BITS Pilani at SemEval-2024 Task 10: Fine-tuning BERT and Llama 2 for Emotion Recognition in Conversation

Dilip Venkatesh[1], **Pasunti Prasanjith**[1], and **Yashvardhan Sharma**[1]

[1]Birla Institute of Technology and Science, Pilani, Rajasthan, India
Email: {f20201203, pasunti.prasanjith, yash}@pilani.bits-pilani.ac.in

Abstract

Emotion Recognition in Conversation **(ERC)** aims to assign an emotion to a dialogue in a conversation between people. The first subtask of EDiReF shared task aims to assign an emotions to a Hindi-English code mixed conversation. For this, our team proposes a system to identify the emotion based on fine-tuning large language models on the `MaSaC` dataset. For our study we have fine tuned 2 LLMs **BERT** and **Llama 2** to perform sequence classification to identify the emotion of the text.

1 Introduction

Emotion can be defined as a conscious mental reaction subjectively experienced as strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body (Merriam-Webster, 2024). In recent times emotion recognition and sentiment analysis has become increasingly popular due to the research developments in natural language processing. Although similar to sentiment analysis, while sentiment analysis aims to classify text as `POSITIVE`, `NEGATIVE` and `NEUTRAL`, ERC aims to identify text as more in-depth emotions like joy, sadness, anger, contempt etc.

Emotion recognition has multiple use cases in the real world. Opinion mining of conversational data posted by users is done at a large scale at big tech companies. Poria et al. (2019) mentions that ERC has major potential to be used in healthcare systems for psychological analysis and education to understand student frustrations. It is important for language models and chat bots to understand the sentiment of an input text to respond accordingly and generate empathetic dialogue systems (Ma et al., 2020).

For the first subtask of the SemEval 2024 Task 10: *Emotion Discovery and Reasoning its Flip in Conversation (EDiReF)* (Kumar et al., 2024) on CodaLab (Pavao et al., 2023), we aim to conduct Emotion Recognition in Conversation on a Hindi-English code-mixed dataset. Our team proposes a system for this where we fine tune two large language models. Namely the transformer based BERT (Devlin et al., 2019) and Llama 2 (Touvron et al., 2023b).

All of our code can be found on GitHub at `github.com/dipsivenkatesh/SemEval-2024-Task-10`

2 Background

2.1 Task and Data Description

The EDiRef shared task[1] consists of three subtasks.

- Emotion Recognition in Conversation **(ERC)** in Hindi-English code-mixed conversations
- Emotion Flip Reasoning **(EFR)** in Hindi-English code-mixed conversations
- **EFR** in English conversations

In this paper we go through our team's system to solve the first sub task.

The first subtask is to perform ERC on the Hindi-English code-mixed `MaSaC` dataset proposed in Bedi et al. (2023). The dataset comprises of around 1,200 multi-party dialogues from the popular Indian TV show *'Sarabhai vs Sarabhai'*[2] and around 15,000 utterance exchanges (primarily in Hindi) between the speakers. The dataset consisted of the utterances by the speaker and the corresponding emotion label given to each utterance. The emotions were *anger, neutral, contempt, sadness, fear, disgust, joy* and *surprise*.

An example of Emotion recognition in conversation can be found in Table 1

[1]https://codalab.lisn.upsaclay.fr/competitions/16769
[2]https://www.imdb.com/title/tt1518542/

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 811–815
June 20-21, 2024 ©2024 Association for Computational Linguistics

Speaker	Utterance	Emotion
Sp1	Aaj to bhot awful day tha! (I had an awful day today!)	Sad
Sp2	Oh no! Kya hua? (Oh no! What happened?)	Sad
Sp1	Kisi ne mera sandwich kha liya! (Somebody ate my sandwich!)	Sad
Sp2	Me abhi tumhare liye new bana deti hun! (I can make you a new one right now!)	Joy
Sp1	Wo great hoga! Thanks! (That would be great! Thanks!)	Joy

Table 1: Hindi-English code-mixed conversation with emotions

2.2 Previous Work

Initially the naive Bayes algorithm was used for subject classification (Maron, 1961), specifically for sentiment analysis the variant, binary multinomial naive Bayes algorithm was proposed. More recently, the way to perform classification tasks in natural language processing is through supervised machine learning.

