Context-Dependent Embedding Utterance Representations for Emotion Recognition in Conversations

Patrícia Pereira[1,2], Helena Moniz[1,3], Isabel Dias[1,2] and **Joao Paulo Carvalho[1,2]**

[1]INESC-ID, Lisbon
[2]Instituto Superior Técnico, University of Lisbon
[3]Faculdade de Letras, University of Lisbon
{patriciaspereira, isabel.h.dias}@tecnico.ulisboa.pt
{helena.moniz, joao.carvalho}@inesc-id.pt

Abstract

Emotion Recognition in Conversations (ERC) has been gaining increasing importance as conversational agents become more and more common. Recognizing emotions is key for effective communication, being a crucial component in the development of effective and empathetic conversational agents. Knowledge and understanding of the conversational context are extremely valuable for identifying the emotions of the interlocutor. We thus approach Emotion Recognition in Conversations leveraging the conversational context, i.e., taking into attention previous conversational turns. The usual approach to model the conversational context has been to produce context-independent representations of each utterance and subsequently perform contextual modeling of these. Here we propose context-dependent embedding representations of each utterance by leveraging the contextual representational power of pre-trained transformer language models. In our approach, we feed the conversational context appended to the utterance to be classified as input to the RoBERTa encoder, to which we append a simple classification module, thus discarding the need to deal with context after obtaining the embeddings since these constitute already an efficient representation of such context. We also investigate how the number of introduced conversational turns influences our model performance. The effectiveness of our approach is validated on the open-domain DailyDialog dataset and on the task-oriented EmoWOZ dataset.

1 Introduction

Emotion Recognition in Conversations (ERC) is useful in automatic opinion mining, emotion-aware conversational agents and assisting modules for therapeutic practices. There is thus an increasing interest in endowing machines with efficient emotion recognition modules.
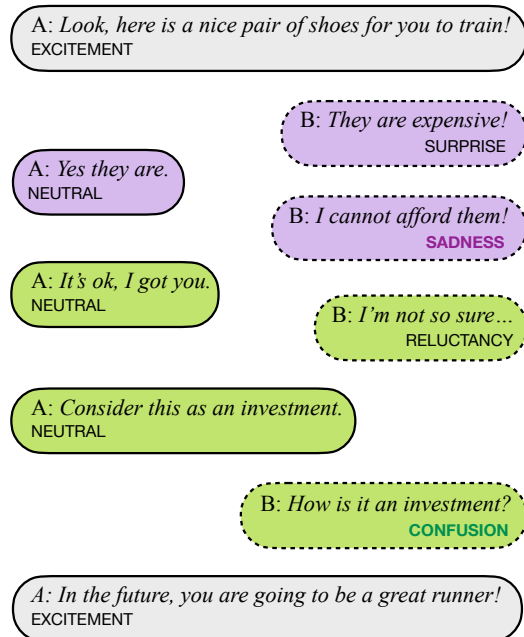

Figure 1: A dialogue in which context is key to infer the associated emotions. To infer the emotions *Sadness* and *Confusion*, knowledge of the present and previous two and three utterances is required, respectively.

Knowledge and understanding of the conversational context, i.e., of the previous conversation turns, are extremely valuable in identifying the emotions of the interlocutors (Poria et al., 2019) (Chatterjee et al., 2019) (Pereira et al., 2022).

Research in automatic emotion recognition using machine learning techniques dates back to the end of the 20th century. However, the use of the conversational context as an auxiliary information for the classifiers, did not appear until publicly available conversational datasets became more common.

State-of-the-art ERC works leverage not only state-of-the-art pre-trained-language models such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), but also deep, complex architectures to model several factors that influence the emotions in the conversation (Pereira et al., 2022). Such fac-

tors usually pertain to self and inter-speaker emotional influence and the context and emotion of preceeding utterances.

In this paper we argue that the powerful representation capabilities of pre-trained language models can be leveraged to model context without the need of additional elaborate classifier architectures, allowing for much simpler and efficient architectures. Furthermore, it is our contention that the Transformer, the backbone of our chosen language model, is better at preserving the contextual information since it has a shorter path of information flow than the RNNs typically used for context modelling. In this line, we rely on the RoBERTa language model and resort to a simple classification module to preserve the contextual information.

