MoE4EDiReF: Mixture of Experts for Emotion Discovery and Reasoning its Flip in Conversation

Katarzyna Szczepaniak, Piotr Andruszkiewicz

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

MoE4EDiReF: Mixture of Experts for Emotion Discovery and Reasoning its Flip in Conversation

Katarzyna Szczepaniak
Warsaw University of Technology
katarzyna.szczepaniak2.stud@pw.edu.pl

Piotr Andruszkiewicz
Warsaw University of Technology
IDEAS Research Institute
piotr.andruszkiewicz@pw.edu.pl

Abstract

Modern artificial intelligence systems are increasingly tasked with solving the problem of emotion recognition in conversation, which is a key element in improving the quality of human-computer interaction. However, recognizing emotions in dynamically conducted conversations remains a challenge. This paper proposes a new approach to the problem based on the Mixture of Experts technique, which allows for the simultaneous execution of both emotion recognition and the identification of utterances that cause emotion flips in conversations. The paper presents the results obtained for current approaches and the proposed method. The experiments were conducted on real-world datasets consisting of transcriptions of conversations. The experimental results indicate a significant improvement compared to other solutions.

1 Introduction

The ability to recognize emotions in dynamic conversations is an emerging challenge for modern AI systems and plays a vital role in advancing human-computer interaction. Traditional emotion classification solutions (Abdul-Mageed and Ungar, 2017; Akhtar et al., 2022) rely on static approaches, overlooking the complexity of emotional dynamics, including emotion changes triggered by specific utterances in conversations.

Integrating emotion recognition and identifying utterances that cause emotion changes into a single task supports the development of natural language processing (Kumar et al., 2024). This task involves two main aspects: emotion recognition in conversation (ERC) (Ghosal et al., 2019; Jiao et al., 2019) and emotion-flip reasoning (EFR) (Kumar et al., 2022), which aims to identify the utterances responsible for changing the emotional state of one of the speakers.

The goal of this paper is to develop a model for solving the tasks of emotion recognition and reasoning its flip in conversation. The proposed approach is based on the Mixture of Experts (MoE) technique (Eigen et al., 2013), which enables the simultaneous modelling of both the ERC and EFR tasks. The paper aims to develop a model that achieves results comparable to or better than currently available solutions.

The conducted research focused on four experimental areas:

1. Identifying the impact of the number and type of gating networks, the number of experts, and the type of experts used on the quality of the model,

2. Investigating the effectiveness of activating only the top k experts during model training,

3. Assessing the impact of the learning rate and the number of epochs on the results using the best models selected in the previous stages of the experiments,

4. Examining the effectiveness of translating the dataset from Hindi to English prior to model training.

2 Related Work

This study builds upon a comprehensive review of existing literature, focusing primarily on the task of Emotion Recognition and Reasoning its Flip in Conversation (EDiReF), recently introduced by Kumar et al. (Kumar et al., 2024). This interdisciplinary task combines emotion classification with the identification of utterances that trigger emotional shifts within conversations. While this area remains relatively new in the research community, early work highlights significant challenges, including the modeling of temporal dependencies, participant interactions, and the dynamics of emotional change over time.

Due to the limited number of publications directly addressing EDiReF, the literature review also encompasses broader research on emotion recognition (ER) and, more specifically, emotion recognition in conversation (ERC). ERC methods commonly fall into three categories:

1. **Knowledge-based techniques:** These use lexicons and rule-based systems such as WordNet (Miller, 1994), SenticNet (Cambria et al., 2016), or ConceptNet (Speer et al., 2016). They are interpretable but limited by linguistic ambiguity and poor scalability.

2. **Statistical methods:** Traditional machine learning models, including Support Vector Machines (SVM) (Chavhan et al., 2010) and Naive Bayes classifiers (Sun et al., 2017), perform well on well-structured datasets. More recently, deep learning approaches like RNNs (Li et al., 2021) and transformer-based models such as BERT (Devlin et al., 2019; Bhat, 2024) have become standard due to their context-aware representations and scalability.

3. **Hybrid methods:** These integrate structured knowledge with data-driven models, enhancing performance and interpretability. Notable examples include (Gievska et al., 2015), which combines affective lexicons with deep learning.

