INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Fine-tuned RoBERTa Model with a CNN-LSTM Network for Conversational Emotion Recognition

Jiachen Luo[1], *Huy Phan*[2*], *Joshua Reiss*[1]

[1]Centre for Digital Music, Queen Mary University of London, UK
[2]Amazon Alexa, Cambridge, MA, USA

jiachen.luo@qmul.ac.uk, huypq@amazon.co.uk, joshua.reiss@qmul.ac.uk

Abstract

Textual emotion recognition in conversations has gained increasing attention in recent years for the growing amount of applications it can serve, e.g., human-robot interactions, recommended systems. However, most existing approaches are either based on BERT-based model which fail to exploit crucial information about the long-text context, or resort to complex entanglement of neural network architectures resulting in less stable training procedures and slower inference time. To bridge this gap, we first propose a fast, compact and parameter-efficient framework based on fine-tuned pre-trained RoBERTa model with a CNN-LSTM network for textual emotion recognition in conversations. First, we fine-tune the pre-tranined RoBERTa model to effectively learn long-term emotion-relevant context information. Second, convolutional neural network coupled with the bidirectional long short-term memory and joint reinforced blocks are utilized to recognize emotion in conversations. Extensive experiments are conducted on benchmark emotion MELD dataset, and the results show that our model outperforms a wide range of strong baselines and achieves competitive results with the state-of-art approaches.

Index Terms:conversational emotion recognition, transfer learning, RoBERTa, CNN LSTM network

1. Introduction

Emotions play a critical role in shaping individual's decision-making process and affects a series of subsequent life choices [1]. Humans convey emotions through various modalities containing the speech, the facial expression, body postures, and etc [2]. With the development of social platforms such as Whatsapp, there is an emerging need to machines to understand human textual emotion in conversations. A growing number of researchers have conducted emotion recognition in conversations based on the text with the hope to help the individuals navigate their decision-making process [3, 4]. Textual emotion recognition in conversations (TERC) aims to the emotion of each utterance from the transcript of a conversation [5]. TERC has been regarded as one of the most challenging and essential tasks in artificial intelligence.

However, there are several challenges when analyzing textual emotion in conversations [6]. Emotion expression in dialogues is easily influenced by surrounding conversational context. Contextual information is the main difference between dialogue emotion analysis and single sentence emotion analysis.

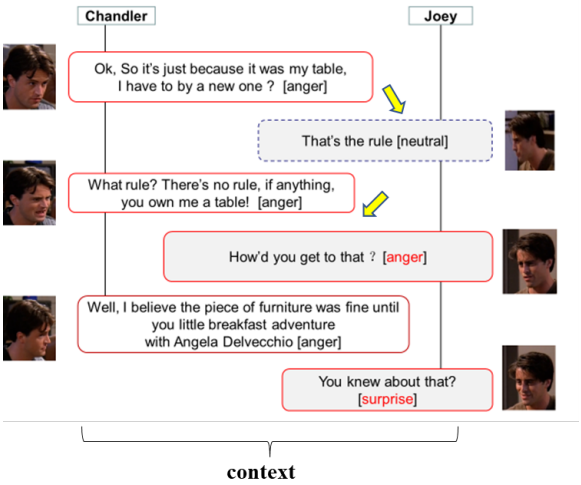

Figure 1: *An example of the context dynamics in dialogue system in MELD dataset.*

It sometimes enhances, weakens, or reverses the raw emotion of utterance. Humans often rely on the context in a dialogue to express emotions [7], which is difficult to be understood by machines. Figure 1 shows an example conversation demonstrates the importance of the context in understanding conversations and detecting implicit emotions. Thus, they are inseparable and cannot be treated independently. There can also be jumps in the multiple speakers' logical flow in the course of interaction.

Textual emotion recognition in a conversation mainly consists of three steps: obtaining context information for an utterance, extracting the influence of the context information for an utterance, and learning emotional features [2, 6, 8, 9]. Existing dialogue emotion analysis like CMN [6] and DialogueRNN [9], employ complicated deep neural network structures to capture context information and model the influence of context information for an utterance. However, designing compact structure to better predict textual emotions is a non-trivial task that researchers often have to face.

