UMUTeam at SemEval-2024 Task 10: Discovering and Reasoning about Emotions in Conversation using Transformers

Ronghao Pan[1], **José Antonio García-Díaz**[1], **Diego Roldán**[2], **Rafael Valencia-García**[1]
[1] Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
[2] DANTIA Tecnología S.L., Parque Empresarial de Jerez 10,
Calle de la Agricultura, 11407, Jerez de la Frontera, Cádiz, España
{ronghao.pan, joseantonio.garcia8, valencia}@um.es
droldan@dantia.es

Abstract

These notes describe the participation of the UMUTeam in EDiReF, the 10th shared task of SemEval 2024. The goal is to develop systems for detecting and inferring emotional changes in the conversation. The task was divided into three related subtasks: (i) Emotion Recognition in Conversation (ERC) in Hindi-English code-mixed conversations, (ii) Emotion Flip Reasoning (EFR) in Hindi-English code-mixed conversations, and (iii) EFR in English conversations. We were involved in all three and our approach is based on a fine-tuning approach with different pre-trained models. After evaluation, we found BERT to be the best model for ERC and EFR and with this model we achieved the thirteenth best result with an F1 score of 43% in Subtask 1, the sixth best in Subtask 2 with an F1 score of 26% and the fifteenth best in Subtask 3 with an F1 score of 22%.

1 Introduction

Emotion, often defined as an individual's mental state associated with thoughts, feelings and behavior, has been categorized in various ways throughout history. Modern classifications include Plutchik's (Plutchik, 1982) eight primary types and Ekman's (Ekman, 1993) emphasis on facial expressions. In Natural Language Processing (NLP), emotion recognition has gained popularity for its applications in opinion mining, healthcare, etc. Although textual emotion recognition has been studied extensively, attention has recently shifted to Emotion Recognition in Conversation (ERC), driven by the availability of conversational data (Yeh et al., 2019) (Chen et al., 2018).

Conversation or dialogue is the main mode of information exchange between individuals, highlighting the prevalence of code-mixed (Kasper and Wagner, 2014), where multiple languages are integrated into the conversation. Despite extensive research on ERC, previous studies have largely focused on monolingual dialogues, neglecting code-mixed conversations. However, in the paper (Kumar et al., 2023a), the authors propose ERC models adapted to code-mixed dialogues, highlighting the need for datasets and resources in this area. Furthermore, they propose to incorporate common sense knowledge to better understand the emotions evoked in the conversation, and present a process to adapt existing English-based common sense knowledge graphs for code-mixed input.

ERC aims to identify emotions in sequences of utterances or dialogues rather than in isolated texts. In many cases, it is necessary to understand the emotional changes in a conversation is necessary in addition to identifying the speaker's emotion. However, understanding the emotional changes in a conversation is an challenging task that requires detailed analysis. Hence, the task of Emotion Flip Reasoning (EFR) (Kumar et al., 2022), which focuses on identifying the cause of a speaker's emotional change in a dialogue.

The EDiReF shared task (SemEval 2024) focuses on discovering and explaining the emotion change in the conversation (Kumar et al., 2024). It is divided into three subtasks: (1) **Subtask 1: ERC in Hindi-English code-mixed conversations**. Given a Hindi-English code-mixed dialog, the goal is to assign an emotion to each utterance from a predefined set of possible emotions (Kumar et al., 2023c); (2) **Subtask 2: EFR in Hindi-English code-mixed conversations**. Given a Hindi-English code-mixed dialog, the goal is to identify the trigger utterance(s) for an emotion flip in a multi-party conversation dialog (Kumar et al., 2022, 2023b); and (3) **Subtask 3: EFR in English conversations**. Given an English conversation, the goal is to identify the trigger utterance(s) for an emotion flip in a multi-party conversation dialog (Kumar et al., 2022, 2023b).

For this task, we propose an approach based on fine-tuning pre-trained Transformer-based models.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 703–709
June 20-21, 2024 ©2024 Association for Computational Linguistics

In a nutshell, fine-tuning is a process by which a pre-trained model, previously trained on a specific task, is adjusted to adapt to a related but different task using a labeled dataset. In addition, a text processing process has been performed where, if possible, past and future conversations or emotions are added to the current user's sentence as input to the model. In this way, the model can have the context of the user's emotion in the past and future states.

These working notes are organized as follows. In Section 2, the reader will find a summary of important details about the task setup. Section 3 gives an overview of our system. Next, Section 4 presents the specific details of our systems. The results are then discussed and presented in Section 5. Finally, the conclusions are presented in Section 7.