The usual approach to model the conversational context has been to produce context independent representations of each utterance and subsequently perform contextual modeling of those representations. State-of-the art approaches start by resorting to embedding representations from language models and employ gated or graph neural network architectures to perform contextual modelling of these embedding representations at a later step. In our much simpler and efficient proposed approach, we produce context-dependent embedding representations of each utterance, by feeding not only the utterance but also its conversational context to the language model. We thus discard the need to deal with context after obtaining the embeddings since these constitute already an efficient representation of such context.

Our experiments show that by leveraging context in this way, one can obtain state-of-the-art results with RoBERTa and a simple classification module, surpassing more complex state-of-the-art models.

2 Related Work

Amongst the first works considering contextual interdependences among utterances is the one by Poria et al. (Poria et al., 2017). It uses LSTMs to extract contextual features from the utterances. These gated recurrent networks make it possible to share information between consecutive utterances while preserving its order.

A more elaborate model also leveraging gated recurrent networks is DialogueRNN (Majumder et al., 2019), which uses GRUs to model the speaker, context and emotion of preceding utterances by keeping a party state and a global state that are used to

model the final emotion representation.

Gated recurrent networks have a long path of information flow which makes it difficult to capture long term dependencies. These can be better captured with the Transformer which a has shorter path of information flow. Its invention in 2017 (Vaswani et al., 2017) led to a new state-of-the-art in several Natural Language Processing tasks.

Amongst the first works leveraging the Transformer is the Knowledge-Enriched Transformer (KET) (Zhong et al., 2019). It uses its self-attention to model context and response. It also makes use of an external knowledge base, a graph of concepts that is retrieved for each word.

Following the invention of Transformers, pre-trained language models brought about another new state-of-the art in 2019. Since their invention, most state-of-the art ERC works resorted to encoder pre-trained language models (Shen et al., 2021a) (Ghosal et al., 2020) (Li et al., 2021).

COSMIC (Ghosal et al., 2020) leverages RoBERTa Large as feature extractor. Furthermore, it makes use of the commonsense transformer model COMET (Bosselut et al., 2019) in order to extract commonsense features. Five bi-directional GRUs model a context state, internal state, external state, intent state, and emotion state that influence the final emotion classification.

Psychological (Li et al., 2021) also uses RoBERTa Large for utterance encoding and COMET. For conversation-level encoding it constructs a graph of utterances to model the actions and intentions of the speaker along with the interactions with other utterances. It uses COMET to introduce commonsense knowledge into the graph edge representations and processes this graph using a graph transformer network.

3 Methodology

We describe how we obtain a contextual embedding representation of the sentence and its context with RoBERTa, how we pool the contextual embeddings, our classification module and how we obtain the emotion labels. These processes can be observed in Figure 2.

3.1 Task definition

Given a conversation, a sequence of u_i utterances with corresponding $emotion_i$ from a predefined set of emotions, the aim of the task of ERC is to correctly assign an emotion to each utterance of the
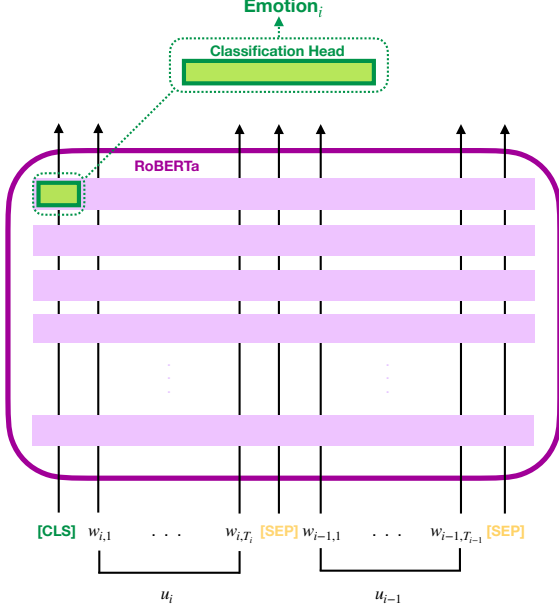
Figure 2: Model architecture. Two utterances are given as input to RoBERTa encoder, of which the CLS token of the last layer is fed to the classification head that predicts the emotion.

conversation. An utterance consists in a sequence of w_{it} tokens representing its T_i words

$$u_i = (w_{i1}, w_{i2}, ..., w_{iT_i}) \quad (1)$$

The usual approach for this task has been to produce context independent representations of each utterance and perform contextual modeling of these. In our approach we produce context-dependent representations of each utterance that represent not only the utterance but also a given number of previous utterances from the conversation.

