CLTeam1 at SemEval-2024 Task 10: Large Language Model based ensemble for Emotion Detection in Hinglish

Ankit Vaidya* , **Aditya Gokhale*** , **Arnav Desai*** , **Ishaan Shukla*** , **Sheetal Sonawane**
Pune Institute of Computer Technology
{ankitvaidya1905, adityangokhale, arnavdesai235, ishaanshukla10}@gmail.com,
sssonawane@pict.edu

Abstract

This paper outlines our approach for the ERC subtask of the SemEval 2024 EdiREF Shared Task. In this sub-task, an emotion had to be assigned to an utterance which was a part of a code-mixed dialogue. The utterance had to be classified into one of the following classes - disgust, contempt, anger, neutral, joy, sadness, fear, surprise. Our proposed system makes use of an ensemble of language specific RoBERTA and BERT models to tackle the problem. A weighted F1-score of 44% was achieved by our system. We conducted comprehensive ablations and suggested directions of future work. Our codebase is available publicly[1].

1 Introduction

Language has been the primary mode of communication for humans since pre-historic times. In linguistics, code-mixing traditionally refers to the embedding of words or phrases into an utterance of another language (Myers-Scotton, 1993). In many multi-lingual societies we see the development of code-mixed languages. Hinglish is one such language which is a linguistic blend of Hindi and English which is spoken primarily in India. Hinglish generally refers to Hindi that is written in the roman script and is used in combination with some English phrases. The variance in spellings and the multiple interpretations of Hindi words, depending on specific contexts, pose challenges for the analysis of language.

The SemEval workshop (co-located with NAACL 2024) explores and advances the current state of semantic analysis to tackle increasingly complex problems in natural language semanitcs. This paper outlines our approach for the Emotion Recognition in Coversation (ERC) (Kumar et al., 2023) sub-task of the Emotion Discovery and Reasoning its Flip in Conversation (EdiREF) (Kumar et al., 2024) shared task. In this sub-task we had to assign a specific emotion to an utterance which the part of a dialogue. Each episode had multiple speakers speaking in Hinglish. We ranked 11^{th} in this subtask achieving a weighted F1-score of 44%. An end-to-end deep learning pipeline that uses an ensemble of transformer-based Hinglish models was used. We converged on the best models to use in the ensemble by rigorous experimentation using the available models. We also analyse the performance of the classification pipeline and present ablations. We also elaborate on the shortcomings of our systems and some future directions of work.

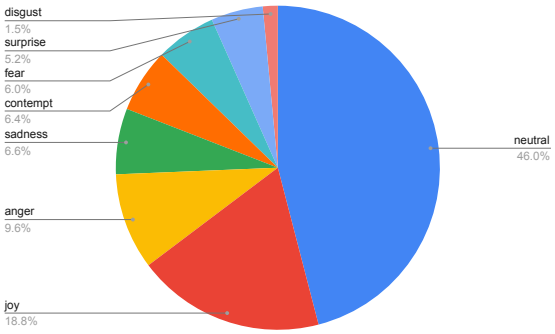

Figure 1: Data Distribution of Training dataset

2 Related Work

Emotion Detection and Sentiment Analysis have been important topics that have been comprehensively studied since the inception of natural language processing. Supervised approaches for Emotion Detection require large datasets which may not be present for low-resource langauges like code-mixed languages.(Orsini, 2015) dates the origin of Hinglish as a language that is widely spoken in India in the post-colonial period. In several works like (Dwivedi and Sukhadeve, 2010), first translation from Hindi-English to English was attempted, however major challenges like non-uniform gram-

* first author, equal contribution
[1] https://github.com/ankit-vaidya19/SemEval24

