# Contrastive Learning for Conversational Emotion Recognition Using Knowledge Enhancement of Large Language Models

Andrew L. Mackey, B. Israel Cuevas and Susan Gauch
*Computer Science and Engineering, University of Arkansas, Fayetteville, Arkansas, U.S.A.*

Keywords: Emotion Analysis, Language Models, Natural Language Processing.

Abstract: Emotion recognition in conversation (ERC) is the task of classifying the emotion of each utterance in a conversation while learning the underlying latent representations. However, the representations for utterances are challenging to produce effectively given semantic and contextual information in the conversation. Large Language Models (LLMs) have demonstrated performance in various forms of emotion classification, including in zero-shot and few-shot settings, but their usage may be curtailed in some settings, particularly in limited resource environments. In this work, we propose a contrastive learning framework for the ERC task that leverages emotional anchors with semantic information encoded from an LLM to facilitate the learning of representations using a lightweight pretrained langauge model (PLM). Experimental results on benchmark ERC datasets demonstrate the effectiveness of our approach to baseline models while simultaneously reducing the inference cost of LLMs.

## 1 INTRODUCTION

Emotion recognition in conversation (ERC) is an active research area in the natural language processing (NLP) community that is concerned with the classification of utterances in a conversation. Unlike the traditional task of classifying a document (i.e. social media post) as being one emotion from a discrete set of possible emotions (i.e. happy, sad, etc.), the ERC task involves conversations where the dynamic interactions create changes between the context, speakers, and dialogue. As demonstrated in Figure 1, the emotion for each state of the conversation can easily shift depending on the state of the dialogue, speaker, utterance context, etc.

In recent years, contrastive learning and knowledge enhancement techniques have demonstrated success as frameworks for representation learning and deep contextual information, respectively. Several approaches have leveraged contrastive learning to learn the latent representations whereby closely or semantically-related representations are pulled closer to one another while pushing dissimilar representations further apart in the latent space. Knowledge enhancements techniques allow for the transfer of knowledge from significantly larger teacher models to smaller models to improve or enhance inputs using techniques such as semantic augmentation, input restructuring, or semantic augmentation. This is particularly advantageous when you require the deployment of models in a resource-constrained environment.

In this paper, we investigate a supervised contrastive learning framework combined with knowledge enhancement techniques for the ERC task on class-imbalanced data. We utilize a pretrained language model with a contrastive learning framework that leverages semantically-enhanced emotion label anchors extracted from an LLM to guide the contextual representations during training. Our study investigates the impact of combining knowledge enhancement with contrastive learning to the ERC task.

## 2 BACKGROUND INFORMATION

The primary approaches for the ERC task in recent times coalesce around sequence-based, graph-based, and knowledge-enhanced methodologies. DialogRNN modeled temporal dynamics and dependencies of dialogue by using RNNs (Majumder et al., 2019). DialogCRN introduced a contextual recurrent network that modeled the dialogue history and temporal dependencies for emotion recognition by utilizing cognitive factors (Hu et al., 2021). DialogGCN is a graph neural network-based approach to the ERC task that uses nodes to model the utterances (Ghosal



Mackey, A. L., Cuevas, B. I. and Gauch, S.
Contrastive Learning for Conversational Emotion Recognition Using Knowledge Enhancement of Large Language Models.
DOI: 10.5220/0013720100004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 330-336
ISBN: 978-989-758-769-6; ISSN: 2184-3228

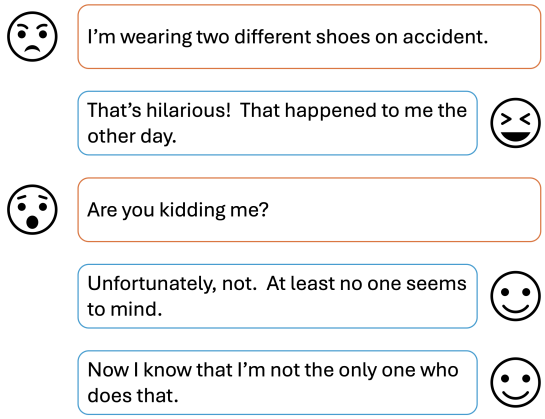


Figure 1: Example conversation and emotion recognition from utterances.

et al., 2019). DAG-ERC uses a directed acyclic graph (DAG) to model the structure within a conversation (Shen et al., 2021). The Knowledge-Enriched Transformer (KET) was proposed as a transformer model that enhanced emotion detection by leveraging external knowledge with the transformer architecture (Zhong et al., 2019).