3.2 Context-dependent feature extraction

For context-dependent feature extraction, we feed as input to RoBERTa the utterance we intend to classify, u_i, concatenated with its conversational context corresponding to the number c of previous utterances in the conversation, $(u_{i-1}, u_{i-2}, ..., u_{i-c})$. Concretely, we feed u_i to the model, preceded by the [CLS] token and suceded by the [SEP] token, followed by the previous turns u_{i-1} up to u_{i-c}, separated by the [SEP] token.

3.3 Pooling

The RoBERTa encoder outputs several layers of embeddings representing the utterance, and in our approach, also the preceding utterances it receives

as input. Each layer comprises several tokens, being the number of tokens the same as the number of input tokens. Each token is a vector with dimension corresponding to the RoBERTa hidden size.

From these embeddings one can extract a suitable representation for the sentence. Choosing all tokens from all layers would yield an extremely memory demanding classification layer and may not yield the best model performance. Thus we choose the first embedding from the last layer L, the [CLS] which is used for classification, as in Equation 2.

$$pooled_i = RoBERTa_{L,[CLS]}(input_i) \quad (2)$$

3.4 Emotion Classification

The classification module that follows RoBERTa is a linear fully connected layer, applying a linear transformation to the pooled encoder output data. Its input size is the RoBERTa encoder hidden size and its output size is the number of emotion classes.

The final label probability distribution is yielded by applying the softmax operation to the output of the classification head and the predicted label is the one with the highest probability:

$$emotion_i = argmax(Softmax(pooled_i W^T + b)) \quad (3)$$

4 Experimental Setup

4.1 Training

Our model is based on RoBERTa-base from the Transformers library by Hugging Face (Wolf et al., 2020). It is trained with the cross-entropy loss with logits. The Adam (Kingma and Ba, 2014) optimizer is used with an initial learning rate of 1e-5 and 5e-5, for the encoder and the classification head, respectively with a layer-wise decay rate of 0.95 after each training epoch for the encoder. The encoder is frozen for the first epoch. The batch size is set to 4. Gradient clipping is set to 1.0. As stopping criteria, early stopping is used to terminate training if there is no improvement after 5 consecutive epochs on the validation set over macro-F1, for a maximum of 10 epochs. The checkpoint used in testing is the one that achieves the highest macro-F1 score on the validation set.

4.2 Evaluation

We evaluate the performance of our model with the macro F1-score. The reported results are yielded

from an average of 5 runs corresponding to 5 distinct random seeds that are kept for a meaningful comparison of all experiments. This average is motivated by the fact that results for the same experiment obtained with different random seeds can have a variability of about 3 in macro F1-score which is a large deviation given that our proposed approach yields an improvement of that magnitude and comparison between state-of-the-art models are based on improvements of less than 1 F1-score. This procedure is in line with several authors that also resort to 5 run averages (Li et al., 2021) (Zhong et al., 2019) (Shen et al., 2021a) (Shen et al., 2021b).

Our code is publicly available[1].

4.3 Datasets

We evaluate our approach on the chit-chat Daily-Dialog (Li et al., 2017) dataset and on the task-oriented EmoWOZ (Feng et al., 2022) dataset.

4.3.1 DailyDialog

DailyDialog is built from websites used to practice English dialogue in daily life. It is labelled with the six Ekman's basic emotions (Ekman, 1999), anger, disgust, fear, happiness, sadness and surprise, or neutral. The publicly available splits of Yanran are used.

4.3.2 EmoWOZ

EmoWOZ is derived from MultiWOZ (Budzianowski et al., 2018), one of the largest multi-domain corpora benchmark dataset for various dialogue tasks. User utterances are annotated with either fear, dissatisfaction, apologetic, abusive, excited, satisfied or neutral emotions.

The statistics and proportion of labels in the datasets are presented in Tables 1 and 2, respectively.