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 365–369

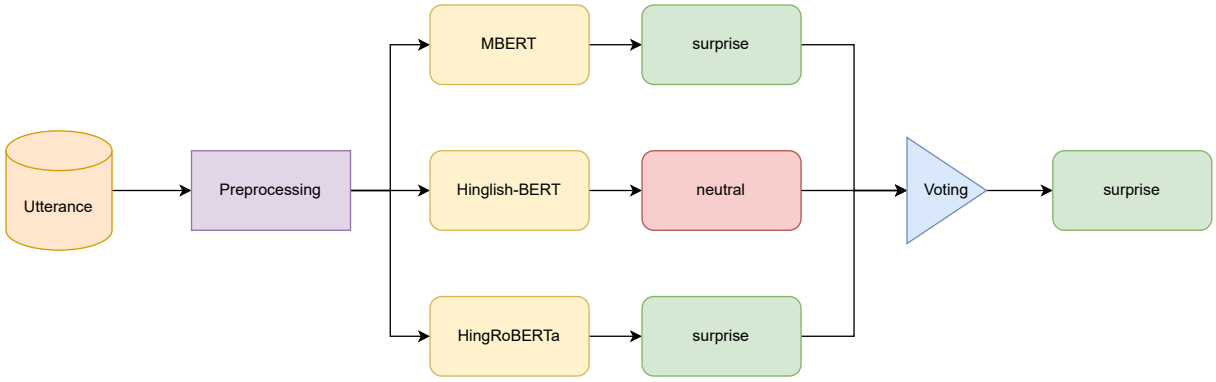

Figure 2: System diagram for Emotion Detection of a sample utterance.

Model	Train F1	Val F1	Test F1
HingBERT	96.72%	45%	43%
HingBERT(LID)	96.67%	44%	42%
HingRoBERTa	96.16%	46%	43%
HingMBERT	96.54%	44%	41%
MBERT	94.76%	41%	40%
Hinglish-BERT	95.76%	42%	41%

Table 1: Comparative results of individual models.

mar and randomised spellings exist could not be overcome.

(Murthy and Kumar, 2021) gives a comprehensive review of modern approaches to detect emotion from text. Extensive work has also been done in the field of Sentiment Analysis of Hinglish text. (Choudhary et al., 2018) made the use of Siamese Networks in order to map the sentences of the code-mixed language and a standard language to a common sentiment space in order to classify the sentences. (Mathur et al., 2018) introduced the Hindi English Offensive Tweets (HEOT) dataset and used a CNN on the embeddings of the data. (Singh and Lefever, 2020) made the use of cross-lingual embeddings obtained using FastText (Bojanowski et al., 2017) and used architectures like CNN, Bi-LSTM and RNN to classify the text. The use of BERT (Devlin et al., 2019) based models was inevitable in this area due to their success in other fields. (Liu et al., 2020) made the use of a pre-trained XLM-RoBERTa (Conneau et al., 2020) and used adversarial examples for the task of sentiment analysis of tweets. However, one thing to note is that most of the prior work has been done on large datasets containing tweets. Due to the large domain shift between analysing tweets and human conversations there was a lack of external training or pre-training data for our task.

For ensemble learning, (Siino et al., 2022) have proposed an ensemble model which generates predictions after the text passes through a vectorisation layer having 2 outputs, one of which is represented as a Bag-of-Words model and provided as input to 3 voters, namely Naive Bayes (NB), Support Vector Machine (SVM) and Decision Tree (DT); and another is a direct input to a CNN. (Kang et al., 2018) proposes a new sentiment analysis method, based on text-based hidden Markov models, that uses word orders without the need of sentiment lexicons. (Miri et al., 2022) proposed use of ensemble feature selection for multi-label text classification which has been used in our approach.

3 Data

The data is in the Hindi-English (Hinglish) code-mixed format which contains words spoken in Hindi but written in the Roman script and English words. The dataset consisted of 343 episodes or dialogues and contained a total of 8,506 utterances which had to classified into eight classes - disgust, contempt, anger, neutral, joy, sadness, fear, surprise. The validation dataset consisted of 46 episodes having 1,354 utterances. The system was then evaluated on a test dataset that contained 57 dialogues consisting of 1,580 utterances. We have illustrated the data distribution in the training dataset in Figure 1. There is acute class imbalance. The class "neutral" contains the most samples (3,123) while the class "disgust" contains the least samples (103). The imbalance ratio was almost 1:30. To mitigate this we tried oversampling to increase size of examples for classes having lower utterances, but they did not improve the performance of the system.