The availability of vast volumes of data enables a deep learning techniques to explore better contextual representation [10, 11, 12, 13, 14, 15, 16]. In the last few years, with the developments of transformer-based architectures BERT such as RoBERTa [17] and DistillBERT [18] for question answering [19], text summarization [20] and other NLP tasks [21]. Unfortunately, the variant BERT model suffers from computational complexities, fixed input length size limitations, and wordpiece embedding problems, and computational complexities [14]. Moreover, compared to the recently published works

*The work was done when H. Phan was at the Centre for Digital Music (C4DM), Queen Mary University of London, UK and prior to joining Amazon.

Thanks to the China Scholarship Council and Queen Mary University of London for funding.

10.21437/Interspeech.2023-463

on TERC [15, 16], they did not design a fast, simple, parameter-efficient architecture to model complex dialogue context cues in different environments with arbitrary turns, local correlations and global contextual information.

In this paper, we convert the previous three-step task into a two-step task. Meanwhile, the compact structure reduces the computational cost. To the best of our knowledge, our proposed approach is first to leverage pre-trained RoBERTa models in conjunction with CNN-LSTM network to predict emotion in a simple, fast, parameter-efficient way. We fine-tune pre-trained RoBERTa to produce a long-term context-sensitive dependency. Moreover, we design a compact CNN-LSTM network with residual block to capture local correlations and global contextual information, and learn emotion-relevant features for better prediction. The main contributions of this paper are as follows:

• We utilized fine-tuned pre-trained RoBERTa models to capture contextual information by dynamically changing the masking pattern to support better understanding of utterances for TERC.

• We designed compact CNN-LSTM network with residual block to model detailed local features and global context dependency in a simple, and parameter-efficient way.

2. Related Works

Textual emotion recognition is a long-standing research topic. Proposed emotion classifiers in conversations include feature-based methods [22], recursive networks [23], convolution neural networks and recurrent neural networks [24]. A substantial number of machine learning approaches are employed to discern the emotions in conversations [25, 26, 27, 28, 29, 30]. These approaches mainly focus on extracting a set of features and training a category emotion polarity classifier. The features of machine learning mainly include n-grams [25], lexicon features [26] and word co-occurrence frequencies [27]. The Maximum Entropy [28], Support vector Machine [29], and Fuzzy Lattice Reasoning methods [30] have all been employed as classifier to predict emotions in conversations. However, these models treated text independently and thus did not obtain the inter-dependence of utterances in dialogues.

To capture the contextual information of utterances, the RNN architecture is a standard way to exploit the sequential relationship. Poria *et al.* used context-dependent LSTM networks to extract contextual information for modeling emotional dynamics [24]. Hazarika *et al.* improved bidirectional contextual LSTM (bc-LSTM), a conversational memory network to capture the self and inter-speaker emotional influence, and the attention mechanism was used to prioritize important utterances [22].

Recently, there exists growing attention over the development of new approaches – the "pre-train and finetune" paradigm has become increasingly popular. Pre-trained language models, such as BERT [9] and XLNet [10], achieve high performance in various tasks (e.g., TERC) by constructing contextualized representation. Pretraining on large-scale unsupervised texts enable these models to learn automatic representation. Prior studies based on BERT simply utilized BERT for textual feature extraction, although BERT is a significantly undertrained model and could be improved. Pre-trained embedding are in high-dimension space, where it is unclear how the leaned feature vectors contribute to the identification of emotion. At the same time, several optimization models based on BERT have also been proposed, including RoBERTa [17] and ALBERT [11], but they are not well explored in the task of utterance-level emotion recognition in dialogue systems. Our work differs from the work mentioned above. We propose to learn emotion-relevant contextual sensitive information via fine-tuned pre-trained RoBERTa. Additionally, our proposed approach adopts a compact and parameter-efficient CNN-LSTM network architecture with residual block to learn hierarchical local and global features to recognize speech emotion.