Contrastive learning has demonstrated success in the field of natural language processing with respect to self-supervised learning frameworks. In contrastive learning, the primary objective is to learn representations where we can distinguish similar and dissimilar samples from one another. The process involves the construction of positive and negative examples and pairings, where positive pairs must share some type of similarity and negative pairs have some differences. Models aim to move positive examples closer to one another by positioning them closer some anchor in the latent embedding space while pushing apart the anchor for dissimilar examples and making them farther apart (Khosla et al., 2020). Prior work as also demonstrated that it is possible to extract useful representations from high-dimensional data in some latent space through Contrastive Predictive Coding (van den Oord et al., 2019).

SimCLR proposed a simple framework for learning visual representations by leveraging a contrastive loss by investigating data augmentations, learnable nonlinear transformations, and the benefits of contrastive learning from varying sizes of batch sizes and training steps (Chen et al., 2020). SimCSE presented a simple contrastive framework that used dropout as a data augmentation approach to advanced sentence embeddings (Gao et al., 2021). With Supervised Prototypical Contrastive Learning (SPCL), the authors leveraged a contrastive learning loss for the ERC task on class-imbalanced data while combining it with curriculum learning (Song et al., 2022). Emotion-Anchored Contrastive Learning (EACL) utilized textual emotion labels that were used to generate emotion anchor representations (Yu et al., 2024).

Several techniques have been proposed to improve or enhance smaller models from larger models, such as knowledge distillation and knowledge enhancement techniques. Knowledge distillation has been demonstrated as a model compression technique where information from a teacher model is transferred to a smaller student model that is more efficient (Bucila et al., 2006). The work presented in (Hinton et al., 2015) demonstrated an ability for the complex model to transfer not just the final predictions, but also on the soft targets that were produced by the teacher model to facilitate the student model learning nuanced knowledge. In NLP, DistilBERT represents a PLM that is a distilled version of BERT which reduces the model's size by 40%, being 60% faster, and retains 97% of its language understanding capabilities (Sanh et al., 2020).

Knowledge enhancement improves the understanding of text inputs by providing additional context from domain-specific sources, tagging from lexicons, restructuring the inputs, etc. In (Qu et al., 2019), the authors enhanced a BERT-based model through a history answer embedding where prior knowledge was necessary in conversational settings. The authors in (Zhang et al., 2019) incorporated the use of knowledge graphs to improve a BERT-based model by providing structured knowledge facts from external sources.

# 3 METHODOLOGY

## 3.1 Definition

Each of the datasets evaluated in this work consists of the following: conversations, speakers, and emotions. The set of conversations $\mathcal{C}$ consists of utterances and emotion labels for each conversation turn. We represent a single conversation $c \in \mathcal{C}$ as a collection of utterances and speakers $c = [(s_1,u_1),(s_2,u_2),...,(s_N,u_N)]$, where $s_i \in \mathcal{S}$ refers to the speaker, $u_i$ is the utterance for the $i^{\text{th}}$ turn, and $\mathcal{S}$ is the set of speakers. We define $\mathcal{E}$ as the set of emotion labels where $\mathcal{E} = \{e_1,e_2,...,e_k\}$ for the corresponding dataset.

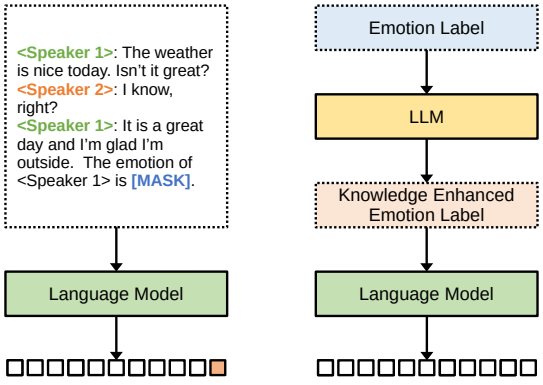


Figure 2: Emotion frequency for the labels in the MELD dataset.

## 3.2 Model Overview

The proposed model for this work features a pretrained language model for learning the representations of the utterances, a knowledge enhancement approach to extract information from large language models to improve ERC task performance of PLMs, the incorporation of semantically-enhanced emotion anchors, and a contrastive learning framework that utilizes these enhanced emotion anchors. In the sections that follow, we will define and outline the purpose of each of these components of our proposed model.