Model 1	Model 2	Model 3	Val F1	Test F1
HingBERT(LID)	HingMBERT	MBERT	45%	42%
HingBERT(LID)	HingBERT	HingMBERT	47%	42%
HingBERT	HingMBERT	MBERT	46%	43%
MBERT	HingMBERT	HingRoBERTa	47%	43%
MBERT	**Hinglish-BERT**	**HingRoBERTa**	**46%**	**44%**
HingMBERT	Hinglish-BERT	HingRoBERTa	46%	44%
HingBERT	MBERT	Hinglish-BERT	44%	44%

Table 2: Results of ensemble pipeline.

4 System Description

The chosen sub-task of emotion detection was a multi-class classification problem which required an utterance to be classified into one of 8 classes. We performed basic pre-processing on the text before passing it to the model. This includes removal of stopwords and punctuation marks from the text, as well as spelling normalisation from the dataset. Due to scarcity of domain specific data related to this task we decided to fine-tune existing transformer-based models to adapt them for our task. Models from (Nayak and Joshi, 2022) like HingBERT, HingRoBERTa, HingMBERT which are based on BERT and RoBERTa (Liu et al., 2019) that were pre-trained on Hinglish data scraped from Twitter were chosen for the task with multilingual models like M-BERT (Devlin et al., 2019). We also chose a variant of BERT (Hinglish-BERT) [2] and a HingBERT variant that was fine-tuned on on the L3Cube-HingLID (Nayak and Joshi, 2022) corpus to include in our system. A linear layer was connected to the pooler output of these models and they were fine-tuned on the dataset. We observed that the performance of the system was enhanced when an ensemble of models was used. We use the method of hard voting to obtain the results from the ensemble. If there was no consensus reached in the ensemble, then the label that the model with the highest F1-score predicted was used as the prediction of the system.

5 Experiments and Results

5.1 Experiments

All the models were used through the HuggingFace (Wolf et al., 2020) library. The data splits that were used during the training and evaluation phase are described in Section 3. The models were fine-tuned for 30 epochs with a learning rate of 1e-5, weight decay of 1e-6 and a batch-size of 32. CrossEntropy loss was used along with the Adam optimizer. We also fixed the seeds to 42. The scoring metric for the task was the weighted F1-score. The scores for the individual models are shown in Table 1. The best performing model checkpoint was chosen according to the epoch-wise validation weighted F1 score. As the individual models had comparable performance on the dataset we decided to create the ensemble by considering all possible combinations of the models. The best performing ensembles and their scores are shown in Table 2.

[2]This model is available here

5.2 Results

The performance of individual models is shown in Table 1 and the performance of the ensemble of models is show in Table 2. The highlighted portion shows our final submission that had a weighted F1-score of 44% consisted of the models MBERT, Hinglish-BERT and HingRoBERTa. We were ranked 11^{th} in the final leaderboard. The difference between our submission and the 5 teams above us was just 1%. We also observed that other combinations also yielded the same result on the dataset as all the models had comparable performance. We also experimented with an ensemble of 5 models (i.e. voting was carried out considering 5 models instead of 3) but the results were similar to our current system and hence, we decided to continue with our current implementation as it is more efficient. The confusion matrix for our submission is illustrated in Figure 3. Note that the confusion matrix has its rows (i.e. true labels axes) normalized according to the number of samples in the class. Here are some observations from our experiments:

1. **The label "anger" has the worst performance:** We observe from Figure 3 that the

label "anger" performs the worst by a significant margin as compared to the rest of the labels despite having relatively more samples compared to some classes. We believe it is due to the fact that the words which characterize anger have a significant overlap with the words that characterize other emotions like "fear" or "contempt".

2. **"joy" vs "surprise" :** We expected the models to confuse these emotions as they are very similar to each other. However, the models rarely confuse these emotions among each other despite the imbalance in the available samples belonging to these two classes. We believe this is due to the fact that these emotions have very distinct appearances in the corpus. We believe that the models captured the subtle difference in the tone that characterize these emotions and thus, could easily differentiate between them.