3. Methodology

Our proposed approach consists of three major modules: A) fine-tuned pre-trained RoBERTa embedding, B) CNN-LSTM network, and C) residual blocks that aims to capture both long-term dependency and local information (see Figure 2).

3.1. Finetuned Pre-trained RoBERTa Encoder

We use the RoBERTa model to produce context independent utterance-level feature vectors. We fine-tune the RoBERTa (Large model) (Uncased: hidden-1024, heads-24, layer-24) for pre-defined emotion prediction from the transcript of the utterances [17]. Let us define an utterance u contained a sequence of Byte-Pair-Encoding (BPE) tokenized tokens $u_1, u_2, \ldots, u_N$, with emotion label Ex. In this setting, the fine-tuned pretrained RoBERTa model is achieved through a sentence classification task. A special token [CLS] is appended at the beginning of the utterance to obtain the input sequence of the model: [CLS], $u_1, u_2, \ldots, u_N$.

The RoBERTa encoder received the transcript of the utterance, passed the [CLS]-appended BPE tokenized utterances to produce activations from the final four layers corresponding to the [CLS] token. Then, these four were concatenated to obtain the context-independent utterance feature vector with a dimension of 4096 (see Figure 2).

3.2. 1D CNN-LSTM Network

We first used a fully connected layer (FC) to reduce the dimension of the RoBERTa embedding. The output features dimension after FC is 1024. Then, two parallel layers of Bi-LSTM and 1D CNN-Bi-LSTM were applied on the output of the embedding layer to learn the long-term contextual dependency and the local information in both forward and backward directions. Additionally, the two stacked Bi-LSTM layers contained 256 units to store long-term context influence. The 1D CNN-Bi-LSTM comprised a convolution operation (kernel size 3, stride 2, and 256 filters) followed by batch normalization, ReLU activation and Bi-LSTM of 256 hidden units. The convolutional layer acted as a projection layer where it mapped the output feature dimension from 1024 to 256. At this point, we combined the two outputs from the Bi-LSTM and 1D CNN-Bi-LSTM using simple concatenation and fed them into residual blocks.

3.3. Residual Blocks

Inspired by ultra-deep CNNs for image recognition, our output was based on the combination of the pre-trained RoBERTa embedding and the concatenated network output mentioned above (Figure 2). In particular, residual connections connected the RoBERTa embedding with two FC layers. The residual block added more feature details after passing each learning layer for training the deep networks more effectively. Finally, the contextual utterance embedding was sent to stacked FC layers to determine the corresponding emotion.

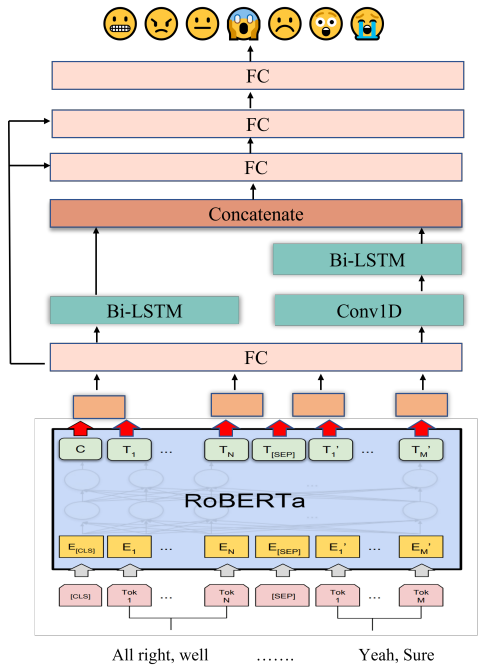

Figure 2: *The overall architecture of the proposed model for textual emotion prediction in conversations.*

3.4. Training

We chose categorical cross-entropy as the measure of classification loss and used L2-regularization to relieve overfitting.

4. Experiments and Results

4.1. Database and Metrics

We evaluated our model on the Multi-modal EmotionLines Dataset (MELD) [24]. MELD is a multi-modal dataset with aligned acoustical, textual, and visual information of each utterance in a conversation. We considered only the textual modality. Statistics of the dataset are presented in Table 1.

