”Emotional Cues Extraction and Fusion for Multimodal Emotion Prediction and Recognition in Conversation”[1].

1. Introduction

Human emotions are important for the communication. In recent years, deep learning has been achieving tremendous success on numerous machine learning applications. In natural language and speech processing areas, conversational AI had a huge growth in recent years — from virtual assistants to customer support chatbots. But there's still a big gap: most systems don't understand how we feel. That's why emotion-aware systems are so valuable — they enable more natural, empathetic, and human-like interactions. In this project, Our goal is to build a multimodal machine learning model that could predict the next emotion in a group conversation using traces of previous audio and text. More specifically, we explore simplified architectures to improve on current existing models that are complex and heavy to train and implement. Light and re-usable architecture are generally more suited for real-life tasks where cost and time are important. This work could serve as the basis for applications such as social robots, therapeutic robots and intelligent tutoring systems.

TODO: add contribution

2. Relate Work

There are three major approaches commonly used in emotion detection models. First is RNN-based models, which are good for temporal sequences but suffer with long-term dependencies and slow processing. The second is Transformer-based models, which excel at capturing long-range dependencies and can process data in parallel — though they are resource-intensive. The third is Graph-based models, which represent conversations as interactions between speakers, allowing them to infer emotions based on relationships. These models are powerful but often complex to implement. Our project mainly leveraged transformer-based models for their strong performance on language and speech tasks.

This study (**?**) explores the extraction and fusion of emotional cues from multiple modalities, including text and audio, to enhance emotion prediction in conversations. The authors propose a transformer-based fusion layer to combine embeddings from different modalities, achieving state-of-the-art results in emotion recognition tasks.

2.1. RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) (**?**)is a transformer-based language model developed by Facebook AI. This model significantly enhances the original BERT architecture through several key modifications in its pre-training methodology. Unlike BERT, RoBERTa was trained on a substantially larger dataset, totaling 160 GB of diverse text and code. This extensive training corpus allowed the model to develop a more robust and generalized understanding of language. One of the critical improvements in RoBERTa is the removal of the Next Sentence Prediction (NSP) objective used in BERT, which was found to not consistently improve downstream task performance. Additionally, RoBERTa employs dynamic masking during pre-training, where the masked tokens change in each epoch, forcing the model to learn more effectively from different parts of the input. The model also benefits from being trained with much larger batch sizes, contributing to more stable and efficient learning. RoBERTa's architecture, being based on the transformer, inherently excels at capturing long-range dependencies within text, making it particularly adept at understanding context. In dialogue, emotions are often conveyed subtly and are deeply intertwined with the surrounding linguistic context. RoBERTa's enhanced semantic capabilities, stemming from its extensive training and architectural refinements, enable it to discern these intricate relationships and make more accurate emotion predictions by considering the nuances of the conversational flow and the interplay of words.

2.2. WavLM

WavLM (**?**), short for Waveform Language Model, represents a significant advancement in self-supervised learning for comprehensive speech processing, developed by Microsoft. This large-scale model is meticulously engineered to learn rich, universal representations directly from raw audio waveforms, eliminating the need for extensive manual feature engineering. Building upon the success of the wav2vec 2.0 framework, WavLM introduces innovative techniques to further enhance the quality of learned representations. A key contribution of WavLM is its emphasis on capturing prosodic features like tone, rhythm, and pitch – elements that carry crucial information about a speaker's emotional state. By explicitly modeling these acoustic characteristics, WavLM gains a distinct advantage in tasks such as emotion analysis, where these subtle vocal cues are paramount. Beyond emotional understanding, WavLM demonstrates its versatility by supporting a wide array of speech-related tasks, including speaker identification, automatic speech recognition, and emotion recognition. This broad applicability makes WavLM an exceptionally valuable component for multimodal fusion approaches, where information from different modalities, such as audio and text, can be combined to achieve a more holistic and accurate understanding of communicative signals. Its ability to extract meaningful features from raw audio across various tasks underscores its potential to significantly improve the performance of speech-centric artificial intelligence systems.

3. Task Definition

Emotion Prediction in Conversation (**EPC**) In a multi-modal multi-party (or dyadic) dialogue containing text and audio D $= (u_1, s_1), (u_2, s_2), ..., (u_N, s_N)$, where (u_i, s_i) represents the ith utterance-speaker pair in the conversation, N is the number of utterances in the dialogue. EPC aims to predict the emotion category label $emotion_{n+1}$ of the future utterance speaker pair (u_{n+1}, s_{n+1}) by given the historical dialogues $(u_1, s_1), (u_2, s_2), ..., (u_n, s_n)$.

