ISDS-NLP at SemEval-2024 Task 10: Transformer based neural networks for emotion recognition in conversations

Claudiu Creangă[2,3,], **Liviu P. Dinu**[1,3]
[1] Faculty of Mathematics and Computer Science
[2] Interdisciplinary School of Doctoral Studies, [3] HLT Research Center
University of Bucharest, Romania
claudiu.creanga@s.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

This paper outlines the approach of the ISDS-NLP team in the SemEval 2024 Task 10: Emotion Discovery and Reasoning its Flip in Conversation (EDiReF). For Subtask 1 we obtained a weighted F1 score of 0.43 and placed 12 in the leaderboard. We investigate two distinct approaches: Masked Language Modeling (MLM) and Causal Language Modeling (CLM). For MLM, we employ pre-trained BERT-like models in a multilingual setting, fine-tuning them with a classifier to predict emotions. Experiments with varying input lengths, classifier architectures, and fine-tuning strategies demonstrate the effectiveness of this approach. Additionally, we utilize Mistral 7B Instruct V0.2, a state-of-the-art model, applying zero-shot and few-shot prompting techniques. Our findings indicate that while Mistral shows promise, MLMs currently outperform them in sentence-level emotion classification.

1 Introduction

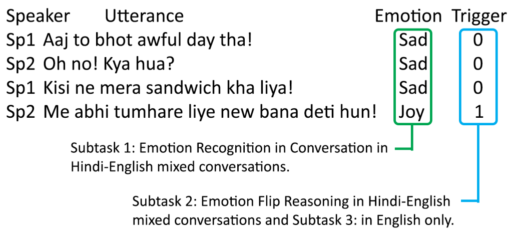
Task 10 from SemEval 2024 competition (Kumar et al., 2024) addresses the complex challenge of identifying the emotions within dialogues (English and Hindi). This task comprises two primary objectives: firstly, assigning an emotion label to each utterance within a dialogue, and secondly, discerning the trigger utterance or utterances responsible for an emotion-flip within the dialogue (Kumar et al., 2022). Emotions play a crucial role in human interaction and one can understand more from a text if one knows the underlying sentiment of the writer. In contexts where disagreements may arise, such as customer service platforms, virtual assistant chats or forums, identifying trigger utterances for emotion flips can help mediate conflicts and prevent escalation. A chatbot dealing with an angry customer would benefit from knowing how to speak in order to generate empathetic responses. If it knows that the chatbot's current sentence can trigger an emotion flip from neutral to anger, the chatbot should refine it, or if the emotion flip is from anger to joy, the chatbot should be more confident in such a response in the future.

Both types of models we tried for Subtask 1 were based on transformers. The first one used BERT-like models and we achieved the best accuracy with them, while the second one is a state of the art causal model (Mistral, (Jiang et al., 2023)) that was tested in zero-shot and few-shot settings with poorer results.

Although in the first task our system worked well, placing 12th in the leaderboard, the other 2 tasks were much harder and we placed 14th on the second subtask. We believed that with a better strategy to prevent overfitting (like under or oversampling), our system would have improved. Our code is open source and available to use on GitHub.

2 Background

The competition had 3 subtasks explained in Figure 1 and we participated in all of them with the best results on subtask 1 where we placed 12th with an F1 score of 0.43.



Speaker	Utterance	Emotion	Trigger
Sp1	Aaj to bhot awful day tha!	Sad	0
Sp2	Oh no! Kya hua?	Sad	0
Sp1	Kisi ne mera sandwich kha liya!	Sad	0
Sp2	Me abhi tumhare liye new bana deti hun!	Joy	1

Subtask 1: Emotion Recognition in Conversation in Hindi-English mixed conversations.

Subtask 2: Emotion Flip Reasoning in Hindi-English mixed conversations and Subtask 3: in English only.