MELD contains more than 13,000 utterances, and 1400 conversations from the TV series "Friends" [24]. Multiple speakers participated in the dialogues. MELD datasets is popular than other becnhmark datasets, and it is close to real-world. The whole dataset is split into three parts: training (9989), validation (1109), and test (2610). Each utterance is annotated with one of seven emotion labels: anger, disgust, sadness, joy, neutral, surprise and fear. Following previous works [12], we used weighted average F1-score as the evaluation metric.

4.2. Compared Methods

We compared the performance of our approach against the performance of the state-of-the-art and baselines on the MELD dataset for textual modality (Table 2). Totally, these approaches employ pre-trained BERT-based embedding for speech emotion recognition tasks. bc-LSTM leverages an utterance-level LSTM to perform context dependent emotion analysis [24], while others are mainly to recognize conversational emotions by graph neural network architecture [31, 32, 34].

Table 1: *Statistics of MELD dataset.*

Dataset	Num. of dialogues			Num. of utterances		
	train	val	test	train	val	test
MELD	1039	114	280	9989	1109	2610

4.3. Textual Features

We used the fine-tuned RoBERTa model to extract context independent representation. RoBERTa is a transformer model pretrained on a large corpus of English data in a self-supervised fashion [17]. The RoBERTa model is optimized on the original BERT by training with dynamic masks, FULL-SENTENCES without Next Sentence Prediction loss, large mini-batches, and a large byte-pair encoding. We fine-tuned the RoBERTa model for conversational emotion recognition from the transcripts of the utterances. The RoBERTa comprised 24-layered transformer encoder blocks with each block containing 24-head self-attention layers and 1024 hidden layers. An utterance was first passed through the pre-trained RoBERTa model to extract activations from the final four layers. Then these four were concatenated to obtain the context independent utterance feature vector with a dimension of 4096.

4.4. Implementation Details

We implemented all the models with the Pytorch 1.11.0 framework. Batch size was set to 32, and dropout rate was 0.3. We used Adam optimizer with an initial learning rate of 0.0001.

5. Results and Discussion

5.1. Overall Results and Discussions

As shown in Table 2, our proposed approach exceed a range of baselines and state-of-the-art models. Overall results of our proposed model and the previous state-of-the-art results are presented in Table 2. The performance of our method reached F1-score of 66.12%. Ablation study (see Table 2) verifies the fine-tuned pre-trained RoBERTa made much effort in capturing long-range contextual representation for better inferring emotions than isolated word embeddings (e.g., n-grams [25], lexi-

Table 2: *Performance on the MELD dataset. The best results are in bold.*

Features	Model	F1 score
BERT	bc-LSTM [24]	60.45
BERT	DialogueRNN [31]	59.21
BERT	DialogueGCN [32]	60.13
BERT	MTL [33]	61.90
RoBERTa	bc-LSTM + att [34]	63.37
RoBERTa	DialogueRNN [31]	63.61
RoBERTa	DialogueGCN [32]	63.02
RoBERTa	UniMSE [36]	65.09
RoBERTa	GRASP [37]	65.06
RoBERTa	SGED [34]	63.34
RoBERTa	SGED + LSTM + att [34]	63.37
RoBERTa	SGED + DialogueGCN [34]	64.55
RoBERTa	SGED + DAG-ERC [34]	65.46
RoBERTa	EmoBERTa [5]	65.61
BERT	our model	62.15
RoBERTa	*w/o* 1D CNN-LSTM in our model	63.86
RoBERTa	*w/o* LSTM in our model	64.7
RoBERTa	*w/o* residual block in our model	63.52
RoBERTa	our model	**66.12**

Table 3: *The class-wise weighted F1-scores on the MELD dataset.*

Proposed method	neutral	surprise	fear	sadness	joy	disgust	anger	w-average F1
text modality	77.32	62.76	32.59	46.35	62.21	32.0	56.18	66.12

con features [26]). Besides, the model's performance was worse when the CNN-LSTM module or residual block was removed, which indicates that these components could effectively capture detailed local features, global context dependency, and extract rich emotion features to boost the overall performance. Moreover, our model with BERT features also outperformed other BERT baselines.