3.1. Dataset

MELD is a popular dataset for tasks involving the analysis of emotions expressed by multiple speakers. It encompasses a collection of over 1400 dialogues and 13,000 speech instances extracted from the television show Friends. Emotion annotation in the dataset includes: neutral, happiness, surprise, sadness, anger, disgust, and fear.

4. Method

4.1. Data Preprocess and Feature Extraction

In this project, we used a multimodal feature extraction method that combines both audio and text information to better predict emotions.

For audio features, we used a pre-trained model called WavLM (wavlm-base-plus) from Hugging Face. First, the audio files (in .mp4 format) were loaded and converted into raw waveforms. These waveforms were then passed into WavLM to generate hidden representations. The output from WavLM captures important speech characteristics like tone, pitch, speaking style, and voice quality, all of which are helpful for recognizing emotions.

For text features, we used the RoBERTa-base model. We read the utterances from a CSV file, then used the RoBERTa tokenizer to split each sentence into tokens that the model can understand. After tokenization, the tokens were passed into the RoBERTa model to get sentence embeddings. These embeddings represent the meaning, sentiment, and emotional hints inside the text.

After extracting features from both audio and text separately, we combined (concatenated) them together into a single feature vector for each sample. This combined feature includes both what was said (text) and how it was said (audio), giving the model a more complete view of the emotional content.

This method allows us to use both modalities at the same time, which improves the model's ability to detect complex emotions that may not be clear from just the audio or just the text alone.

In addition, before extracting features, we combined the audio clips and merged the text utterances for each conversation. Instead of treating each utterance separately, we processed the whole conversation as a single input. For audio, we concatenated the waveform files belonging to the same dialogue into one long audio sequence. For text, we joined the utterances from the same conversation into one long text string. This way, the models could capture the full context of the conversation — including how emotions develop over time — rather than just focusing on isolated sentences. After combining, we extracted features from the full conversation-level audio and text, and then fused them together for downstream emotion prediction.

TODO: add image

4.2. Reference baseline method

Our baseline architecture is from study (**?**). As seen in Figure 1, they propose an architecture using both text and audio modalities. They also extract a mel-spectrogram as a third modality from the audio. Their first steps are to extract features from modalities using pre-trained models (i.e. RoBERTa, WaveLM and a Spectrum Extraction Tool)

Figure 1. Schema of the baseline architecture. The left side is the main framework of the model. (a) multi-modal fusion module, the green part represents the pre-trained language transformer layer, the yellow part represents the pre-trained vision transformer layer, the pink part is trainable, and other parts are frozen.

4.2.1. KWRT

On textual side, they add a bloc of Knowledge-Based Word Relation Tagging (KWRT) before extracting their features. This block aims to give more word relation cues that context alone could not provide. This technique leverage external resources to identifies relationship between two words. Usually this is done through some human involvement in the labelling or learning process. In their specific use-case, their architecture use ConceptNet (**?**), a knowledge graph that connects words and phrases of natural language with labeled edges.

4.2.2. Prosody Enhancement

As for the waveform, the second modality, they use what they call a ”Prosody Enhancement”. Prosody is the study of intonation, stress and rhythm of spoken audio. They enhance the emotional cues given by the audio through a fine-tuned encoder (**?**). Their claim is that this module extract and amplify emotional cues.

4.2.3. Transformer module

TODO: explain Transformer module

4.2.4. Fusion module

In the two-step multi-modal fusion part, they first fuse the text and audio features enhanced in Knowledge-based Word Relation Tagging (KWRT) module and a Prosody Enhancement (PE) module , then combine the initial fusion features with the mel-spectrogram extracted from the audio to generate the final multi-modal fusion representation.

They use two pre-trained transformer models. For the language features,they utilized BERT (Bidirectional Encoder Representations from Transformers), a seminal model proposed by Devlin et al. in 2018. BERT introduced a deep bidirectional training approach using transformers, allowing the model to understand the context of a word based on both its left and right surroundings. For the vision features, they use Vision Transformer (ViT), introduced by Dosovitskiy et al. in 2021. This model applies the transformer architecture directly to sequences of image patches. ViT divides an image into fixed-size patches, linearly embeds them, and feeds the sequence into a standard transformer encoder. This design enables the model to capture global dependencies among patches, leading to state-of-the-art performance on image classification tasks when trained at scale.

4.3. First architecture draft

Even though the baseline model yield respectable results, having such complexity and that many components can be complex to implement, train and deploy. Hence, our main idea is to build a simpler architecture that result in similar performance with fewer parameters and less complexity.