2 Background

Sentiment Analysis (SA) is the study of human attitudes and feelings in specific situations, focusing on understanding emotions expressed through speech, voice, facial expressions and behavior. It typically identifies positive, negative and neutral emotions (Fu et al., 2023). In contrast, Emotion Recognition (ER) attempts to identify more nuanced emotions such as joy, hate and disgust, and modern classifications include Plutchik's (Plutchik, 1982) eight primary types and Ekman's (Ekman, 1993) emphasis on facial expressions. Emotion recognition spans text, audio and video modalities and differs from sentiment analysis in that it considers the context and interdependence between speakers within a conversation.

Multimodal emotion recognition has become an important research topic, mainly due to its potential applications in many challenging tasks such as dialog generation, user behavior understanding, multimodal interaction, and others. Therefore, a conversational emotion recognition system can be used to generate appropriate responses by analyzing the user's emotions. According to (Poria et al., 2019), ERC poses several challenges such as modeling the conversational context, emotion shifts of interlocutors, and others, which make the task more challenging. Recent works propose solutions based on multimodal memory networks (Hazarika et al., 2018). However, they are mostly limited to dyadic conversations and are therefore not scalable to ERC with multiple interlocutors. Furthermore, previous studies have largely focused on monolingual dialogues, neglecting code-mixed conversations (Kumar et al., 2023a).

In a conversation, utterances generally depend on the context of the conversation. This is also true for the emotions associated with them. In other words, the context acts as a set of parameters that can influence a person to make an utterance while expressing a certain emotion. This context can be modeled in different ways, for example using Recurrent Neural Networks (RNN) and Memory Networks (Hazarika et al., 2018) (Serban et al., 2017). Public datasets available for multimodal emotion recognition in conversation, such as IEMOCAP (Busso et al., 2008) and SEMAINE (McKeown et al., 2010), have facilitated a significant number of research projects, but they also have limitations due to their relatively small number of total utterances and the lack of multipart conversations.

Understanding the emotional flips in a conversation requires a detailed analysis. This is where Emotional Flip Reasoning (EFR) comes in, which focuses on identifying the cause of a speaker's emotional flip in a dialogue. The EFR process consists of three stages (Kumar et al., 2022): identifying the utterance in which the emotional flip occurs, identifying the triggers responsible for the change, and assigning psychologically motivated instigator labels to the triggers to explain the emotional flip. Therefore, the EFR task has the potential to improve the user experience in a conversational dialog system, especially in the generation of empathetic responses (Lin et al., 2019), (Ma et al., 2020).

In recent years, with the rapid development in the field of NLP, many pre-trained models based on Transformer have emerged. These models are trained on large corpora of unlabeled text and, due to their transfer learning capability, can be adapted to different tasks such as classification, translation, response generation without the need of a large training corpus. For example, (García-Díaz et al., 2023) and (García-Díaz and Valencia-García, 2022) demonstrated the effectiveness of Transformers-based models for identifying hate speech and satire. Therefore, in this study, different pre-trained models were evaluated for the ERC and EFR tasks.

The models evaluated are: (1) XLM-RoBERTa-base (Conneau et al., 2019); (2) DeBERTa-V3-base(He et al., 2021); and (3) BERT (Devlin et al., 2018). For the ERC and EFR tasks, we evaluated the basic version and the version without the mask, which removes the accent markers.

3 System overview

Figure 1 shows the general architecture of our approach for the three subtasks, which is mainly divided into two modules: data processing and fine tuning.

In the processing module, for Subtask 1 (ERC), we first translated the statements into English, since most language models are pre-trained in English and have shown good performance in the emotion identification and sentiment analysis tasks. They were then grouped by user, as this provides a coherent context for analyzing their emotional state, rather than adding conversational context from other speakers. Therefore, by examining all the interventions made by the same speaker, we gain a deeper understanding of their emotional state at the time of the target intervention. Furthermore, we believe that adding more context could introduce noise and reduce the performance of the models. Once grouped, for each current statement of the user, the previous statement was concatenated with the next by a semicolon. For example, for statement U3 from a particular user, the input to the model would be `U2;U3;U4`. For subtasks 2 and 3 (EFR) in addition to concatenating the previous and subsequent statements from the same user for each current statement, the emotion of each statement is added. For example, for statement U3, the input to the model would be `U2-e2;U3-e3;U4-e4`, where *e* represents the user's emotion at that moment. The figure 2 shows examples of processing for the user *Ross* in a specific conversation.

