RACAI at SemEval-2024 Task 10: Combining algorithms for code-mixed Emotion Recognition in Conversation

Sara Niță and **Vasile Păiș**

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy

Bucharest, Romania

saramaria.nita9@gmail.com, vasile@racai.ro

Abstract

Code-mixed emotion recognition constitutes a challenge for NLP research due to the text's deviation from the traditional grammatical structure of the original languages. This paper describes the system submitted by the RACAI Team for the SemEval 2024 Task 10 - EDiReF subtasks 1: Emotion Recognition in Conversation (ERC) in Hindi-English code-mixed conversations. We propose a system that combines a transformer-based model with two simple neural networks.

1 Introduction

Emotion recognition in conversation (ERC) (Kumar et al., 2023) is a crucial task in conversational artificial intelligence research that aims to identify the emotion of each utterance in a conversation. ERC proves useful in applications such as opinion mining and empathetic dialog systems. However, many of the existing models and datasets for emotion recognition are single-language. But, proliferating mixed language interactions have boosted interest in code-mixed natural language processing (NLP) tasks.

The present work describes the system that participated in the shared task "Emotion Discovery and Reasoning its Flip in Conversation (EDiReF)", task 10, organized at SemEval 2024 (Kumar et al., 2024). The EDiReF shared task is made up 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. Out of these subtasks, our team participated only in sub-task (i).

Many current approaches for diverse NLP tasks, including ERC, relies on the application of large language models (LLMs) and fine-tuning them on a specific dataset. For this work we were interested in determining how existing language resources, such as emotion lexicons, could be used to complement and improve the predictions of LLM-based approaches. For this reason, our final system, as detailed in Section 4.2, is an ensemble of a BERT-based implementation and traditional feature-based approaches, employing an emotion lexicon. Apart from the emotion lexicon, we did not use any external datasets. Only the dataset provided by the task organisers was used as an emotion annotated dataset. We also took into account the requirement expressed by the task organizers, that no data from task 2 or task 3 can be used to train/evaluate task 1.

This paper is organized as follows: Section 2 provides related work, Section 3 briefly presents the task and describes the dataset, Section 4 gives an overview of the participating system, including pre-processing and architecture, Section 5 presents the results, and Section 6 gives conclusions.

2 Related work

Wang et al. (2020) recognizes the importance of ERC for developing empathetic machines in a variety of areas. The authors model the ERC task as sequence tagging where a Conditional Random Field (CRF) layer is leveraged to learn the emotional consistency in the conversation. Experiments are performed on three datasets: IEMOCAP (Busso et al., 2008), DailyDialogue (Li et al., 2017), and MELD (Poria et al., 2019). The authors ackowledge an imbalanced data distribution in some of the ERC datasets, similar to the distribution provided for the current task (as described in Section 3).

Ghosal et al. (2019) propose Dialogue Graph Convolutional Network (DialogueGCN), a graph neural network based approach to ERC. The authors test the approach on a number of datasets, including IEMOCAP and MELD, showing good results.

Song et al. (2022) employ a Supervised Prototypical Contrastive Learning (SPCL) loss for the ERC task. In this case, the SPCL aims to solve the imbal-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1032–1037
June 20-21, 2024 ©2024 Association for Computational Linguistics

anced classification problem through contrastive learning. Their approach further improve results on the IEMOCAP and MELD datasets, achieving F1 scores of 69.74% and 67.25%, respectively.

De Bruyne et al. (2022) evaluate the language-dependence of an mBERT-based emotion detection model. Experiments included the Hindi and English languages. Their findings suggest that there could be evidence for the language-dependence of emotion detection performance.

Datasets and systems for emotion recognition have been proposed for other languages as well. For example, for the Romanian language, Ciobotaru and Dinu (2021) introduced the RED dataset for emotion detection in Romanian tweets. Colhon et al. (2016) showed that particular Romanian language words, such as negations, intensifiers and diminishers, affect the detected polarity of the sentiments described in natural language texts. Furthermore, Tăiatu et al. (2023) introduced RoBERTweet, a BERT-like LLM for Romanian language. The authors also describe a system using the RoBERTweet model for emotion detection outperforming previous general-domain Romanian and multilingual language models.