Research in ERC typically adopts either supervised or unsupervised approaches. Supervised methods include DialogueRNN (Majumder et al., 2018), DialogueGCN (Ghosal et al., 2019), and newer transformer-based systems like BERT-ERC (Qin et al., 2023) and ERC-DP (Wang et al., 2024). These methods require large annotated corpora but demonstrate state-of-the-art results.

Conversely, unsupervised methods such as clustering techniques (e.g., SCCL (Yang et al., 2023) and DeepEmoCluster (Lin and Busso, 2024)) and probabilistic models like Hidden Markov Models (HMMs) (Nwe et al., 2003; Schuller et al., 2003) are useful for low-resource settings, though generally less accurate.

The EDiReF task itself was first modeled by Kumar et al. (Kumar et al., 2022), who proposed a two-stage architecture combining emotion recognition and flip reasoning. They used a Masked Memory Network (MMN) for utterance-level ERC and a Transformer-based model (TX) for instance-level EFR. MMN hierarchically encodes contextual information from past utterances, while TX models cause-effect relationships across the dialogue. Their system outperformed several baselines, including CMN (Hazarika et al., 2018b), ICON (Hazarika et al., 2018a), DialogueGCN (Ghosal et al., 2019), and AGHMN (Jiao et al., 2019). Additionally, they released MELD-FR, a benchmark dataset derived from MELD, annotated for flip reasoning tasks.

The EDiReF task was introduced as a shared task at SemEval 2024 (Kumar et al., 2024) (Semantic Evaluation is a series of research workshops focused on natural language processing, aimed at advancing knowledge in the field of semantic analysis) with three subtasks: the ERC task on a Hindi dataset (referred to as task A), the EFR task on a Hindi dataset (referred to as task B), and the EFR task on an English dataset (referred to as task C).

This setup allowed for multilingual evaluation and benchmarking. The top systems employed large language models (LLMs) such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT (Radford et al., 2019), and Zephyr (Tunstall et al., 2023), often fine-tuned or instruction-tuned for the task. Approaches like prompting and few-shot learning were also explored. The best-performing systems achieved F1 scores of 0.70, 0.79, and 0.76 for tasks A, B, and C, respectively, indicating their high effectiveness in the context of the task being solved. In the case of task C, this also represents an improvement over the results obtained in the original paper (0.33 for MMN and 0.45 for TX).

3 Methodology

In this section, the idea behind the EDiReF task is explained, followed by a description of the datasets used for this task. Next, the applied method is discussed, and the evaluation methods used to evaluate the performance of the model, which make it possible to accurately measure the effectiveness of the proposed solution and compare it with existing solutions, is presented.

3.1 Problem Definition

In the context of the SemEval 2024 workshop (Kumar et al., 2024), the problem of emotion recognition and the identification of utterances causing emotional changes in conversation was defined as the ERC task and the EFR task on a dataset composed of the phonetic transcriptions of conversations in the Hindi language and on a dataset in

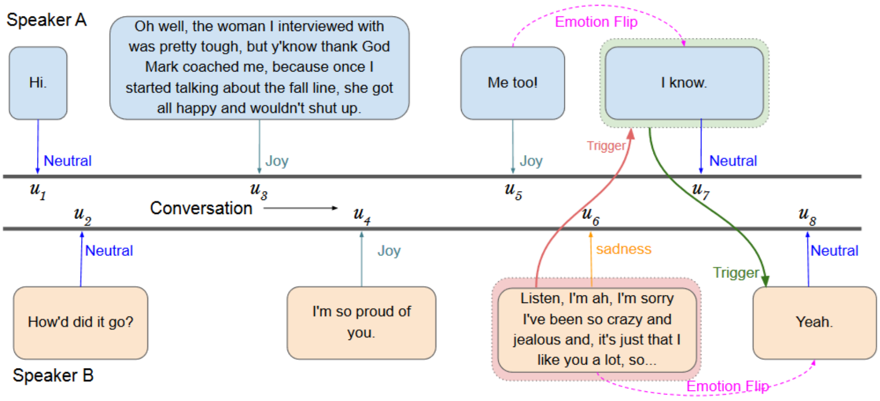
Speaker A
Hi.
Oh well, the woman I interviewed with was pretty tough, but y'know thank God Mark coached me, because once I started talking about the fall line, she got all happy and wouldn't shut up.
Emotion Flip
Me too!
I know.
Neutral
Joy
Joy
Trigger
Neutral
u_1
u_3
u_5
u_7
u_2
Conversation
u_4
u_6
u_8
Neutral
Joy
sadness
Trigger
Neutral
How'd did it go?
I'm so proud of you.
Listen, I'm ah, I'm sorry I've been so crazy and jealous and, it's just that I like you a lot, so...
Yeah.
Speaker B
Emotion Flip

Figure 1: An example of a conversation with assigned emotion labels and utterances marked as contributing to emotional flips (Kumar et al., 2024).