Figure 1: Three sub-tasks explained

2.1 Dataset

The dataset contains English and Hindi code-mixed conversations for Subtask 1 and 2 and English only conversations for Subtask 3 (Table 1). The dataset is quite small, except for the training dataset for

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 649–654
June 20-21, 2024 ©2024 Association for Computational Linguistics

Table 1: Datasets sizes used in this competition by tasks.

	Subtask 1	Subtask 2	Subtask 3
Train	8506	98777	35000
Dev	1354	7462	3522
Test	1580	7690	8642

Subtask 2 and 3. If we were to combined them for Subtask 1, our F1 score would reach 0.97, but it wasn't allowed. This fact shows that with more data our model would do really well. The dataset is based on MELD, a known emotion recognition dataset, which was then augmented with triggers for the emotion-flip task.

There were 8 distinct emotions to predict: neutral, anger, surprise, fear, joy, sadness, disgust, and contempt. By far the most predominant emotion is neutral, followed by joy and anger (Figure 2). If we look at Subtask 2, most often the emotion flips are from neutral to joy or anger (Figure 3).

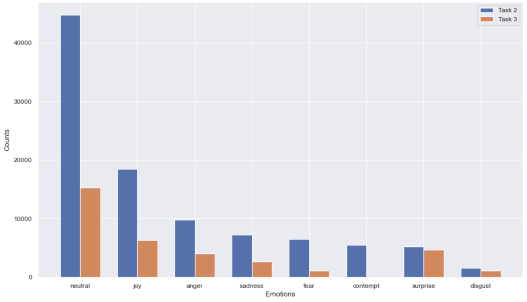


Figure 2: Emotion Distribution Comparison between Task 2 and Task 3

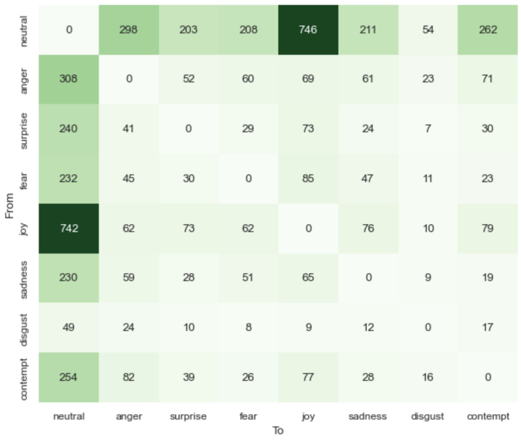


Figure 3: Task 2: Emotion-flip counts

2.2 Previous Work

Since the release of the first small datasets for emotion recognition in 1992 (Ekman, 1992), the field has evolved substantially, marked by significant contributions from big companies in the form of extensive datasets (Demszky et al., 2020). In the beginning, lexicon based methods were used in which there was a manually curated dictionary which associates words with specific emotions. The algorithm was simply picking the most expressed emotion according to the dictionary. This method had severe limitations because it was ignoring context, sentence structure and negations which can flip a sentiment. Today, state of the art models are based on transformer architecture and use either Masked Language Modelling (BERT based models (Devlin et al., 2018)) or Causal Modelling (GPT (Brown et al., 2020)) which can capture dependencies and nuances missed by word-level approaches.

While traditional emotion recognition tasks are well-established, research on emotion flip recognition is still in its early stages because it is a new task within the field of emotion analysis. Research (Kumar et al., 2021) has found that a transformer based classifier with 6 encoder layer (EFR-TX) works well, obtaining an F1 score of 40 when trained on MELD-FR dataset and tested on IEMOCAP-FR dataset.

3 System overview

We tried two approaches, both of them based on transformer architecture: Masked Language Modelling and Casual Modelling. We chose these two architectures because of their recent successes in NLP.

3.1 Masked Language Modelling

We used pre-trained BERT-like models in a multilingual setting so that it can tokenise Hindi sentences. These pre-trained models will give us the features from sentences and then we pass them through a classifier which will do the prediction for each task Figure 4.