Table 1: Statistics of the datasets

	DailyDialog	EmoWOZ
Dlg type	Chit-chat	Task-oriented
# Dlgs	13,118	11,434
# Turns	102,979	167,234
Avg turns in dlg	7.9	14.6

Table 2: Proportion of labels in the datasets

DailyDialog			
Ang	Disg	Fear	Hap
1.0%	0.3%	0.2%	12.5%
	Sad	Sur	Neu
	1.1%	1.8%	83.1%
EmoWOZ			
Fear	Diss	Apol	Abus
0.5%	6.1%	1.0%	0.2%
	Exc	Sat	Neu
	1.2%	21.0%	70.1%

From Table 1 it can be noted that EmoWOZ has almost double the amount of average turns per dialogue than DailyDialog.

From Table 2 it can be observed that both datasets are imbalanced, not only for its dominant majority neutral class, but also for the relative imbalance between minority classes. Therefore, we have opted to use the macro-F1 score for evaluation in order to promote consistent performance across all classes.

5 Results and Analysis

5.1 Iterating towards the ideal approach

We have performed extensive experiments in order to obtain our ideal model architecture. From experimenting different approaches to pool the various layers of embeddings RoBERTa provides to choosing which classification module to employ withing a wide variety of deep learning architectures, we put forward our experiments in this subsection.

5.1.1 Fine-tuning

Fine-tuning, the modification of the pre-trained RoBERTa's weights along with the classification head during training with the target dataset, is a determinant procedure for the success of our approach.

In our experiments we observed that if we did not fine-tune the language model and just trained the classification head, the model would always predict the majority neutral class. This supports the notion that pre-trained-language models are useful for a wide variety of tasks but need to be fine-tuned for the specific task at hand.

5.1.2 Pooling

We have performed experiments with several pooling alternatives. From average pooling, max pool-

ing, concatenation of the CLS token of more than 1 last layers to the concatenation of the CLS token with the result from average pooling. All these pooling alternatives resulted in lower performance than choosing the CLS token of the last layer. This might suggest a high representative power for the CLS token, which is proposed for classification, and discards the need for directly considering other tokens for this task.

5.1.3 Classification module

We have also performed alternative experiments with other classification modules than our simple classification head. These consisted in passing the pooled embeddings through Recurrent Neural Networks (Elman, 1991), uni (Hochreiter and Schmidhuber, 1997) and bi-directional (Graves et al., 2005) Long Short-Term Memory Networks and a Conditional Random Field (Lafferty et al., 2001) before feeding them to the classification head. Performance was lower in all alternative experiments when compared to our main approach of using a simple classification head. These results may indicate that our approach leveraging RoBERTa's representational power for context suffices and there is no apparent need for modelling the context with complex classification modules, after obtaining our context-dependent embedding utterance representations.

5.2 Overall Performance

For each of the datasets, we have performed experiments without introducing any context ($c = 0$) to introducing 4 previous conversation turns ($c = 4$), for which the overal performance operationalized by the macro-F1 metric is reported in Table 3. Our results are an average of 5 runs.

Table 3: Model performance in macro F1-score with the introduction of c conversational turns

	DailyDialog macro-F1	EmoWOZ macro-F1
c=0	48.52	58.66
c=1	50.31	62.32
c=2	50.44	64.98
c=3	**51.23**	**65.33**
c=4	50.46	63.28

It can be observed that introducing previous conversational context turns leads to an increase in macro-F1 score. As hypothesised, providing no context is never the best option. This shows that the introduction of an adequate number of context turns directly as the language model input significantly improves model performance. In general performance increases with the introduction of each additional context turn up to the ideal number of turns and then it decreases. Overall, it can be concluded that the ideal number of introduced context turns for ERC in both datasets is 3.

5.3 Performance on each emotion label

For each dataset, we report the results on each individual emotion label and also present the confusion matrices for the best determined c value. Our results are an average of 5 runs.

The individual emotion label F1-scores for the DailyDialog dataset are presented in Table 4.

It can be observed that for more than half of the labels, Anger, Fear, Sadness and Neutral, the ideal context to be provided is 3 turns which maximise their F1-scores, and also the macro-F1 score on Table 3, and for the other labels the ideal context is 4 turns for Disgust, 2 turns for Happiness and 1 turn for Surprise. As expected, providing no context is never the best option.

The confusion matrix for $c = 3$ corresponding to the highest macro-F1 score is displayed on Figure 3, in which the label nomenclature and order is the same as in table 4 but with neutral as the first label.

This matrix indicates that majority of the errors are due to classifying utterances as neutral instead of assigning a non-neutral emotion. The classifier also displays some confusion in discerning between Happiness and Surprised.