English.

Figure 1 presents an example conversation and the key idea behind the EDiReF task. The figure shows a conversation between two speakers – Speaker A and Speaker B. For each statement (labeled consecutively from u_1 to u_8), an emotion is assigned. Additionally, the statements u_6 and u_7 are highlighted as those causing emotional changes.

The model solving the EDiReF task should assign to each utterance in the provided conversation an emotion label from the set of emotions, as well as a label indicating whether the statement causes an emotional change in the conversation partner.

3.1.1 Emotion Recognition in Conversation

In the ERC task, the model receives textual utterances as input, and its goal is to predict the emotion label for each utterance. Each conversation is represented as a list of tuples $D_{ERC} = \{(s_1, u_1), (s_2, u_2), \ldots, (s_n, u_n)\}$, where s_i denotes the speaker of the utterance u_i. The goal of the model is to predict the emotion e_i for each utterance u_i.

3.1.2 Emotion-Flip Reasoning

In the EFR task, the model analyzes the utterances presented in the form of a list of tuples $D_{EFR} = \{(s_1, u_1, e_1), (s_2, u_2, e_2), \ldots, (s_n, u_n, e_n)\}$, where e_i denotes the emotion present in the utterance u_i spoken by the speaker s_i. The goal is to identify the utterances t_i that trigger an emotion change in the conversation partner and label them as triggers with a value of 1. Otherwise, a value of 0 is assigned.

3.2 Datasets

Two datasets were used to train and verify the models. Both datasets were extended with labels needed for the EFR task. The annotations were prepared by professionals in the field of dataset annotation.

3.2.1 MELD

The Multimodal EmotionLines Dataset (MELD) (Poria et al., 2018) is a dataset in English that is used for the ERC task. In the SemEval 2024 workshop and for the purposes of this work, a modified version of the MELD dataset, i.e. the MELD-FR dataset, was used.

There are seven emotion labels in the dataset – disgust, joy, surprise, anger, fear, neutral and sadness. The neutral label is assigned to those utterances that do not show the emotionality typical of the other emotions.

3.2.2 MaSaC

MaSaC (Bedi et al., 2021) is a set of conversations in Hindi using the phonetic transcription of the language. The conversations are from the TV series *Sarabhai vs Sarabhai* and for the purposes of the SemEval Workshop have been annotated with labels covering eight emotions and labels indicating triggers in the conversations. Seven of the eight labels overlap with labels from the MELD-FR dataset. An additional label in the MaSaC dataset is *contempt*.

3.3 Mixture of Experts Technique

Mixture of Experts (MoE) is a machine learning technique that uses multiple networks, called experts, to divide the problem space into smaller problems (Eigen et al., 2013; Du et al., 2021). During training, each expert is trained, but during testing, depending on the input, only a subset of experts is activated, allowing the trained model to generate answers faster.

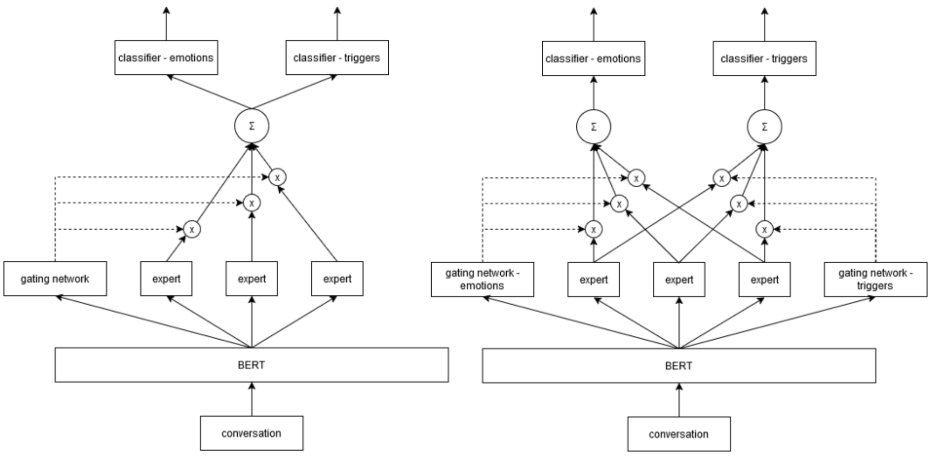

Figure 2: Model architectures with a single (left) and dual (right) gating network used for solving the EDiReF task.