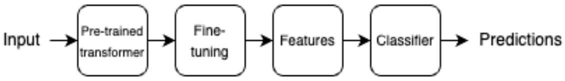


Figure 4: Model architecture

3.1.1 Input

Analysis of dialogue sentences reveals a predominantly short length, with a sharp decline in frequency after 30 tokens (see Figure 5). To optimize performance, various maximum sequence lengths were tested, with 55 tokens yielding the best results (Figure 6). Data preprocessing, (such as lemmatization, removing punctuation or stopwords) didn't help the model learn better so we kept the input as is. Probably this is because punctuation and stopwords contain useful information that the models is able to learn.

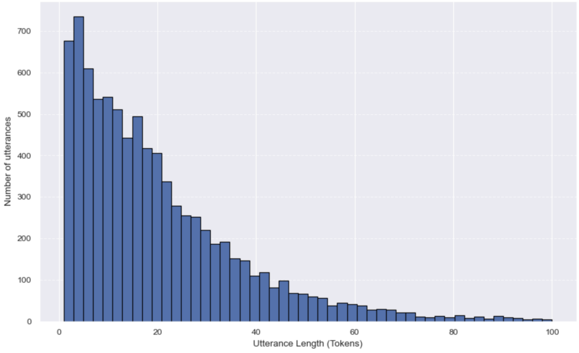


Figure 5: Distribution of utterances lengths.

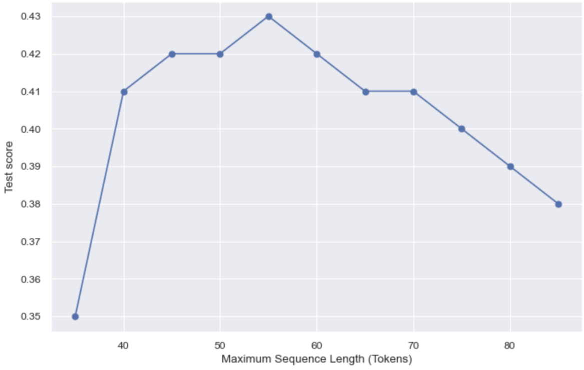


Figure 6: Best model score with different maximum utterances lengths.

3.1.2 Output

Selecting the optimal hidden state layer is crucial for leveraging the pre-trained model's results. Our experiments demonstrated that using the final layer's output yielded the strongest performance, with accuracy declining in earlier layers. For MLM-type models, the [CLS] token encodes the features, which is what we pass to our classification layer.

Among various classifiers tested (Table 2), fully connected layers excelled, likely due to their ability to model complex, non-linear relationships. The top-performing model employed a fully connected layer with 0.5 dropout and a Softmax activation function.

Table 2: Test scores of different classifiers.

Classifier	Extra features	Score
Fully Connected	Dropout(0.5)	0.43
Fully Connected	Dropout(0.7)	0.42
Fully Connected	Dropout(0.2)	0.40
Fully Connected	-	0.40
RandomForest	-	0.23
LogisticRegression	-	0.21
KNeighbors	-	0.20

3.1.3 Fine-tuning

The large pre-trained language models we employed offer a robust foundation for understanding language in general. Through fine-tuning, we adapt them to the nuances of our emotion recognition task. Inspired by the strategy presented in (Sun et al., 2020), we initially train only the classifier with a larger learning rate (5e-5) and a warm-up period of 10,000 steps over 'k' epochs (we tried a range of 'k' from 1 to 10). Subsequently, we fine-tune both the classifier and the transformer's final layer using a smaller learning rate (2e-5). Our goal in freezing the transformer weights at first, and then training them with a reduced learning rate, is to minimize the risk of overfitting.

3.2 Causal Modelling

Given the success of generative models we also tried Mistral 7B Instruct V0.2 which is believed to be state of the art in its category of models (Jiang et al., 2023). These type of LLMs have had success in a large number of NLP tasks, but seem to still lag Masked Language Models in sentence classification.