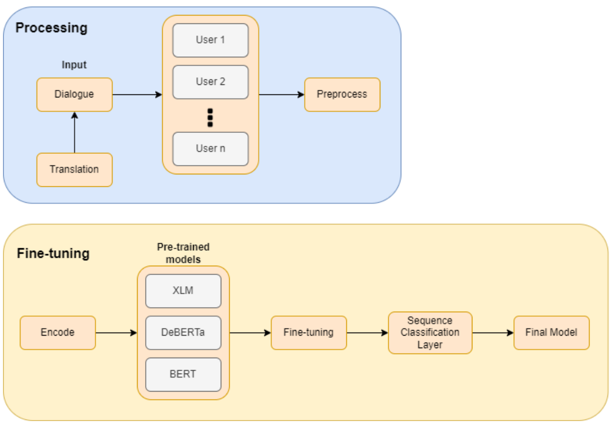

Figure 1: System architecture

In the fine tuning module (see Figure 1), the inputs are first tokenized according to the tokenizers of the pre-trained models. Next, the pre-trained model is loaded as the basis for the classification task. Next, a sequence classification layer is added on top of the pre-trained model. This layer takes the last hidden state generated by the pre-trained model and performs classification based on the labels of the specific classification task. In this case, we used the sequence classification layer from the *Transformers*[1] library for each pre-trained model. Finally, the tuning is performed out and a performance is evaluated using the validation set.

4 Experimental setup

To train the three subtasks, we used the data set provided by the organizers, which consists of a training set and a validation set. In Figure 3 and Table 1 we can see the distribution of the training and validation sets for the three subtasks.

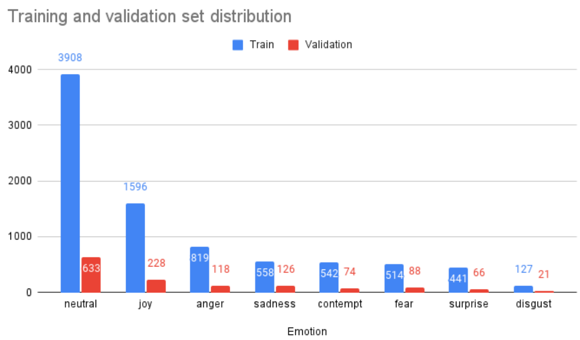

Figure 3: Training and validation set distribution of Subtask 1.

Table 1: Training and validation set distribution of Subtask 2 and 3.

Set	Triggers	No triggers
Subtask 2		
Train	6542	92235
Validation	434	7028
Subtask 3		
Train	5575	29416
Validation	494	3027

For all three subtasks (1, 2, 3), we used the same fine-tuning hyperparameters, namely: a batch size of 8 for both training and validation, 10 epochs, a learning rate of 2e-5, and a weight decay of 0.01. During training, we used the weighted F1 as a reference. To evaluate the three subtasks, the organizers

[1] https://github.com/huggingface/transformers

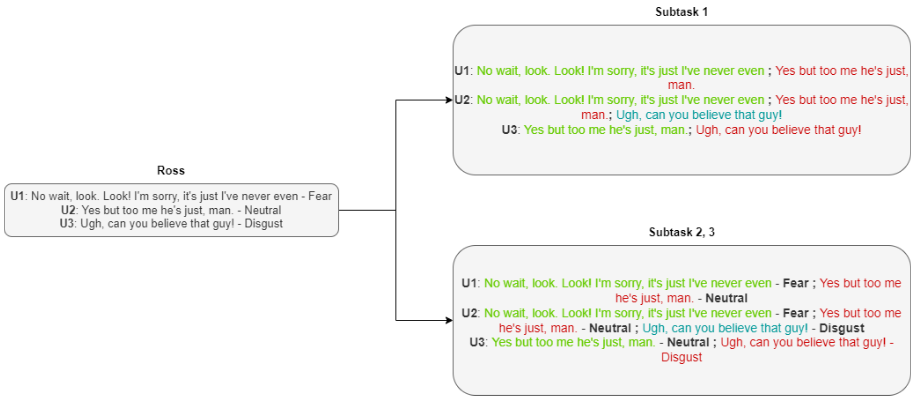

Figure 2: Examples of processing for subtasks 1, 2, and 3.

used the weighted F1, an evaluation metric used in classification problems that takes into account the class imbalance in the data. While the traditional F1 score calculates the harmonic mean of precision and recall for all classes equally, the weighted F1 score weights these measures according to the number of samples in each class.

5 Results

Table 2 shows the results obtained on the test set with different models for Subtask 1 on the ERC. We can see that the XLM-R model obtained the best result with a weighted F1 score of 42.878%, followed by BERT with a weighted F1 score of 42.691%.