3.3.1 Expert

An expert is a single, specialized model or module that is designed to solve a specific task. Experts operate independently and are responsible for generating results for their part of the problem.

3.3.2 Gating Network

A gating network is a mechanism that decides which experts should be activated for a given input. It works by assigning weights to experts, indicating which of them are best suited to solve a given task.

3.4 Methods of Model Evaluation

The F1 Score was used to assess the quality of the model. The F1 Score is the harmonic mean of precision and recall, balancing these two measures. It is a metric often used in natural language processing to compare solutions.

4 Experimental Setup

This section discusses the model architecture and the set of hyperparameters used in the experiments and presents the details of the experiments conducted to evaluate the effectiveness of the MoE technique in the tasks of recognizing emotions and utterances causing emotion change in conversations.

4.1 Model Architecture

Two model architectures employing the Mixture of Experts technique were designed for the EDiReF task. The first architecture uses a single gating network, while the second incorporates two separate gating networks, each dedicated to a different task: emotion recognition in conversation and identification of utterances that trigger emotional changes. Introducing a second gating network enables a more precise alignment of experts with the specific requirements of each task, potentially improving both emotion recognition and the detection of emotion triggers. Figure 2 (left) illustrates the architecture with a single gating network. In this version, conversations are passed to the model input, where a BERT model generates embedding vectors. These vectors are then sent to both the experts and the gating network, which computes a weight vector used to aggregate the responses from individual experts. The weighted outputs are then combined and forwarded to the classifiers responsible for emotion and trigger classification. Figure 2 (right) shows the dual-gate architecture. Each gating network is assigned to a different task: one to emotion recognition (ERC) and the other to emotion trigger recognition (EFR). This approach allows each gating network to be trained independently in the context of its specific task, enabling a more tailored selection of experts to meet distinct task requirements.

In both datasets, conversations are stored as lists of strings. Since BERT requires a single string as input, all utterances in a conversation are concatenated, separated by the special `[SEP]` token. This enables the model to better capture the dynamics of the conversation.

The preprocessed conversation is then passed to BERT, which produces a vector representation in the embedding space. For the MELD dataset and the translated MaSaC dataset, the `bert-base-cased`[1] model is used. For the original Hindi MaSaC dataset, the

[1] https://huggingface.co/google-bert/bert-base-cased

	Gate type	N. of gates	Expert type	Number of experts	ERC	EFR
M1	Linear	1	Linear	2	88.9	**79.8**
M2	Linear	1	MLP	2	83.8	79.0
M3	Linear	1	LSTM	2	74.2	78.5
M4	MLP	2	Linear	2	**89.6**	79.2
M5	MLP	1	Linear	4	89.2	79.7
M6	MLP	1	LSTM	4	72.8	79.3
M7	MLP	2	Linear	4	89.4	79.4
M8	MLP	2	MLP	4	83.2	79.2

Table 1: Comparison of the results obtained for the MELD dataset and the different setups of gating networks and experts.

	Gate type	Number of gates	Expert type	Number of experts	ERC	EFR
S1	Linear	1	Linear	2	**91.8**	89.5
S2	Linear	1	LSTM	2	65.5	88.2
S3	MLP	1	Linear	2	90.2	89.5
S4	MLP	1	LSTM	4	64.6	**89.8**
S5	MLP	1	Linear	8	90.4	89.2
S6	MLP	1	MLP	8	83.0	89.3
S7	Linear	2	Linear	8	90.6	**89.8**
S8	MLP	2	MLP	8	81.3	89.7

Table 2: Comparison of the results obtained for the MaSaC dataset and the different setups of gating networks and experts.

`bert-base-multilingual-cased`[2] model is employed.

The gating networks and experts may take the form of a linear layer or a multilayer perceptron. Experts can also be implemented using long short-term memory (LSTM) networks.