3.2.1 Prompting

In Causal Modelling, how you prompt the model significantly influences its performance. We tested different prompting strategies in both zero-shot and few-shot settings:

- **Zero-Shot Learning:** Here, we provide the model with a single example and ask it to predict the emotion without any additional references. For zero-shot learning the best prompting technique was: "[INS] Given the following sentence: `{sentence}. ###` Predict which emotion is expressed. Chose one of the following options: neutral, anger, surprise, fear, joy, sadness, disgust, and contempt. Answer in one word only. Answer: [\INS]"

- **Few-Shot Learning:** In this setting, we give the model several examples – one for each emotion – along with their corresponding labels. This leverages the model's in-context learning ability, potentially boosting its performance for unseen samples. For few-shot learning the best prompting technique was: "[INS] This is an example of a sad sentence: `{sentence} {repeat for every emotion}`. ### Predict the emotion of the following sentence: sentence. Chose one of the following options: neutral, anger, surprise, fear, joy, sadness, disgust, and contempt. Answer in one word only. Answer: [\INS]"

4 Experimental setup

4.1 Data Split Strategy

We employed a classic data split approach:

- **Initial Development**: We combined the training and development sets and shuffled the data. Subsequently, we used 70% for training, 10% for validation, and the remaining 20% as a held-out test set.

- **Competition Test Set Release**: Upon the competition's test set release, we directly evaluated our models using the platform. To maximize training data, we trained on the combined training set with a 20% validation split.

- **Final Model**: Once we selected our best model, we re-trained it on the entire dataset without validation. This re-training didn't yield significant improvements

4.2 Subtask 1

We'll focus on Subtask 1, where we achieved strong results. The key hyperparameters used:

- **Batch Size**: A batch size of 64 provided the best balance. Smaller sizes hurt performance, while larger sizes exceeded our memory constraints.

- **Fine-Tuning**: We trained for 4 epochs with frozen model weights, followed by 3 epochs with only the last layer unfrozen (as detailed in section 3.1.3).

- **Classifier**: Our classifier used 128 neurons, 0.5 dropout, and a softmax activation.

- **Optimization**: We used cross-entropy loss, the AdamW optimizer, and experimented with different learning rates (see section 3.1.3).

- **Evaluation**: We measured performance using the MulticlassF1Score with 8 classes and 'macro' averaging.

5 Results

Our top-performing model (Table 3) was a fine-tuned FacebookAI/xlm-roberta-large (Conneau et al., 2019). This highlights the superiority of fine-tuned Masked Language Models (MLMs) over Mistral for sentence classification tasks. The results suggest that smaller Causal models remain less effective than fine-tuned MLMs in this domain. We also see that few-shot Mistral is worse than zero-shot, probably because too much data in the prompt confuses the model.

Model	Train	Validation	Test
xlm-roberta	0.74	0.57	**0.43**
mdeberta-v3	0.74	0.56	0.42
bert-multi	0.66	0.48	0.35
Mistral zero-shot	-	-	0.32
Mistral few-shot	-	-	0.31
distilbert-multi	0.6	0.47	0.29

Table 3: Results for Subtask 1 - Masked Language Models and Causal Models (Mistral).

In terms of number of epochs, our best model was overfitting when finetuned for too many epochs (Table 4) and we finally trained for 4 + 3 epochs.

Frozen	Fine-tunning	Training	Validation	Test
3	2	0.67	0.53	0.4
3	3	0.71	0.55	0.42
4	3	0.74	0.57	**0.43**
4	4	0.78	0.60	0.42
4	5	0.85	0.45	0.38
5	3	0.71	0.56	0.42

Table 4: Finding the optimal number of epochs to avoid overfitting. First column contains epochs when training only the classifier. Second columns contains epochs when training the classifier and the last transformer layer.

5.1 Error analysis

Our confusion matrix (Figure 7) reveals that the model overpredicts the ’neutral’ emotion, likely due to its prevalence in the training data. This created a bias, leading the model to misclassify instances of other emotions as ’neutral’. While we attempted to mitigate this with class weights in the loss function, it proves insufficient. In the future, we should explore more robust techniques like oversampling or undersampling to address the class imbalance.