Table 1: Frequency metrics for the IEMOCAP and MELD datasets by the number of utterances, dialogues, and label classes.

| | **IEMOCAP** | | **MELD** | |
|---|---|---|---|---|
| | Uttr. | Dia. | Uttr. | Dia. |
| Train | 4,810 | 100 | 9,989 | 1,038 |
| Val | 1,000 | 20 | 1,109 | 114 |
| Test | 1,523 | 31 | 2,610 | 280 |
| Total | 7,333 | 151 | 13,708 | 1,432 |
| Classes | 6 | | 7 | |

## 3.3 Context Encoding

We adopt a contrastive learning framework with emotion anchors by utilizing pretrained language models along with a prompt-based approach to implement masked language modeling following previous work (Song et al., 2022). Our prompt-based contextual representations are formed at utterance time $t$ by using turns $(s,u) \in \{(s_j,u_j) \mid t-k \leq j \leq t\}$ where $k$ represents the window length of most recent turns. The emotion for utterance $u_t$ is predicted by using the following prompt:

$$c_t = [s_{t-k}, u_{t-k}, ..., s_t, u_t, p_t] \quad (1)$$

$$p_t = \text{"For } u_t, s_t \text{ feels } \langle\text{MASK}\rangle\text{"}. \quad (2)$$

The last hidden state of the $\langle$MASK$\rangle$ token as the representation for the utterance. The model attends to the target sentence when training in this manner so that it is able to produce usable representations.

## 3.4 Contrastive Learning and Knowledge Enhancement

Our model leverages a supervised contrastive learning framework that utilizes both learned contextual representations and semantically-enhanced emotion anchors from an LLM. A batch of $N$ conversation examples $\mathbf{X} = \{x_1, x_2, ..., x_N\}$ where $\mathbf{X} \in \mathbb{R}^{n \times \ell}$ where $n$ represents the batch size and $\ell$ is the maximum length of the input. The last hidden state of the input of the language model is obtained in:

$$\mathbf{Z} = \text{PLM}(\mathbf{X}) \quad (3)$$

We use the hidden state of the $\langle$MASK$\rangle$ token $\mathbf{Z}_{\langle\text{MASK}\rangle}$ and feed this into a multilayer perceptron (MLP) network to obtain the representations for the utterances:

$$\mathbf{R} = \text{MLP}_{\text{CL}}(\mathbf{Z}_{\langle\text{MASK}\rangle}) \quad (4)$$

Prior work leveraged anchors where PLMs were used to encode the emotion labels (Yu et al., 2024). In our approach, we leverage LLMs to expand the semantic and contextual representations of each emotion label in the set of emotions **emo** to form an enhanced emotion label representation **emo**$'$. We obtain the embedding representations from the LLM for **emo**$'$ and use an MLP network to obtain a set of parameterized representations for our model as $\mathbf{R}'$:

$$\mathbf{emo}' = \text{LLM}_1(\mathbf{emo}) \quad (5)$$

$$\mathbf{R}' = \text{MLP}(\text{LLM}_2(\mathbf{emo}')) \quad (6)$$

We let $\text{sim}(\mathbf{z}_i, \mathbf{z}_k)$ be some similarity function for inputs $\mathbf{z}_i$ and $\mathbf{z}_k$, where the use of cosine similarity employed for this task. For the given batch representations $\mathbf{R}$ along with semantically-enhanced and encoded emotion labels $\mathbf{R}'$, we combine the representations together to form $T = \mathbf{R} \cup \mathbf{R}'$ for use with the

contrastive learning loss that leverages the anchors to improve the alignment of representations. We define $\text{Pos}(i)$ to return all members in the representations $T$ with the same emotion as member $i$. We define $\tau$ to represent a temperature hyperparameter for the loss function.

$$f(\mathbf{z}_i, \mathbf{z}_k) = \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau) \tag{7}$$

$$\mathcal{L}_{\text{CL}} = \sum_{i=1}^{n+|emo|} -\log \frac{\sum_{r_j \in \text{Pos}(i)} f(r_i, r_j)}{|\text{Pos}(i)| \sum_{r_j \in T} f(r_i, r_j)} \tag{8}$$

The effects from the $\mathcal{L}_{\text{CL}}$ function can be observed in the movements of related representations becoming nearer and unrelated representations becoming more distant. In addition, the anchors serve as a guide when learning the representations for the utterances while also learning how to increase the distance between emotion anchor representations. A cross entropy loss is combined with the supervised contrastive loss to improve the model's discriminative capabilities:

$$\hat{\mathbf{y}} = \text{Softmax}\big(\text{MLP}_{\text{CE}}(\mathbf{Z}_{\langle \text{MASK} \rangle})\big) \tag{9}$$

$$\mathcal{L}_{\text{CE}} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{|\text{emo}|} y_{ik} \log(\hat{y}_{ik}) \tag{10}$$

The final loss function uses the $\lambda$ hyperparameter to serve as a weighted average between the supervised contrastive loss and the cross entropy loss functions.