4.2 Hyperparameters

In all experiments we used the same set of the following hyperparameters, which ensures consistency of results and enables direct comparison of performance of different model configurations. The batch size was set to 32. To reduce overfitting, dropout regularization with a value of 0.1 was applied. The loss function for the emotion recognition task was CrossEntropyLoss. For the trigger identification task, BCEWithLogitsLoss was used.

4.3 Conducted Experiments

In order to investigate the effectiveness of the Mixture of Experts technique, three stages of experiments were designed. Each stage examines different aspects of the MoE technique. Additionally, it was also assessed whether translating the MaSaC dataset into English before training improves the results for the final model. Furthermore, the performance of the best models with that of other existing approaches was compared. The F1 measure in the form of a percentage value was used as a criterion for evaluation and comparison.

4.3.1 Impact of Type and Number of Gating Networks and Experts

During the first stage of experiments, it was examined how the quality of the model is affected by the use of either one or two gating networks with different architectures. A single linear layer and a multilayer perceptron were investigated.

The influence of different numbers of experts with different architectures was also examined. A single linear layer, a multilayer perceptron and a long short-term memory network were investigated as possible expert architectures. The choice of the number of experts was motivated as follows: with two experts, each could specialize in one task (ERC or EFR). In the case of four experts, one could handle EFR, while the rest of the experts would handle ERC divided into positive, negative and neutral emotions. With eight experts, one could be assigned to EFR, and the rest to individual emotions

[2] https://huggingface.co/google-bert/bert-base-multilingual-cased

	Gate type	Number of gates	Expert type	Number of experts	Top k	ERC	EFR
M9	MLP	1	Linear	4	1	86.8	**80.1**
M10	MLP	1	Linear	4	2	87.9	79.6
M5	MLP	1	Linear	4	4	**89.2**	79.7
M11	MLP	2	Linear	4	1	88.2	**79.6**
M12	MLP	2	Linear	4	2	88.4	79.1
M7	MLP	2	Linear	4	4	**89.4**	79.4

Table 3: Comparison of the results obtained for the MELD dataset and the different number of experts activated during training.

	No.G.	No.E.	Top k	ERC	EFR
S9	1	2	1	90.8	89.0
S1	1	2	2	**91.8**	**89.5**
S10	2	8	1	88.5	89.5
S11	2	8	2	90.4	89.4
S12	2	8	4	88.4	89.4
S7	2	8	8	**90.6**	**89.8**

Table 4: Comparison of the results obtained for the MaSaC dataset and the different number of experts activated during training. All configurations with linear gate type and linear expert type. No.G. - Number of Gates, No.E. - Number of Experts.

in ERC. In this set of experiments, the models were trained for 5 epochs with a learning rate of $2\cdot10^{-5}$.

Selected results obtained within the first set of experiments investigating the impact of type and number of gating networks and experts on the MELD and MaSaC datasets are presented in Table 1 and 2, respectively. The best results for ERC and EFR are marked in bold.

Models M4 and M1, which achieved the highest F1 scores for ERC and ERF, respectively, utilized a double (M4) and a single (M1) gating network, implemented as a multilayer perceptron and four experts implemented as individual linear layers. The remaining four best-performing model configurations, two per task, also used linear layers as experts, suggesting that for the characteristics of the MELD dataset, using simpler experts was beneficial. No significant difference was observed between approaches employing dual gating networks and those using a single gate, nor was there an improvement in performance when using a more complex gating architecture such as a multilayer perceptron.

The results on the MaSaC dataset support the hypothesis that when using the mixture of experts technique, a very simple architecture—such as a linear layer—is sufficient to achieve strong performance (F1 score of 91.8 in the case of S1 for ERC and 89.5 for EFR). An interesting difference between the results for the MELD and MaSaC datasets is the relationship between the number of experts used and the achieved performance. In the case of MaSaC, a significant number of high-performing models employed 8 experts, whereas for the MELD dataset, a smaller number of experts was preferred (with 4, and in some cases even just 2, experts yielding better results).

4.3.2 Impact of Activating Top k Experts During Training

To compare the efficiency of activating the top k experts, i.e. limiting the number of activated experts, two models, which achieved the best result in the first set of experiments, were selected for each dataset.

The summary of the results on the MELD dataset is presented in Table 3. The results for the MaSaC dataset are presented in Table 4.