However, three factors make it considerably harder to model textual emotion recognition on the MELD dataset. First, the average length of MELD utterances is much shorter than other benchmark dataset in emotion recognition tasks (e.g., IEMOCAP [35]). Second, the average number of turns in MELD dataset is 10, and most of the participants only speak one or two utterances per dialogue. Finally, for a party-dependent model (DialogueRNN [31]), it is hard to capture intra- and inter-information between speakers. We found that the performances of the party-dependent model (DialogueRNN [31]) was slightly better than the party-ignorant model (bc-LSTM [24]) on the MELD dataset (see Table 2).

There exists a lot of short utterances in MELD dataset, contributing to less emotion-relevant information. For example, a single word "okay" is ambitious, and it can express different emotions: joy, neutral and anger, respectively. Strong emotions like joy and anger might be identified by punctuation marks (e.g., '!'). These symbols provide limited and low-quality information for inferring emotion categories in conversations. In our work, we achieved good performance in few emotion categories: joy and anger (Table 3). Firstly, fine-tuned RoBERTa by dynamically changing the mask pattern and training the model longer, with bigger batches, over more data is useful in limited and unbalanced MELD dataset as knowledge transfer should contribute to better contextual-level emotion representation. It not only helps to handle the shortage of data but also speeds up training and improves the performance of the prediction model. Secondly, we employ a compact structure 'CNN-LSTM network with residual block' to better learn local and global emotion-related contextual information in more parameter-efficient and less time consuming for training. The designed network, combining two parallel layers of bi-LSTM and 1D CNN-Bi-LSTM, can take advantage of the strength of both networks while overcoming their shortcomings. Textual information in conversations has time-varying properties and requires more sophisticated analysis to reflect them. The designed networks utilize the strength of CNN and LSTM to capture the emotion-relevant time dynamics. Specifically, a possible explanation is that introducing a convolution window may focus the learning on local contexts with neighboring word occurrence patterns captured by the LSTM, and thus result in the correct understanding of the global contextual information. Additionally, the residual module adds more emotion-relevant feature details after passing each learning layer for training the deep networks more effectively.

5.2. Case Studies

Figure 1 showed a conversation snippet with annotated labels from the MELD dataset classified by our proposed method. In this snippet, character Joey is initially in a neutral state while Chandler acts as a speaker in the beginning, Then, Chanlder changes her focus and question character Joey state. Chanlder makes Joey feel anger. This snippet reveals that the interaction of emotion between participants in conversations. Our proposed method showed good ability in detecting the emotion shifting from neutral to anger. The CNN-LSTM framework got better results since it was more sensitive to the context information and had a better context compositionality ability in general.

5.3. Error Analysis

By predicting emotions in conversations, we found that the model errors were mainly caused by the following aspects. Firstly, the MELD dataset is limited and unbalanced and most of published work performed poorly for minor classes (e.g., disgust, fear, and sadness) [5, 24, 31, 32, 33, 34]. Our systems being trained on an imbalanced dataset perform well on dominant emotions (i.e. neutral), and also achieve better on minor classes (surprise, fear, sadness, joy, disgust, and anger) than other published work [5, 24, 31, 32, 33, 34]. However, there are still room for improvement for under-represented ones. Data augmentation, transfer learning, and resampling techniques are potential strategies to explore. Secondly, human convey emotions through various modalities including speech, facial expression, body postures, etc. It is natural to introduce acoustic, visual and other modalities to help text information in short utterances (e.g., 'Oh'). Collating and combining information from multiple modalities in human communication to discern the perceived emotions is beneficial: different modalities can complement or augment each other to consolidate richer emotion-relevant information. Thirdly, similar emotion categories were often misclassified (e.g., anger).