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{CE}} + (1 - \lambda) \cdot \mathcal{L}_{\text{CL}} \tag{11}$$

# 4 EXPERIMENTAL DESIGN

## 4.1 Setup

The language models used for experiments include BERT, RoBERTa, and ModernBERT from the HuggingFace Transformers library. The PyTorch framework was used on a single NVIDIA A6000 GPU. The OpenAI GPT-4o LLM was used for knowledge enhancement tasks. We use the AdamW optimizer, a dropout rate of 0.1, maximum length of 512, temperature $\tau = 0.1$, and a learning rate of $1e^{-5}$.

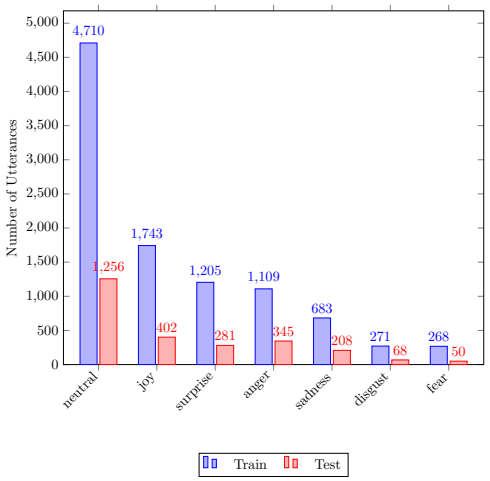


Figure 3: Emotion frequency for the labels in the MELD dataset.

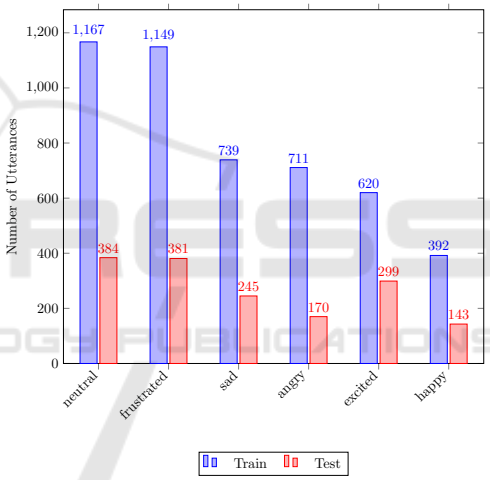


Figure 4: Emotion frequency for the labels in the IEMOCAP dataset.

## 4.2 Datasets

Experiments are conducted on two major benchmark datasets: MELD and IEMOCAP (Poria et al., 2019) (Busso et al., 2008).

**MELD.** The MELD (Multimodal EmotionLines Dataset) is a multimodal emotion recognition dataset that contains utterances and conversations extracted from the TV show Friends (Poria et al., 2019). Each utterance contains an emotion label from one of the following: surprise, anger, neutral, sadness, disgusting, joy, and fear. The emotion distribution between the training and testing sets can be found in Figure 3.

Table 2: Comparison of Weighted F1 Average Metric for IEMOCAP and MELD Datasets. The bold font indicates the best performance.

| **Baseline Models** | **IEMOCAP** | **MELD** |
|---|---|---|
| BERT (Devlin et al., 2019) | 64.87 | 63.45 |
| RoBERTa (Liu et al., 2019) | 63.98 | 64.62 |
| ModernBERT (Warner et al., 2024) | 66.11 | 61.80 |
| ChatGPT 3-shot (Zhao et al., 2023) | 48.58 | 58.35 |
| **Experimental Models** | **IEMOCAP** | **MELD** |
| BERT+ECL | 65.28 | 64.91 |
| RoBERTa+ECL | 67.72 | 66.31 |
| ModernBERT+ECL | **71.25** | **65.67** |

**IEMOCAP.** The IEMOCAP dataset consists of 151 videos of two speakers per session. These clips are spread across five sessions per actor and include both scripted and improvised dialogues (Busso et al., 2008). The dataset is multimodal, providing video recordings of the actors' facial expressions and body language. Each segment is annotated for the presence of the following emotions: excited, frustrated, neutral, sad, happy, and angry. The emotion distribution between the training and testing sets can be found in Figure 4.