We decided to remove some of the proposed modules to keep only the core modules that can grasp the context of a conversation. Intuitively, we pruned the spectrogram extraction, the third modality, as WaveLM already has a significant capacity to treat the audio cues. Also, KWRT and Prosody Enhancement modules are not in our proposal, see Figure 2. This first draft is however not really detailed and require specific solution to fuse both modalities without losing each extracted features. In Figure 2, we supposed N is fixed, but in reality, N differ for each conversation and even between both modalities. The next sections cover three different approaches we developed in parallel to tackle these issues and compare the results.

4.4. Convolution Neural Network

Our first approach TODO

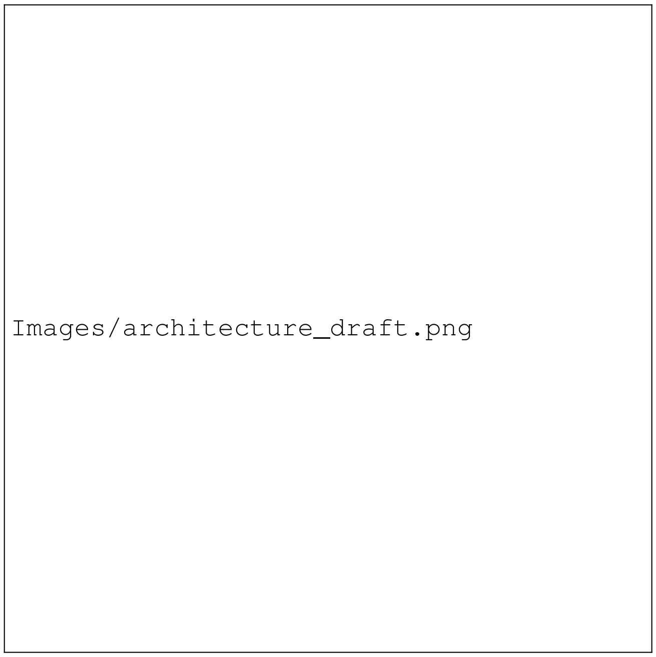

Figure 2. First draft of our proposed simplified architecture illustrating the core modules used where N is the length of extracted features sequence.

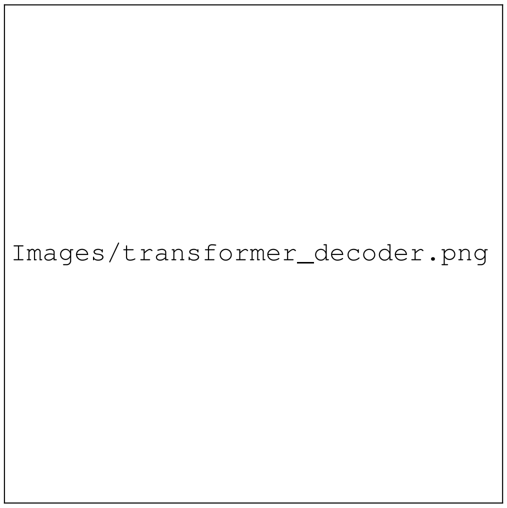

Figure 3. Overview of the joint utterance decoding and emotion classification architecture. Textual and audio features extracted from RoBERTa and WavLM are merged and pooled to create utterance embeddings. A Transformer-based utterance decoder predicts the next utterance representation, which is then classified into emotion categories using an MLP. Training minimizes a weighted sum of cross-entropy loss and cosine similarity loss.

4.5. Decoder Transformer & MLP

TODO

Our second architecture consists of two components: an utterance decoder and an emotion classifier, trained jointly (Figure 4). Each utterance is represented as a sequence of feature vectors (dimension 1536) extracted with RoBERTa and WavLM. A self-attention pooling first compresses each utterance into a single embedding. The resulting dialogue representation is passed to a transformer decoder (utterance decoder) that predicts the embedding of the next utterance at each timestep. To maintain causality, future utterances are masked during decoding, and learnable positional embeddings are added to the input sequence.

The predicted embeddings are used for two tasks. First, we minimize a cosine similarity loss ($\mathcal{L}_{cosine}$) against the pooled ground-truth embeddings of the next utterance. Second, the predicted embeddings are input into a multi-layer perceptron (MLP) classifier. This classifier contains layer normalization, a hidden GELU layer, dropout and a final output layer to predict emotion labels using cross-entropy loss ($\mathcal{L}_{CE}$) with class weighting and label smoothing. The total training objective is a weighted sum of the classification loss and the cosine similarity loss:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{cosine}\mathcal{L}_{cosine}$$

where λ_{cosine} controls the contribution of the similarity loss. By jointly modeling next-utterance prediction and emotion recognition, the architecture captures both semantic continuity and emotional dynamics across dialogues.