6. Conclusion

We proposed a fine-tuned RoBERTa model along with CNN-LSTM network for tackling utterance-level emotion recognition in a dialogue system. As the core of the fine-tuning part, the pre-trained RoBERTa was designed to dynamically change mask pattern and train the model over more data. The CNN-LSTM network with residual block was capable of extracting log-term contextual dependency and modeling local information in multi-turns conversations. Our experimental results outperformed current state-of-the-art models on the MLED in most cases, which indicated that the proposed model had the potential to effectively use the target utterance's context to capture the emotional nature in conversations.

Despite the promising results, our proposed approach may still present some limitations in terms of the diversity of the benchmark datasets, complex cognition process, and the availability of more modality. Future work will focus on the following directions: how to differentiate emotions in more benchmark datasets, how to consider cognitive perspective, how to differentiate similar emotions, and how to incorporate multimodal information for TERC task.

7. References

[1] J.H. Lin, S.R. Pan, C.S. Lee, S. Oviatt, "An explainable deep fusion network for affect recognition using physiological signals," *In Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp.2069-2072.

[2] Y. Wang, W. Song, W. Tao, A. Liotta, D.W. Yang, X.L. Li, S.Y. Gao, Y.X. Sun, W.F. Ge, W. Zhang, W.Q. Zhang, "A systematic review on affective computing: emotion models, databases, and recent advances," *Information Fusion*, 2022, vol.83,pp.19-52.

[3] Y.T. Lan, W. Liu, B.L. Lu, "Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism," *International Joint Conference on Neural Networks (IJCNN)*, 2020, pp.1-12.

[4] Y.T. Lan, W. Liu, B.L. Lu, "Multi-label and multimodal classifier for affective states recognition in virtual rehabilitation," *IEEE Transactions on Affective Computing*, 2022, pp.1183-1194.

[5] T. Kim, P. Vossen, "EmoBERTa: speaker-aware emotion recognition in conversation with RoBERTa,"*arXiv:2108.12009*, 2021, pp.1-7.

[6] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.P. Morency, R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," *In Proceedings of the conference. Association for Computational Linguistics*, 2018, pp.1-23.

[7] J.W. Deng, F.J. Ren, "A survey of textual emotion recognition and its challenges,"*IEEE TRANSACTIONS ON JOURNAL*, 2021, pp.1-20.

[8] P. Pereira, H. Moniz, J.P. Carvallo, "Deep emotion recognition in textual conversations: a survey,"*arXiv:2211.09172*, 2022, 1-25.

[9] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, N. Onoe, "M2FNet: multi-modal fusion network for emotion recognition in conversation", *arXiv:2206.02187*, 2022, pp. 1-10.

[10] K. Yang, D. Lee, T. Whang, S. Lee, H. Lim, "EmotionX-KU: BERT-Max based contextual emotion classifier,"*arXiv:1906.11565*, 2019, pp. 1-6.

[11] Z.Y. Fu, W.C.S. Zhou, J.J. Xu, H. Zhou, L. Li, "Contextual representation learning beyond masked language modeling," *arXiv:2204.04163*, 2022, pp. 2701-2704.

[12] I. Lauriola, A. Lavelli, et al. "An introduction to deep learning in natural language processing: models, techniques, and tools," *Neurocomputing*, 2022, vol.470, pp.443-456.

[13] T. Tambe, C. Hooper, L. Pentecost, T.Y. Jia, E.Y. Yang, M. Donato, V. Sanh, P.N. Whatmough, A.M. Rush, D. Brooks, G.Y. Wei, "EdgeBERT: sentence-Level energy optimizations for latency-aware multi-task NLP inference," *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, 2021, pp.830-844.

[14] H.Q. Yang, J.P. Shen, "Emotion dynamics modeling via BERT," *International Joint Conference on Neural Networks (IJCNN)*, 2021, pp.1-8.

[15] V.S. Kodiyala, R.E. Mercer, "Emotion recognition and sentiment classification using BERT with data augmentation and emotion lexicon enrichment," *20th IEEE International Conference on Machine Learning and Applications*, 2021, pp.191-198.