### 4.3 Metrics

Due to the imbalance that exists between the different target classes as seen in (Lee and Lee, 2022), (Yu et al., 2024), and (Song et al., 2022), we report the results using the weighted F1 score in the sections that follow.

## 5 RESULTS

The results for our proposed methods and baseline experiments are reported in Table 2. The mean weighted F1 score is reported after $n = 5$ successive runs of each experiment. As demonstrated in the results, we observed that our experimental models outperform the baseline pretrained langauge models on both the IEMOCAP and MELD datasets. For the BERT models, we observe a difference of $\Delta = 0.41$ and $\Delta = 1.46$ for the IEMOCAP and MELD datasets, respectively. For the RoBERTA models, we observe a difference of $\Delta = 3.74$ and $\Delta = 1.69$ for the IEMOCAP and MELD datasets, respectively. For the ModernBERT models, we observe a difference of $\Delta = 5.14$ and $\Delta = 3.87$, respectively.

We observe that the choice of pretrained language model with the proposed constrastive learning framework affects the overall performance, but all evaluated pretrained language models demonstrate an improvement in performance when combined with the contrastive learning framework. The mean performance gain across all datasets and PLMs is $\bar{\Delta}_{\text{ALL}} = 2.718$ ($s_{\Delta} = 1.800$) where the performance gain for the IEMOCAP dataset is $\bar{\Delta}_{\text{IEMO}} = 3.097$ ($s_{\Delta} = 2.430$) and for the MELD dataset is $\bar{\Delta}_{\text{MELD}} = 2.34$ ($s_{\Delta} = 1.33$).

In comparison to the model proposed by (Zhao et al., 2023), we observe an improvement from our best performing model ModernBERT+ECL over the previous work by a large margin of $\Delta = 22.67$ and $\Delta = 7.32$ for the IEMOCAP and MELD datasets, respectively. This may be due to the limitations of the original experiment under a few shot prompt approach and further exploration is needed to understand whether more recent version of the models demonstrate improved performance for the ERC task.

## 6 CONCLUSION

In this paper, we presented an approach that combined a contrastive learning framework with knowledge enhancement from large language models to improve representation learning for the ERC classification task. Our experiments demonstrated that using LLM-generated anchors as guidance led to transferable representations that could be leveraged in a resource-constrained environment. We also demonstrate that the proposed framework would be effective across different pretrained language models.

Our findings suggest that LLMs can be leveraged as a source to extract the semantic representations for emotion labels that can be used in a contrastive learning framework. Future work can explore further refinements in the knowledge enhancement process to improve the label imbalances to provide the PLMs with additional context to improve underperforming

label classification. Furthermore, contrastive learning methods could potentially be expanded by exploring the relationship of hard negative selection based on emotion relationships. Lastly, multimodal data could be incorporated to further enhance the model's performance.

## REFERENCES


Bucila, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535–541, New York, NY, USA. Association for Computing Machinery.

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

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Gao, T., Yao, X., and Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. (2019). Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network.

Hu, D., Wei, L., and Huai, X. (2021). DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052, Online. Association for Computational Linguistics.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Lee, J. and Lee, W. (2022). CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5669–5679, Seattle, United States. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. (2019). Dialoguernn: an attentive rnn for emotion detection in conversations. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Qu, C., Yang, L., Qiu, M., Croft, W. B., Zhang, Y., and Iyyer, M. (2019). Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1133–1136, New York, NY, USA. Association for Computing Machinery.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Shen, W., Wu, S., Yang, Y., and Quan, X. (2021). Directed acyclic graph network for conversational emotion recognition. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560, Online. Association for Computational Linguistics.

Song, X., Huang, L., Xue, H., and Hu, S. (2022). Supervised prototypical contrastive learning for emotion recognition in conversation. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

van den Oord, A., Li, Y., and Vinyals, O. (2019). Representation learning with contrastive predictive coding.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., and Poli, I. (2024). Smarter, better, faster, longer:

A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Yu, F., Guo, J., Wu, Z., and Dai, X. (2024). Emotion-anchored contrastive learning framework for emotion recognition in conversation. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4521–4534, Mexico City, Mexico. Association for Computational Linguistics.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zhao, W., Zhao, Y., Lu, X., Wang, S., Tong, Y., and Qin, B. (2023). Is chatgpt equipped with emotional dialogue capabilities?

Zhong, P., Wang, D., and Miao, C. (2019). Knowledge-enriched transformer for emotion detection in textual conversations. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.