Hazarika et al. (2018b) proposes a conversational memory network (CMN), a method that uses memories to capture inter-speaker dependencies. This was further improved with Interactive Conversational memory Network (ICON) a multimodal method that models the self- and inter-speaker emotional influences into global memories (Hazarika et al., 2018a). The Interaction-Aware Attention Network (IANN) (Yeh et al., 2019) incorporates the contextual information through a novel attention mechanism. It works by by leveraging inter-speaker relation modeling, however it uses distinct memories for each speaker. This is solved with DialougeRNN (Majumder et al., 2019) a method based on RNNs that keeps track of the individual states of speakers throughout conversation. This is then used for emotion classification.

The discovery of Large Language Models (LLMs) have brought in a huge transformation to the field of natural language processing. This is due to to the reasoning and understanding capabilities of these powerful models such as GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023a). Fine tuning of these pre-trained LLMs have showed their versatility and effectiveness across a variety of tasks.

For this task we fine tune 2 models. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), a model to pre-train bidirectional representations by jointly conditioning on both left and right context in all layers. Due to this, the model can be fine tuned with just one layer to achieve state of the art performance. We also use the Llama 2 7 billion parameter model (Touvron et al., 2023b). We choose the Llama 2 model due to it's state of the art performance on various NLP benchmarks. Due to the large size of Llama 2 we fine tune this model using Parameter Efficient Fine Tuning Methods (Mangrulkar et al., 2022). We do this with Low-Rank Adaptation of Large Language Models (LoRA) (Hu et al., 2021) which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture. This reduces the number of trainable parameters.

2.3 Evaluation Metrics

The systems used were evaluated with the weighted F1 score metric.

$$\text{Weighted F1} = \sum_{i=1}^{N} \left(\frac{\text{support}_i}{\text{total support}} \right) \cdot \text{F1}_i \quad (1)$$

$$\text{F1}_i = 2 \cdot \frac{\text{precision}_i \cdot \text{recall}_i}{\text{precision}_i + \text{recall}_i} \quad (2)$$

$$\text{where,} \quad \text{precision}_i = \frac{TP_i}{TP_i + FP_i} \quad (3)$$

$$\text{recall}_i = \frac{TP_i}{TP_i + FN_i} \quad (4)$$

and support_i is the number of true instances of class_i and total support is the total number of instances across all classes

3 System Overview

3.1 BERT

We fine-tune the BERT base model (cased) (Devlin et al., 2019) for the emotion classification task with 8 labels. We load the model and train it using the HuggingFace Transformers library (Wolf et al., 2020). The input text is tokenized with the *bert-based-case* tokenizer.

3.1.1 Model Architecture

The model uses the existing BERT base cased architecture. The final layer of the model (the output layer) is altered to match the 8 classes in the classification task.

3.1.2 Loss Function

For this model we use the **Cross Entropy Loss** between the outputs of the model predictions and the actual labels to optimize the system.

3.2 Llama 2

We fine-tune the Llama 2, 7 billion parameter model (Touvron et al., 2023b) in a similar way in which we fine-tune BERT. We load the model and train it using the HuggingFace Transformers library (Wolf et al., 2020). The input text is tokenized with the *meta-llama/Llama-2-7b-hf* tokenizer.

3.2.1 Model Architecture

Llama 2 model architecture is similar in structure to its predecessor LLaMA (Touvron et al., 2023a) with a context length increase from 2048 to 4096 tokens and usage of Grouped-Query Attention instead of Multi-Query Attention. It is an auto-regressive language model that uses optimized transformer architecture.

3.2.2 Loss Function

We use custom loss function that combines the F1 score and Cross-Entropy Loss to form a single loss value that takes into account both the precision and recall, along with the class imbalances.