Regardless of the model configuration, those that activated all available experts during training achieved better results than models that activated only the top-k experts. Additionally, no improvement in training speed was observed when using top-k activation, and thus this approach is not recommended.

Similarly, for the MaSaC dataset, the use of top-k expert activation did not yield any benefits in terms of higher F1 scores or shorter training times.

4.3.3 Impact of Learning Rate and Number of Epochs

In the final stage of the experiments, three model configurations that had so far achieved the best results were used and the influence of learning rates of $2\cdot10^{-5}$, $3\cdot10^{-5}$, $4\cdot10^{-5}$ and $5\cdot10^{-5}$ and the number of epochs of 3, 5 and 7 was investigated. These parameters were selected based on the recommendations from the article on the BERT model (Devlin

	L. rate	N. of epochs	ERC	EFR
M13	$2\cdot10^{-5}$	3	74.9	78.8
M14	$5\cdot10^{-5}$	3	87.9	79.4
M5	$2\cdot10^{-5}$	5	89.2	**79.7**
M15	$5\cdot10^{-5}$	5	93.3	78.1
M16	$2\cdot10^{-5}$	7	92.1	79.4
M17	$3\cdot10^{-5}$	7	**93.9**	79.4
M18	$5\cdot10^{-5}$	7	**93.9**	78.9

Table 5: Comparison of the results obtained for the MELD dataset and the different learning rates and number of epochs.

	L. rate	N. of epochs	ERC	EFR
S13	$2\cdot10^{-5}$	3	64.3	**90.0**
S14	$5\cdot10^{-5}$	3	80.3	89.3
S1	$2\cdot10^{-5}$	5	91.8	89.5
S15	$5\cdot10^{-5}$	5	96.6	89.7
S16	$2\cdot10^{-5}$	7	95.8	**90.0**
S17	$4\cdot10^{-5}$	7	**97.7**	89.4
S18	$5\cdot10^{-5}$	7	97.0	89.5

Table 6: Comparison of the results obtained for the MaSaC dataset and the different learning rates and number of epochs.

et al., 2019), indicating them as some of the optimal ones.

The summary of the results on the MELD dataset is given in Table 5. The model used in the experiments had a single MLP-type gating network and 4 linear experts, which all were activated during training.

The results obtained on the MaSaC dataset are shown in Table 6. The model used in the experiments had a single linear gating network and 2 linear experts, both of which were activated during training.

Models M15 and M16, which were trained for 3 epochs, achieved the worst results among all considered configurations. In particular, model M15, trained with a learning rate of $2 \cdot 10^{-5}$, had the lowest performance, reaching an F1 score of 76.85. Models M17 and M18, trained for 7 epochs with learning rates of $3 \cdot 10^{-5}$ and $5 \cdot 10^{-5}$ respectively, achieved similar results – the average F1 score for the ERC task was 93.9 for both models, and the average F1 score for the EFR task differed by only 0.5 percentage points.

It can therefore be concluded that increasing the number of epochs from 5 to 7, as well as raising the learning rate from $2 \cdot 10^{-5}$ to $3 \cdot 10^{-5}$ or even $5 \cdot 10^{-5}$, has a positive effect on the final results.

For the MaSaC dataset, as well as for the MELD dataset, models trained for 7 epochs generally achieved better results than those trained for only 3 or 5 epochs, regardless of the learning rate value.

4.3.4 Impact of Translating MaSaC Dataset into English

The pyhinavrophonetic[3] library and a variant of the NLLB-200 model developed by Facebook were used to translate the MaSaC dataset into English. The script iterated through the MaSaC dataset, converting each conversation from the phonetic script of the Hindi language to the official alphabet, and then passed it to the NLLB-200 model, which can translate text between multiple languages.

	No.G.	No.E.	Top k	ERC	EFR
MaSaC					
T1	1	2	2	97.7	89.4
T2	2	8	2	97.6	89.4
T3	2	8	2	**98.2**	88.8
T4	2	8	8	97.3	**89.7**
Translation of MaSaC					
T1	1	2	2	96.7	87.9
T2	2	8	2	96.7	88.5
T3	2	8	2	96.7	88.5
T4	2	8	8	96.7	87.6

Table 7: Comparison of the results obtained for the MaSaC dataset and its translation. All configurations with linear gate type and linear expert type. No.G. - Number of Gates, No.E. - Number of Experts.