4.6. Encoder Transformer & MLP

Our third architecture consists of the following components:

4.6.1. Input Representation

Each utterance is embedded as a fixed-size vector of 1536 dimensions which is the concatenation of WaveLM features $x_1 \in R^{768}$ for the audio modality and RoBERTa features $x_2 \in R^{768}$. Due to the conversational context, a sliding window of adjacent utterances can be aggregated; the window with the best results was 2.

4.6.2. Transformer Encoder

The core of the architecture is a single-layer Transformer Encoder. The encoder processes the 1536-dimensional input using multi-head self-attention (8 heads) and a feed-forward network with an intermediate dimensionality of 1048. Dropout regularization with a rate of 0.45 is applied both within the self-attention mechanism and the feed-forward sublayers to mitigate overfitting.

Formally, for an input feature vector $x \in R^{1536}$, the Transformer Encoder computes:

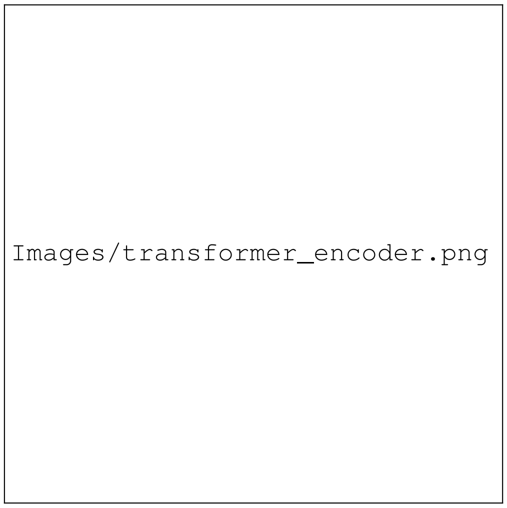

Figure 4. Overview of the joint utterance encoding and emotion classification architecture. Encoder transformer finds a representation of the utterance multimodal features and later classifies it using a fully connected layer. Training minimizes a weighted sum of cross-entropy loss.

$$x' = TransformerEncoder(x)$$

where self-attention enables the model to reweigh and recontextualize different feature dimensions relative to each other.

4.6.3. Output Layer

The output of the Transformer Encoder is passed through a fully connected linear layer that maps the representation to the target emotion space:

$$\hat{y} = Linear(x') \in R^7$$

where 7 corresponds to the number of emotion classes: *anger, disgust, fear, joy, neutral, sadness*, and *surprise*.

4.6.4. Loss Function

Training is performed using a weighted cross-entropy loss with label smoothing (smoothing factor = 0.1). Class imbalance is addressed by assigning inverse-frequency weights to each emotion class based on their occurrence in the training data, ensuring that majority classes (i.e. Neutral) are adequately penalized during optimization.

TODO

5. Experiments

For our experiments, we utilize pre-trained wavLM and RoBERTa models to extract 768-dimensional features, then use serval different architectures to predict next emotion. architecture 1 architecture 2 architecture 3

TODO:

Experiments: Comparisons to baselines using metrics (explain what the metrics are), with visualizations in terms of graphs, tables etc. (if you can show more, e.g., qualitative results with images - the better).

6. Optional sections if needed

6.1. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure **??**. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX). Always place two-column figures at the top or bottom of the page.

6.2. Algorithms

If you are using LaTeX, please use the “algorithm” and “algorithmic” environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

6.3. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 1. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Algorithm 1 Bubble Sort

```
Input: data $x_i$, size $m$
repeat
  Initialize $noChange = true$.
  for $i = 1$ to $m - 1$ do
    if $x_i > x_{i+1}$ then
      Swap $x_i$ and $x_{i+1}$
      $noChange = false$
    end if
  end for
until $noChange$ is $true$
```

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

Data set	Naive	Flexible	Better?
Breast	95.9± 0.2	96.7± 0.2	√
Cleveland	83.3± 0.6	80.0± 0.6	×
Glass2	61.9± 1.4	83.8± 0.7	√
Credit	74.8± 0.5	78.3± 0.6	
Horse	73.3± 0.9	69.7± 1.0	×
Meta	67.1± 0.6	76.5± 0.5	√
Pima	75.1± 0.6	73.9± 0.5	
Vehicle	44.9± 0.6	61.5± 0.4	√

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

References

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.