Table 2: Evaluation of different pre-trained models in test set of Subtask 1.

Model	W-R	W-P	W-F1
Subtask 1			
XLM-R	**44.9367**	42.1941	**42.878**
DeBERTa	43.5443	41.0664	41.7686
BERT	44.8734	**42.4540**	42.6910

Table 3 shows the results of Subtask 2, which is an EFR task, on a dataset of Hindi-English code-mixed conversations. The evaluation metric is the F1 score of the triggers, and it can be seen that BERT is the only model that obtained a score greater than 0, with 25.8721% in F1 score. The XLM-R and DeBERTa models were not able to predict emotion change triggers well because they were fine-tuned with the same hyperparameters, so it may be necessary to use different hyperparameters, such as a smaller learning rate. Therefore, as a future line, it is proposed to perform hyperparameter tuning to fine-tune the models to achieve better performance.

Regarding Subtask 3, which has the same objective as Subtask 2, but on a dataset of English code-mixed conversations, it can be observed that BERT and DeBERTa are the only two models that have obtained an F1 score greater than 0, with 22.4764% for BERT and 17.1111% for DeBERTa (see Table 3).

Table 3: Evaluation of different pre-trained models in test set of Subtask 2 and 3.

Model	Recall	Precision	F1
Subtask 2			
XLM-R	0.0	0.0	0.0
DeBERTa	0.0	0.0	0.0
BERT	**21.3942**	**32.7206**	**25.8721**
Subtask 3			
XLM-R	0.0	0.0	0.0
DeBERTa	13.1737	24.4057	17.1111
BERT	**19.3328**	**87.9103**	**22.4764**

Therefore, we have chosen the BERT model for this task, since it outperforms the other models in all three subtasks, except for the first, where it is 0.187% worse than XLM-RoBERTa, which does not exceed 1%. In this case, we have obtained the thirteenth position in Subtask 1, the sixth in Subtask 2 and the fifteenth in Subtask 3.

6 Error analysis

For error analysis, we extracted the confusion matrix from BERT using the Subtask 1's test sets. A confusion matrix is a tool used in error analysis, especially in classification scenarios, by illustrating the performance of a model in predicting true class labels compared to the model-predicted classes.

In Figure 4, we can see that our system tends to confuse the *Neutral* emotion in the ERC task, due to the unbalanced training set provided by the organizers, where the *Neutral* emotion occupies the highest percentage. Furthermore, the disgust emotion was not correctly identified in any case.

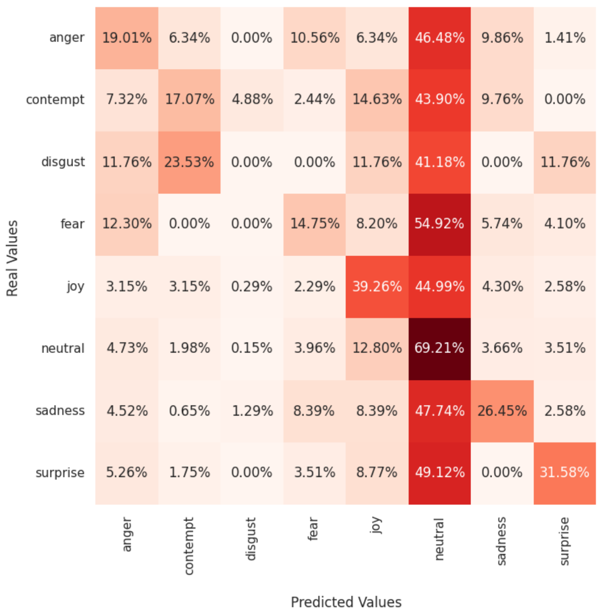

Figure 4: BERT confusion matrix in the test set of subtask 1.

Table 4 shows a classification report of our model in the EFR task of Hindi-English code-mixed conversation (Subtask 2). We can see that our system tends to identify instances as "No triggers" and has a higher recall due to the imbalance in the training set, which contains more instances of "no triggers". As for Subtask 3, the same phenomenon occurs as in Subtask 2, as shown in Table 5.

Table 4: BERT's classification report of Subtask 2 in the test set.

	Precision	**Recall**	**F1**
No triggers	95.5918	97.4842	96.5287
Triggers	32.7206	21.3942	25.8721
Macro avg	64.1562	59.4392	61.2004
Weighted avg	92.1907	93.3680	92.7065

Table 5: BERT's classification report of Subtask 3 in the test set.