Four best model configurations from the previous experiments were used. Each model was trained for 7 epochs. Models T1, T3, and T4 with a learning rate of $4\cdot10^{-5}$, and model T2 with a learning rate of $3\cdot10^{-5}$. The results are summarized in Table 7.

For both datasets, the models achieved very good results. However, when comparing paired configurations (i.e., T1 on the MaSaC dataset and T1 on the translated version of MaSaC), it becomes evident that the preliminary translation of the Hindi dataset into English did not yield significant benefits. Each model trained on the translated MaSaC dataset achieved lower average F1 scores for the ERC task, the F1 score for the EFR task, or the average F1 score across both tasks.

[3] Available at: https://pypi.org/project/pyhinavrophonetic/, Last accessed: 08.01.2025.

	ERC								EFR
	Dg	Jy	Sr	An	Fr	Ne	Sa	Avg	Trigger
MMN	20,2	48,7	50,4	42,9	9,80	71,9	29,6	55,7	33,4
TX	0,00	4,00	5,00	1,90	0,00	61,2	0,00	29,5	44,8
GAVx	-	-	-	-	-	-	-	-	76,0
MoE-1	84,9	93,4	93,6	**91,8**	90,4	95,0	**94,4**	93,9	79,4
MoE-2	**86,1**	**93,8**	**94,0**	91,4	**92,8**	**95,2**	92,8	**94,0**	**79,7**

Table 8: Comparison of the results obtained for the MELD dataset (Dg: disgust, Jy: joy, Sr: surprise, An: anger, Fr: fear, Ne: neutral, Sa: sadness).

	ERC									EFR
	Dg	Jy	Sr	An	Fr	Ne	Sa	Co	Avg	Trigger
TW-NLP	-	-	-	-	-	-	-	-	46.0	79.0
FeedForward	-	-	-	-	-	-	-	-	51.0	77.0
UCSC NLP	-	-	-	-	-	-	-	-	45.0	79.0
MoE-3	**98.0**	97.7	**96.8**	97.2	96.7	98.4	96.0	96.2	97.7	**89.4**
MoE-4	96.5	**98.4**	96.5	**97.4**	**97.9**	**98.8**	**97.2**	**97.2**	**98.2**	88.8

Table 9: Comparison of the results obtained for the MaSaC dataset (Dg: disgust, Jy: joy, Sr: surprise, An: anger, Fr: fear, Ne: neutral, Sa: sadness, Co: contempt).

4.3.5 Comparison of the Results for the MELD Dataset

The results obtained by Kumar et al. (Kumar et al., 2022), the GAVx team (Nguyen and Zhang, 2024) that achieved the best results in the SemEval 2024 workshop for task C, and our own results (models MoE-1 and MoE-2) are summarized in Table 8.

MoE-1 uses a single MLP-type gating network and 2 linear experts, and MoE-2 uses 2 MLP-type gating networks and 4 linear experts. Both models were trained for 7 epochs with a learning rate of $3 \cdot 10^{-5}$.

The comparison shows that, for the ERC task, each of the proposed models outperformed the models introduced by Kumar et al. Unfortunately, it was not possible to compare these results with those of the GAVx teams due to the unavailability of their scores.

However, comparison is possible for the EFR task. In this case, the proposed models also achieved better results than those reported in the original paper introducing the EDiREF task. Furthermore, they outperformed the top participants of the SemEval 2024 workshop, exceeding the first-place result by at least 3.1 percentage points.

4.3.6 Comparison of the Results for the MaSaC Dataset

The results obtained by the TW-NLP (Tian et al., 2024), FeedForward (Shaik et al., 2024) and UCSC NLP (Wan et al., 2024) teams, which achieved the best results in the SemEval 2024 workshop for tasks A and B, and our own results (models MoE-3 and MoE-4) are presented in Table 9.

MoE-3 uses a single linear gating network and 2 linear experts, and MoE-4 uses 2 linear gating networks and 8 linear experts, of which the best 2 were activated during training. Both models were trained for 7 epochs with a learning rate of $4 \cdot 10^{-5}$.

For the MaSaC dataset, it is not possible to compare classification results for individual emotions; however, it is possible to compare the average F1 scores for the ERC and EFR tasks. In both cases, each of the proposed models achieved better results than the best-performing model presented at the SemEval 2024 workshop. For ERC, the difference is at least 46.3 percentage points, while for EFR, the difference amounts to 9.8 percentage points.