Laki and Yang (2023) explore sentiment analysis with neural models for the Hungarian language. The authors try to solve the class imbalance problem either by removing examples from the highly represented class (while keeping the same number of examples as the least represented class) or by duplicating examples from the least represented class. In addition they explore data augmentation by means of machine translation and cross-lingual transfer. Different Hungarian language LLMs, especially BERT-like LLMs, are considered for the experiments. Üveges and Ring (2023) introduce HunEmBERT, a fine-tuned BERT-like model for classifying sentiment and emotion in political communication in the Hungarian language.

Apart from neural network models and datasets, lexicons constitute another type of useful resources for sentiment analysis. This type of resources have been created for different languages. Lupea and Briciu (2019) introduced the Romanian Emotion Lexicon (RoEmoLex v.3). It contains associations between a series of words and eight basic emotions (Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust) and two sentiment orientations (Positivity and Negativity). Initially translated from an English version, it now contains additional tags, including derived emotions, part-of-speech, additional polarity scores and conceptual category information. It was also expanded with synonyms of the original terms and new words and phrases.

Mohammad and Turney (2010, 2013) propose the NRC Word-Emotion Association Lexicon (EmoLex). It contains English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing. The authors assess that despite some cultural differences, the majority of the affective norms are stable across languages. Thus, the lexicon is also provided in over 100 languages by automatic translating the English terms using Google Translate.

Various datasets for sentiment classification, including those mentioned in this section, suffer from a class imbalance problem. Frameworks for data augmentation, such as NL-Augmenter (Dhole et al., 2023), have been proposed, allowing automatic enrichment of less represented classes. Chawla et al. (2002) proposed SMOTE, a synthetic minority over-sampling technique. Their approach combines under-sampling of the majority class with a special form of over-sampling the minority class. The minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement, thus reducing the potential overfitting.

3 Dataset and task

The goal of the emotion recognition task is to classify a given sentence from a dialogue into one of eight emotion states: the seven universal human emotions as described by Dr. Paul Ekman (Ekman, 1992) ("anger", "surprise", "contempt", "disgust", "fear", "joy", "sadness") and "neutral". The dataset files, with splits for training, validation, and testing, were provided in JSON format. The records contain fields for the name of the episode the lines were taken from, a list of speakers, the actual dialogue (list of sentences called "utterances"), and a list with the emotions attributed to each line ("emotions" or "labels"). The utterances included some unrecognized characters that needed to be removed. The training dataset contains 343 entries (8,506 utterances), the validation dataset contains 46 entries (1,354 utterances), while the test dataset contains 57 entries (1,580 utterances).

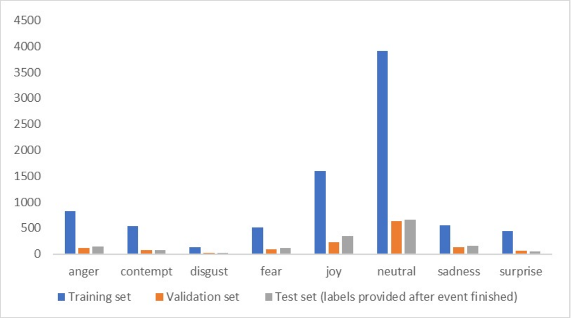

Figure 1: Emotion distribution for train, validation and test sets

The labels distribution for the train, validation and test splits is given in Figure 1. Similar to other emotion recognition datasets, such as IEMOCAP or MELD, as reported in Section 2, there is a class imbalance present in the task dataset as well. Many sentences are marked as being "neutral", while the next class, considering the number of samples, is "joy". The least represented class is "disgust".

4 Methodology

4.1 Pre-Processing

In the pre-processing stage, all blank characters, including new lines, tabs, and other unrecognized UTF-8 characters, were transformed into regular spaces. Dialogues were split into individual sentences and duplicates removed from the training set.

Given the observation of De Bruyne et al. (2022) regarding the possible language-dependence of emotion detection performance, combined with the existence of a large number of emotion lexicons in the English language, individual sentences were completely translated into English, removing any Hindi text (including roman script). For this purpose, we employed the GoogleTranslator from the deep_translator library.