The individual emotion label F1-scores for the EmoWOZ dataset are presented in Table 5.

It can be observed that for 4 of the labels, Dissatistfied, Excited, Satisfied and Neutral, the ideal context to be provided is 4 turns which maximise their F1-scores. Regarding the other labels the ideal context is 2 for Fear, 3 for Abusive, and surprisingly 0 turns for Apologetic, which might indicate that this emotion is very explicit in this dataset.

The confusion matrix for $c = 3$ corresponding to the highest macro-F1 score is displayed on Figure 4, in which the label nomenclature and order is the same as in table 5 but with neutral as the first label.

This matrix indicates that majority of the errors are due to classifying utterances as neutral instead of assigning a non-neutral emotion, as in happens with the DailyDialog dataset.

Table 4: Model performance on each individual emotion label on the **DailyDialog** dataset with the introduction of c conversational turns

	Ang	Disg	Fear	Hap	Sad	Sur	Neu
c=0	37.47	32.32	36.69	59.42	33.16	49.60	90.99
c=1	40.18	29.28	39.43	61.26	38.30	**52.66**	91.06
c=2	43.26	33.91	36.52	**61.98**	33.63	52.23	91.12
c=3	**43.51**	33.22	**39.44**	61.12	**38.43**	51.50	**91.42**
c=4	42.00	**34.52**	34.65	61.97	37.18	51.70	91.18

Table 5: Model performance on each individual emotion label on the **EmoWOZ** dataset with the introduction of c conversational turns

	Fear	Diss	Apol	Abus	Exc	Sat	Neu
c=0	35.72	45.18	**74.93**	25.21	46.96	90.09	92.53
c=1	32.97	57.97	72.47	42.97	47.07	89.75	93.01
c=2	**38.91**	66.24	73.37	44.79	48.13	89.73	93.74
c=3	37.89	68.02	72.49	**47.73**	47.64	89.76	93.81
c=4	35.15	**69.57**	73.00	30.09	**50.89**	**90.23**	**94.03**

Figure 3: Confusion Matrix for the **DailyDialog** dataset with the introduction of $c=3$ conversational turns

Figure 4: Confusion Matrix for the **EmoWOZ** dataset with the introduction of $c=3$ conversational turns

It is worth noting that our results are an average of 5 runs and the final model is determined via performance on the validation set. Therefore, the fluctuation in individual label F1-scores does not hinder the representativity of our results and these fluctuations may occur between results from the other reported state-of-the-art models.

5.4 Comparison with state-of-the-art

We further compare our approach to other state-of-the-art approaches that also resort to the RoBERTa or BERT pre-trained-language models. This allows for a fair comparison between approaches

given that using this language model brings great performance increases when compared to using other means of utterance feature extraction. Regarding DailyDialog results, we compare our approach to COSMIC (Ghosal et al., 2020), RoBERTa and RoBERTa DialogueRNN, implemented by the authors of COSMIC, and the Psychological model (Li et al., 2021), all models described in Section 2. Concerning the performance on the EmoWOZ dataset, we compare out approach to COSMIC, BERT and BERT DialogueRNN, tested by the authors of EmoWOZ (Feng et al., 2022), since for this dataset the authors obtained a more suitable

Table 6: Comparison with state-of-the-art works

	DailyDialog macro-F1	EmoWOZ macro-F1
RoBERTa (Ghosal et al., 2020) / BERT (Feng et al., 2022)	48.20	55.80
RoBERTa (Ghosal et al., 2020) / BERT (Feng et al., 2022) + DlgRNN	49.65	57.10
ContextBERT (Feng et al., 2022)	-	59.70
COSMIC (Ghosal et al., 2020) / (Feng et al., 2022)	51.05	61.12
Psychological (Li et al., 2021)	**51.95**	-
CD-ERC (Ours)	51.23	**65.33**

uterrance representation using BERT instead of RoBERTa. Results are displayed on table 6 and are an average of 5 runs.

Regarding performance on the DailyDailog dataset, our approach outperforms not only the simple RoBERTa/BERT, but also RoBERTa/BERT in a more elaborate gated neural network model such as DialogueRNN and COSMIC. The Psychological model has a slightly higher performance than ours. It may be due to the fact that it leverages a large commonsense knowledge base and an elaborate classifier architecture, while we opted for a minimalistic classification module. Concerning performance on the EmoWOZ dataset, our approach outperforms all baselines by a wide margin, setting a new state of the art for task-oriented emotion datasets.