[16] P. Kumar, B. Raman, "A BERT based dual-channel explainable text emotion recognition system,"*Neural Networks*, 2022, vol.150, pp.392-407.

[17] Y. Liu, M. Ott, N. Goyal, J.F. Du, M. Joshi, D.Q. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv:1907.11692*, 2019, pp. 1218–1227.

[18] V. Sanh, L. Debut, J. Chaumod, T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv:1910.01108*, 2019, pp. 1-5.

[19] M. Namazifar, A. Papangelis, G. Tur, D. Tur, "Language model is all you need: natural language understanding as question answering," *arXiv:2011.03023*, 2020, pp. 1-5.

[20] H.T. Vu, M.T. Nguyen, V.C. Nguyen, M.T. Tien, V.H. Nguyen, "Label correlation based graph convolutional network for multi-label text classification," *International Joint Conference on Neural Networks (IJCNN)*, 2022, pp.1-8.

[21] G.H. Qin, Y.K. Feng, B.V. Durme, "The nlp task effectiveness of long-range transformers," *arXiv:2202.07856*, 2022, pp. 3774–3790.

[22] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, "ICON: interactive conversational memory network for multimodal emotion detection,"*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp.2594-2604.

[23] J.R. Chowdhury, C. Caragea, "Modeling hierarchical structures with continuous recursive neural networks," *arXiv:2106.06038*, 2021, pp. 1-15.

[24] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria , R. Mihalcea, "MELD: a multimodal multi-party dataset for emotion recognition in conversations,"*Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019, pp.527-536.

[25] A. Chowanda, R. Sutoyo, S. Tanachutiwat, "Exploring text-based emotions recognition machine learning techniques on social media conversation", *Procedia Computer Science*, 2021, pp.821-828.

[26] qQ. Jin, C. Li, S. Chen, H. Wu, "Speech emotion recognition with acoustic and lexical features," *In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp.4749-4753.

[27] S.H. Park, B.C. Bae, Y.G. Cheong, "Emotion recognition from text stories using an emotion embedding model," *In2020 IEEE international conference on big data and smart computing (BigComp)*,2020, pp.579-583.

[28] Y.H. Rao, H.R. Xie, J. Li, F.M. Jin, F.L. Wang, Q. Li, "Social emotion classification of short text via topic-level maximum entropy model," *Information Management*, 2016, pp.978-86.

[29] J.H. Hsu, M.H. Su, C.H. Wu, Y.H. Chen, "Speech emotion recognition considering nonverbal vocalization in affective conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*2021, vol.29, pp.1675-86.

[30] K.C. Yao, L.B. Zhang, T.J. Luo, D.W. Du, Y.J. Wu, "Non-deterministic and emotional chatting machine: learning emotional conversation generation using conditional variational autoencoders," *Neural Computing and Applications*, 2021, vol.33, pp.5581-5589.

[31] N. Majumder, S. Poria, "DialogueRNN: An attentive RNN for emotion detection in conversations," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol.33, no.01, pp.6818-6825.

[32] D. Ghosal, N. Majumder, "DialogueGCN: a Graph convolutional neural network for emotion recognition in conversation,"*arXiv:1908.11540*, 2019, pp. 154-164.

[33] J. Li, M. Zhang, D, "Multi-task learning with auxiliary speaker identification for conversational emotion recognition,"*arXiv:2003.01478*, 2020, pp.1-7.

[34] Z. Lian, B. Liu, "DECN: dialogical emotion correction network for conversational emotion recognition,"*Neurocomputing*, 2021, 454, pp.483-495.

[35] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, 2008, vol.42, pp. 335-359.

[36] G.M. Hu, T.E. Lin, Y. Zhao, G.M. Lu, Y.C. Wu, Y.B. Li, "UniMSE: towards unified multimodal sentiment analysis and emotion recognition," *In Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 7837–7851.

[37] J.Y Son, J.S Kim, J.W Lim, "GRASP: guiding model with relational semantics using prompt for dialogue relation extraction,"*In Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 412–423.