5 Conclusions

The presented experimental results show that the use of the Mixture of Experts technique is an effective solution for the task of recognizing emotions and reasoning its flip in conversation. The solution achieved higher results than the current solutions, which means that the intended goal of the work was achieved.

It was observed that using a smaller number of experts with a simpler architecture, such as a linear layer or a simple multilayer perceptron, and training the model for a larger number of epochs with a lower learning rate has a positive effect on the final performance of the model. It was not observed that changing the number or type of gating network or increasing the number of experts significantly improved the results. However, using the activation of top k experts during training worsened the result, so it is not recommended to use this solution.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. 2022. All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE Transactions on Affective Computing*, 13:285–297.

Manjot Bedi, Shivani Kumar, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, 14:1363–1375.

Siddhanth Bhat. 2024. Emotion classification in short english texts using deep learning techniques. *ArXiv*, abs/2402.16034.

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Bjoern Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2666–2677, Osaka, Japan. The COLING 2016 Organizing Committee.

Yashpalsing Chavhan, Manikrao Dhore, and Yesaware Pallavi. 2010. Speech emotion recognition using support vector machines. *International Journal of Computer Applications*, 1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, and 8 others. 2021. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. *CoRR*, abs/1312.4314.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. 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 154–164, Hong Kong, China. Association for Computational Linguistics.

Sonja Gievska, Kiril Koroveshovski, and Tatjana Chavdarova. 2015. A hybrid approach for emotion detection in support of affective interaction. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2015:352–359.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2019. Real-time emotion recognition via attention gated hierarchical memory network. In *AAAI Conference on Artificial Intelligence*.

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 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1933–1946, Mexico City, Mexico. 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.

Dongdong Li, Jinlin Liu, Zhuo Yang, Linyu Sun, and Zhe Wang. 2021. Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert Systems with Applications*, 173:114683.

Wei-Cheng Lin and Carlos Busso. 2024. Deep temporal clustering features for speech emotion recognition. *Speech Communication*, 157:103027.

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*, abs/1907.11692.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and E. Cambria. 2018. Dialoguernn: An attentive rnn for emotion detection in conversations. In *AAAI Conference on Artificial Intelligence*.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Vy Nguyen and Xiuzhen Zhang. 2024. GAVx at SemEval-2024 task 10: Emotion flip reasoning via stacked instruction finetuning of LLMs. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 326–336, Mexico City, Mexico. Association for Computational Linguistics.

Tin Nwe, S.W. Foo, and Liyanage De Silva. 2003. Speech emotion recognition using hidden markov models. *Speech Communication*, 41:603–623.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, E. Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *ArXiv*, abs/1810.02508.

Xiangyu Qin, Zhiyu Wu, Jinshi Cui, Ting Zhang, Yanran Li, Jian Luan, Bin Wang, and L. xilinx Wang. 2023. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. *ArXiv*, abs/2301.06745.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI blog*.

Björn Schuller, Gerhard Rigoll, and Manfred Lang. 2003. Hidden markov model-based speech emotion recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2:401–404.

Zuhair Hasan Shaik, Dhivya Prasanna, Enduri Jahnavi, Rishi Thippireddy, Vamsi Madhav, Sunil Saumya, and Shankar Biradar. 2024. FeedForward at SemEval-2024 task 10: Trigger and sentext-height enriched emotion analysis in multi-party conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 745–756, Mexico City, Mexico. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*.

Shiliang Sun, Chen Luo, and Junyu Chen. 2017. A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36:10–25.

Wei Tian, Peiyu Ji, Lei Zhang, and Yue Jian. 2024. TW-NLP at SemEval-2024 task10: Emotion recognition and emotion reversal inference in multi-party dialogues. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 311–315, Mexico City, Mexico. Association for Computational Linguistics.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *ArXiv*, abs/2310.16944.

Neng Wan, Steven Au, Esha Ubale, and Decker Krogh. 2024. UCSC NLP at SemEval-2024 task 10: Emotion discovery and reasoning its flip in conversation (EDiReF). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1492–1497, Mexico City, Mexico. Association for Computational Linguistics.

Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. Emotion recognition in conversation via dynamic personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5711–5722, Torino, Italia. ELRA and ICCL.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, 14:3269–3280.