4.2 Overall system architecture

The system is comprised of two parts: one being a multilingual BERT LLM and the other consisting of a Decision Tree and a Random Forest classification algorithms, employing additional features.The final result was obtained by running the three sets of predictions from the models through a voting system. If two or all three models predict the same emotion, then this becomes the final prediction, but if they each give different results, then the BERT prediction is chosen as the final prediction, because when taken separately, BERT has better results then either decision trees or random forest, as shown in in Table 1. A diagram of the entire system is given in Figure 2.

4.3 Decision Tree and Random Forest

To aid in feature construction, the text was lemmatized by employing the WordNet (Fellbaum, 1998) lemmatizer available in the NLTK library[1].

From the translated sentences a set of handcrafted features were produced, some of which were binary features associated with each one of the seven emotions. Through the use of an English lexicon, the NRC-Emotion-Lexicon-Wordlevel[2] (Mohammad and Turney, 2013), the feature was either marked as "1", if the emotion was the most commonly found one among the meanings of the words in a given sentence, or as "0". The lexicon unfortunately did not contain any data about the “contempt” value of words. As a unified resource for both Hindi and English was not successfully found, the translation previously performed was necessary. The other features were: length, the number of sentences in an utterance, punctuation (for full stop, question mark, exclamation mark and ellipsis), ratio of words from the lexicon that were predominantly positive or negative and the confidence of the lexicon. The confidence was computed based on the number of words belonging to different classes which were found in the lexicon for a given sample.

For the decision tree predictions only, new examples were synthesized using a SMOTE pipeline (Chawla et al., 2002) due to the imbalanced nature of the dataset.

4.4 BERT

The LLM used for training the system was bert-base-uncased. This was chosen due to our assumption that a smaller model may benefit more from additional resources, such as an emotion lexicon. The LLM classifier has two additional linear layers, with 2,048 and 1,024 cells respectively, employing ReLU and tanh activation functions respectively. These are followed by a final class prediction head. The model was trained for at least 5 epochs and a maximum of 20 epochs, with early stopping, when there was no improvement for 3 epochs. During the first 3 epochs, the LLM was frozen and only

[1] https://www.nltk.org/

[2] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

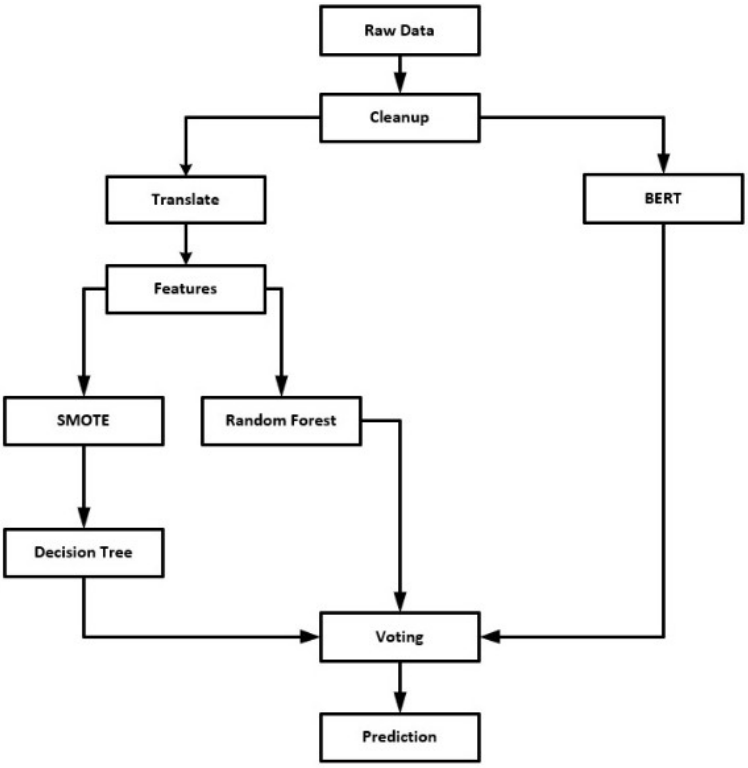

Figure 2: System architecture.

the last linear layers were actually trained. A batch size of 6 was used. The learning rates for the LLM and the other layers were kept separated. The best hyper-parameters were found to be encoder learning rate 1.0e-05, and linear layers learning rate 3.0e-05. The final model training lasted 18 epochs.