4 Experimental Setup

4.1 Dataset Splits

We load the `MaSaC` dataset (Kumar et al., 2023) train, validation and test splits provided to us by the EDiReF shared task organizers using huggingface datasets library (Lhoest et al., 2021). The train set consists of 8506 utterances along with their corresponding label (emotion). The validation set consists of 1354 utterances and the respective label. For final evaluation we are provided with an unlabeled test set of 1580 utterances, to which we must predict the emotion for submission.

4.2 Preprocessing data

Before we pass the inputs to the large language model, we must preprocess the data to an acceptable input format for the large language model, for this we tokenize the datasets.

- **BERT:** For the BERT model we use the pretrained BERT tokenizer *bert-base-cased*. This takes the text of the utterance and generates the *input ids*, *token type ids* and *attention mask*. To make sure all the input sequences have the same length we use maximum length padding. Longer sequences are truncated to the maximum allowable length of the BERT model.

- **Llama 2:** The text for the Llama model is tokenized with the *meta-llama/Llama-2-7b-hf* tokenizer. While tokenizing it is ensured that a space is added before the first token of a given text. The pad token and pad token id are set to the EOS[3] token and EOS token id. While tokenizing, we truncate the longer sequences to the maximum allowable length of the Llama model.

4.3 Training/Fine-tuning

We use the NVIDIA A100 GPUs available on Google Colab for fine-tuning the models.

We load the *bert-base-cased* on HuggingFace for fine-tuning. For the BERT model we use a data loader of batch size 32 while shuffling the data each epoch to not learn any unintended patterns. We use the AdamW optimizer for training (Loshchilov and Hutter, 2019). We set the initial learning rate to be 5×10^{-5} and use a linear learning rate scheduler across the entire duration of training. We then train the model for 4 epochs.

The Llama 2 model is available as *meta-llama/Llama-2-7b-hf* on HuggingFace. We load this model for fine-tuning. Similar to the BERT model, we use a data loader with shuffling for the Llama 2 model, but with a batch size of 16. The AdamW optimzer (Loshchilov and Hutter, 2019) is used while training. Due to the large size of the Llama 2 model, we fine tune the model with PEFT (Mangrulkar et al., 2022) and LoRA (Hu et al., 2021). The LoRA configuration we setup for parameter efficient fine-tuning is as follows. We set the task type as sequence classification, the rank of decompostion matrix (r) is set to 16, the alpha parameter to scale the learned weights (lora alpha) is set to 16 as advised by the LoRA paper. The dropout probability of the LoRA layers is set to 0.05. We do not add any bias term to LoRA layers. We apply LoRA to the projection layers for the query and value components in the attention mechanism of the transformer. We then fine-tune the model for 10 epochs with a learning rate of

[3]End of Speech

1×10^{-4}, warmup ratio of 0.1, maximum gradient norm of 0.3 and a weight decay of 0.001.

5 Results

For evaluation, the organizers rank the system based on weighted F1 score. This is due to the classes being highly imbalanced in the data distribution. The BERT model which was submitted to the leader board achieved a 0.42 weighted F1 score to get 14th place [4]. The performance of all the models can be found in Table 2

	Validation Set	Test Set
BERT	0.43	0.42
Llama 2	0.42	0.41

Table 2: Weighted F1 Scores

Acknowledgements

I would like to thank the organizers of the *EDiReF - SemEval 2024 Task 10* shared task for conducting this competition and organizing this task. I would also like to thank the faculty and research scholars at BITS Pilani for assisting me in my work.

References

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, 14(2):1363–1375.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

[4]https://codalab.lisn.upsaclay.fr/competitions/16769#results

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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, Ramaneswaran S, Md Akhtar, and Tanmoy Chakraborty. 2023. 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.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

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

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik

Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

M. E. Maron. 1961. Automatic indexing: An experimental inquiry. *J. ACM*, 8(3):404–417.

Merriam-Webster. 2024. Emotion. https://www.merriam-webster.com/dictionary/emotion. Accessed: 2024-02-08.

OpenAI. 2023. Gpt-4 technical report.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2023. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

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.