5.5 Case Studies

On Table 7 we can compare the performance of our contextual classifier when considering the ideal 3 context turns on both datasets versus not considering any context at all.

In the first example, from the DailyDialog dataset, A offers B assistance, so B asks A to view the apartment, to which A sadly apologizes informing B that B will not be able to view it. The classifier that does not consider context classifies this last apology as neutral. However, given the context of the conversation, A should not be neutral since A is unable to assist B which was A's initial purpose. The contextual classifier is able to consider this, thus correctly classifying A's utterance with the emotion Sadness.

In the second example, also from the DailyDialog dataset, A gives B a good idea to which B happily reacts and thanks A. A happily reacts to B's acknowledgments, especially since B mentioned A's was a "wonderful idea". The classifier that does not consider context classifies A's final reac-

tion to B as neutral, since A's utterance is a merely "No problem. Good luck", not being able to recognize A's positive reaction to B's acknowledgements. The contextual classifier, however, having this utterances into account, correctly classifies A's final reaction with the emotion Happiness.

In the last example, from the EmoWOZ dataset, B is merely answering A's question of what day B would like to travel. The classifier that does not consider context takes into account the words "please" and "vacation" which bias the classification towards the emotion Excited. The contextual classifier might grasp that "please" is used as a polite expression and "vacation" is just the object of the phrase, thus correctly classifying the utterance as neutral.

6 Conclusions and Future Work

In this work we have leveraged context-dependent embedding utterrance representations for Emotion Recognition in Conversations. Our approach of producing context-dependent representations of each utterance contrasted with the usual approach of producing context independent representations of each utterance and subsequently performing contextual modeling of these. It consisted in feeding a variable number of previous conversational turns appended to the utterance to be classified as input to the state-of-the-art pre-trained-language model RoBERTa, to which we appended a simple classification module. We further investigated how the number of introduced conversational turns influenced our model performance. We concluded that the introduction of an adequate number of context turns directly as the language model input significantly improves model performance.

Furthermore, we attained state-of-the-art results on the widely used DailyDialog dataset and established a new state-of-the-art by a wide margin on the EmoWOZ dataset, which are usually yielded by

Table 7: Case studies comparing the performance of our contextual classifier ($c = ideal = 3$) with the no-context classifier ($c = 0$)

Turn	Gold	$c = 0$	$c = 3$
A: Can I help you ?	Neu	Neu	Neu
B: I would actually like to view the apartment for rent today .	Neu	Neu	Neu
A: **I ' m sorry , but you won ' t be able to view it today .**	**Sad**	**Neu**	**Sad**
A: Maybe you should look around for an outlet .	Neu	Neu	Neu
B: That is a wonderful idea .	Hap	Hap	Hap
A: Outlets have more reasonable prices .	Neu	Neu	Neu
B: Thank you for your help .	Hap	Hap	Hap
A: **No problem . Good luck**	**Hap**	**Neu**	**Hap**
A: On what day would you like to travel?	-	-	-
B: **Saturday, please. I'm thinking just a short vacation over the weekend.**	**Neu**	**Exc**	**Neu**

more elaborate classifiers resorting to larger state-of-the-art pre-trained-language models and more complex classification modules.

For future work, from adequately capturing the conversation context, the focus of our approach, to capturing several other factors that influence the emotions in the conversation, such as self and inter-speaker emotional influence and the emotion of preceeding utterances, various architectures comprising not only state-of-the art language models for embeddings but also combining our context-dependent embedding utterance representation with more elaborate classification modules can be used.

Finally, we put forward important ethical aspects pertaining to Emotion Recognition in Conversations. These are, for example and not limited to, whether an ERC module should be developed or used for a certain purpose, which data to collect, the subjects behind the data, diversity, inclusiveness, privacy, control and possible biases and misuses of the application (Mohammad, 2022). Research taking into account these aspects will benefit the community with better ERC modules for current and novel applications.

7 Acknowledgements

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 39–48.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Jeffrey L Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7(2):195–225.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. Emowoz: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion

identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. In *International conference on artificial neural networks*, pages 799–804. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1204–1214, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. 2022. Deep emotion recognition in textual conversations: A survey. *arXiv preprint arXiv:2211.09172*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. 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)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. In *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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *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.