5 Results and discussion

Results are given in Table 1 for the test dataset, in terms of weighted precision, recall, and F1 scores. In this case, the weighted recall is equal to the accuracy measure. The baseline is computed on the assumption that all results are neutral. As expected, due to the class imbalance, this provides the best accuracy. Decision tree and random forest classifiers provide results worse than BERT alone and even worse than the baseline approach. This translates into words not being found in the lexicon or words that may mean different things in context, while the lexicon does not take into account the context. Even though the voting mechanism favors the BERT prediction, it seems it actually decreases all the metrics. It is however worth observing the precision score associated with the random forest classifier employing features generated based on the lexicon which is quite high (only 4% under the precision offered by the LLM predictor).

System	P	R	F1
BERT	**36.2**	37.6	**35.0**
DT	7.8	16.7	10.5
RF	0.326	16.7	18.2
Voting	35.2	33.9	30.9
Baseline	17.1	**0.42**	0.24

Table 1: Results on the test dataset.

6 Conclusion and future work

The proposed system tried to combine a lexicon approach with a LLM prediction, considering that a manually created emotion lexicon could complement the LLM predictions. Nevertheless, even though the precision given by the random forest classifier based on features derived using the lexicon is surprisingly good, the recall is significantly lower, thus resulting in an overall lower F1 score, even in the face of a LLM with a reduced number of parameters.

In accordance with open science principles, the code for the described system is made available open source in its own GitHub repository[3].

The class imbalance problem was tackled only with the SMOTE technique. However, as mentioned in Section 2, different frameworks for data augmentation are available. Future work may include experiments with other data augmentation techniques for the minority classes.

As documented in Section 2, different authors have shown improvements using language-specific and domain-specific LLMs. For this work, we focused on a single BERT LLM. Other LLMs, with more specificity or a larger number of parameters, may provide better results. However, the question regarding the possible enhancement of predictions using additional resources, such as emotion lexicons, still remains valid.

Limitations

The current system implementation makes use of English-only emotion lexicons. The system architecture does not take into account long messages that surpass the direct capability of the LLMs used.

Ethics Statement

We do not foresee ethical concerns with the research presented in this paper. However, it is important to acknowledge that unintended bias might be present in the dataset and this could be reflected in the resulting models. Furthermore, since the emotion lexicons have been created by people they capture various human biases which may be reflected in the final system.

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(4):335–359.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Alexandra Ciobotaru and Liviu P. Dinu. 2021. RED: A novel dataset for Romanian emotion detection from tweets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 291–300, Held Online. INCOMA Ltd.

Mihaela Colhon, Mădălina Cerban, Alex Becheru, and Mirela Teodorescu. 2016. Polarity shifting for romanian sentiment classification. In *2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6.

Luna De Bruyne, Pranaydeep Singh, Orphee De Clercq, Els Lefever, and Veronique Hoste. 2022. How language-dependent is emotion detection? evidence from multilingual BERT. In *Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 76–85, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kaustubh Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahadiran, Simon Mille, Ashish Shrivastava, Samson Tan, Tongshang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondřej Dušek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Tanya Goyal, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honoré, Ishan Jindal, Przemysław K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxine Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Meunnighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Păiș, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicholas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, Yiwen Shi, Haoyue Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Zijie J. Wang, Gloria Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyu Wu, Witold Wydmanski, Tianbao Xie, Usama Yaseen, Michael A. Yee, Jing Zhang, and Yue Zhang. 2023. NL-Augmenter: A framework for task-sensitive natural language augmentation. *Northern European Journal of Language Technology*, 9(1):1–41.

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

[3] `https://github.com/SaNita9/ediref2024-subtask-1`

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

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.

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.

László Laki and Zijian Yang. 2023. Sentiment analysis with neural models for hungarian. *Acta Polytechnica Hungarica*, 20:109–128.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Mihaiela Lupea and Anamaria Briciu. 2019. Studying emotions in romanian words using formal concept analysis. *Computer Speech & Language*, 57:128–145.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Iulian-Marius Tăiatu, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Florin Pop. 2023. Robertweet: A bert language model for romanian tweets. In *Natural Language Processing and Information Systems*, pages 577–587, Cham. Springer Nature Switzerland.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195, 1st virtual meeting. Association for Computational Linguistics.

István Üveges and Orsolya Ring. 2023. Hunembert: A fine-tuned bert-model for classifying sentiment and emotion in political communication. *IEEE Access*, 11:60267–60278.