	Precision	**Recall**	**F1**
No triggers	87.9103	91.7570	89.7924
Triggers	26.8409	19.3328	22.4764
Macro avg	57.3756	55.5449	56.1344
Weighted avg	79.6494	81.9602	80.6866

7 Conclusion

We have described the UMUTeam's participation in the 10th shared task 10 of SemEval 2024, the goal of which was to develop models for detecting and reasoning about the emotion change in the conversation. The task consists of three subtasks: (i) Emotion Recognition in Conversation (ERC) in Hindi-English code-mixed conversations, (ii) Emotion Flip Reasoning (EFR) in Hindi-English code-mixed conversations, and (iii) EFR in English conversations.

For all three subtasks, we used the fine-tuning approach of pre-trained models and performed a text processing process where, where possible, previous and future conversations or emotions are added to the current user's sentence as input to the model. In terms of results, our system achieved the thirteenth best result in Subtask 1 with an F1 of 43%, the sixth best in Subtask 2 with an F1 of 26%, and the fifteenth best in Subtask 3 with an F1 of 22%.

The study of emotional shifts provides a valuable insights for understanding psychographic characteristics in author profiling in the political context. Political communication is inherently intertwined with emotional appeals, and the ability to identify patterns of emotional shifts provides insight into the psychological makeup of political authors. Therefore, we plan to further validate the effectiveness of emotion flip inference by applying it to our PoliticES 2022 and 2023 datasets (García-Díaz et al., 2022; Garcia-Díaz et al., 2023) thus, contributing to a more comprehensive understanding of the ideologies, motivations, and communication strategies of political figures.

Acknowledgments

This work is part of the research projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way to make Europe and

LT-SWM (TED2021-131167B-I00) funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. In addition, this work was funded by the Spanish Government, the Spanish Ministry of Economy and Digital Transformation through the Digital Transformation through the "Recovery, Transformation and Resilience Plan" and also funded by the European Union NextGenerationEU/PRTR through the research project 2021/C005/00149877. Mr. Ronghao Pan is supported by the "Programa Investigo" grant, funded by the Region of Murcia, the Spanish Ministry of Labour and Social Economy and the European Union - NextGenerationEU under the "Plan de Recuperación, Transformación y Resiliencia (PRTR)".

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Yao Fu, Shaoyang Yuan, Chi Zhang, and Juan Cao. 2023. Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods. *Electronics*, 12(22).

José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, and Rafael Valencia-García. 2023. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, 9(3):2893–2914.

José Antonio Garcia-Díaz, Salud María Jiménez-Zafra, María-Teresa Martín-Valdivia, Francisco García-Sánchez, Luis Alfonso Ureña-López, and Rafael Valencia-García. 2023. Overview of politices at iberlef 2023: Political ideology detection in spanish texts. *Procesamiento del Lenguaje Natural*, 71:409–416.

José Antonio García-Díaz, Salud María Jiménez-Zafra, María-Teresa Martín Valdivia, Francisco García-Sánchez, L Alfonso Ureña-López, and Rafael Valencia-García. 2022. Overview of politices 2022: Spanish author profiling for political ideology. *Procesamiento del Lenguaje Natural*, 69:265–272.

José Antonio García-Díaz and Rafael Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, 8(2):1723–1736.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Gabriele Kasper and Johannes Wagner. 2014. Conversation analysis in applied linguistics. *Annual Review of Applied Linguistics*, 34:171–212.

Shivani Kumar, Md Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024. Semeval 2024 – task 10: Emotion discovery and reasoning its flip in conversation (ediref). In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Shivani Kumar, Md Shad Akhtar, Tanmoy Chakraborty, et al. 2023a. From multilingual complexity to emotional clarity: Leveraging commonsense to unveil emotions in code-mixed dialogues. *arXiv preprint arXiv:2310.13080*.

Shivani Kumar, Shubham Dudeja, Md Shad Akhtar, and Tanmoy Chakraborty. 2023b. Emotion flip reasoning in multiparty conversations. *IEEE Transactions on Artificial Intelligence*.

Shivani Kumar, Ramaneswaran S, Md Akhtar, and Tanmoy Chakraborty. 2023c. From multilingual complexity to emotional clarity: Leveraging commonsense to unveil emotions in code-mixed dialogues. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9638–9652, Singapore. Association for Computational Linguistics.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Gary McKeown, Michel F. Valstar, Roderick Cowie, and Maja Pantic. 2010. The semaine corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo*, pages 1079–1084.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions.

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.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, 1.

Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee. 2019. An interaction-aware attention network for speech emotion recognition in spoken dialogs. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6685–6689. IEEE.