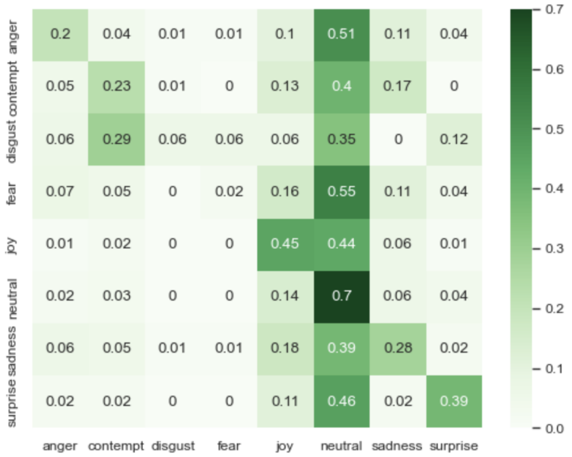


Figure 7: Confusion matrix. On y-axis true labels, on x-axis predicted labels. Values are normalised.

As seen in the emotion accuracy chart (Figure 8), the model performs best on the dominant ’neutral’ class, along with well-represented emotions like ’joy’ and ’sadness’. Conversely, the model struggles to predict the ’disgust’ emotion, which aligns with its under-representation in the training data. This suggests a direct correlation between dataset frequency and model proficiency for each emotion.

6 Conclusion

Overall, our system achieved encouraging results in Subtask 1, despite exhibiting some overfitting for dominant labels. While performance on the emotion-flip detection tasks (Subtasks 2 and 3) highlights areas for improvement, we still placed in the first half of the leaderboard. Looking ahead, we plan to investigate hybrid transformer-LSTM architectures for a more nuanced understanding of emotion-flip triggers. Additionally, enriching the data by incorporating a broader conversational context through multi-turn analysis could enhance our model’s capabilities. Not least, even though we tried Mistral, there are newer causal models like Mixtral (Jiang et al., 2024) and Solar (Kim et al., 2023) which could perform better at this type of task.

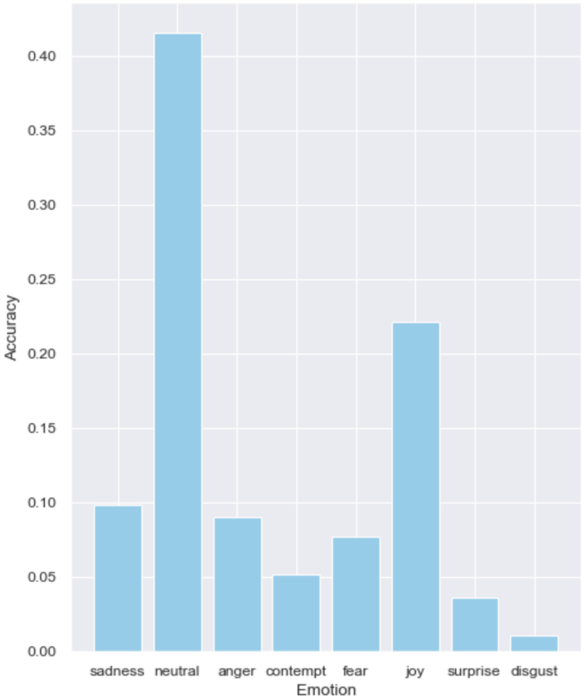


Figure 8: Accuracy by emotion. Accuracy directly correlates with the frequency of each emotion in the training set.

Acknowledgements

This work was partially supported by a grant on Machine Reading Comprehension from Accenture Labs and by the POCIDIF project in Action 1.2. “Romanian Hub for Artificial Intelligence”.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. SOLAR 10.7B: Scaling large language models with simple yet effective depth upscaling.

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, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer.

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.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?