3. **Failure to capture nuance in negative emotions:** We observe that the overall confusion among negative emotions is higher than the positive emotions. We think that this is due to the fact that many of these emotions have very nuanced differences which the model could not capture due to the scarcity in examples belonging to some of these emotions.

4. **This is a scalable system:** Due to the robust pre-training of the models used, the system could be trained to classify new emotions as well. One could use this system in a continual learning setup in order to increase its capabilities.

6 Conclusion

This paper aims to describe our approach for the ERC sub-task of the 2024 EdiREF Shared Task. We conducted experiments with multiple transformer based models like HingBERT, HingBERT and MBERT. We also show that an ensemble of these models has the best performance on the evaluation dataset with a weighted F1-score of 44%. We foresee several future directions. One direction can be to develop and use more sophisticated methods for ensembling. Another direction is the generation or collection of such data which is more relevant in a real-world scenario in low-resource languages.

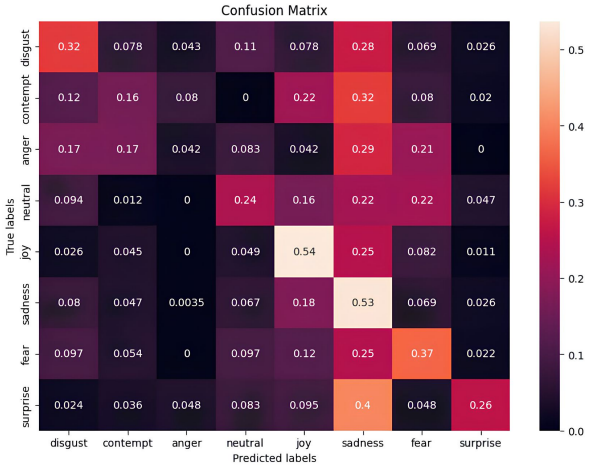

Figure 3: Confusion matrix of system on the Test dataset.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. 2018. Sentiment analysis of code-mixed languages leveraging resource rich languages.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Sanjay Dwivedi and Pramod Sukhadeve. 2010. Machine translation system in indian perspectives. *Journal of Computer Science*, 6.

Mangi Kang, Jaelim Ahn, and Kichun Lee. 2018. Opinion mining using ensemble text hidden markov models for text classification. *Expert Systems with Applications*, 94:218–227.

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.

Jiaxiang Liu, Xuyi Chen, Shikun Feng, Shuohuan Wang, Xuan Ouyang, Yu Sun, Zhengjie Huang, and Weiyue Su. 2020. Kk2018 at SemEval-2020 task 9: Adversarial training for code-mixing sentiment classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 817–823, Barcelona (online). International Committee for Computational Linguistics.

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.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

Mohsen Miri, Mohammad Bagher Dowlatshahi, Amin Hashemi, Marjan Kuchaki Rafsanjani, Brij B Gupta, and Wadee Alhalabi. 2022. Ensemble feature selection for multi-label text classification: An intelligent order statistics approach. *International Journal of Intelligent Systems*, 37(12):11319–11341.

Ashritha R Murthy and K M Anil Kumar. 2021. A review of different approaches for detecting emotion from text. *IOP Conference Series: Materials Science and Engineering*, 1110(1):012009.

Carol Myers-Scotton. 1993. Duelling languages: Grammatical structure in codeswitching.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Francesca Orsini. 2015. Dil maange more: Cultural contexts of hinglish in contemporary india. *African Studies*, 74(2):199–220.

Marco Siino, Ilenia Tinnirello, Marco La Cascia, et al. 2022. T100: A modern classic ensemble to profile irony and stereotype spreaders. In *CLEF (Working Notes)*, pages 2666–2674.

Pranaydeep Singh and Els Lefever. 2020. LT3 at SemEval-2020 task 9: Cross-lingual embeddings for sentiment analysis of Hinglish social media text. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1288–1293, Barcelona (online). International Committee for Computational Linguistics.

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