# Hidetsune at SemEval-2024 Task 10: An English Based Approach to Emotion Recognition in Hindi-English code-mixed Conversations Using Machine Learning and Machine Translation

**Hidetsune Takahashi**
Waseda University
takahashi78h@toki.waseda.jp

## Abstract

In this system paper for SemEval-2024 Task 10 subtask 1 (ERC), I present my approach to recognizing emotions in Hindi-English code-mixed conversations. I train a SpaCy model with English translated data and classify emotions behind Hindi-English code-mixed utterances by using the model and translating them into English. I use machine translation to translate all the data in Hindi-English mixed language into English due to an easy access to existing data for emotion recognition in English. Some additional data in English are used to enhance my model. This English based approach demonstrates a fundamental possibility and potential of simplifying code-mixed language into one major language for emotion recognition.

## 1 Introduction

SemEval 2024 Task 10 (Kumar et al., 2024) calls for assigning emotions to Hindi-English code-mixed conversations (ERC) and reasoning emotion flips (EFR) in Hindi-English code-mixed conversations and in English conversations. I participate in subtask 1 (ERC), for which Hindi-English code-mixed utterances are given in text for participants to recognize emotions behind them. An emotion from disgust, contempt, anger, neutral, joy, sadness, fear and surprise is assigned to each utterance as the correct emotion associated with it.

My methodology has its basis on emotion detection in English. I translate all the development data into English using machine translation (Adep, n.d.), and use the data to train a SpaCy model (Kömeçoğlu, 2023). I also use data by Nikam (Nikam, n.d.) to enhance my model, where utterances in English and their previously assigned emotions are given as a CSV file. He mainly introduces his own model, but I made use of the data only.

I leverage an easy access to emotion-assigned data in English and its linguistic simplicity. Therefore, the point of my approach is to test whether or not machine translation of Hindi-English code-mixed conversations into English can contribute to fundamental accuracy for emotion recognition with lower complexity and an easier access to additional data.

The result shows that combination of machine translation and machine learning works with reasonable accuracy for emotion detection considering its simplicity and numerous data in English that can be implemented in future studies.

My code is available on GitHub [1].

## 2 Background

The subtask I participate in focuses on emotion recognition in Hindi-English code-mixed conversations. Given Hindi-English code-mixed utterances ("kuchh karo sahil please kuchh karo. mera rosesh adopt ho karke chala gaya na to me, i know this sounds horribly melodramatic, monishaish, par me mar jaaungi. i swear mein mar jaaungi" for example) as input, the subtask requires assigning an emotion to each of them as output. Episode name and speakers' names are given as input as well, but they are not used for my solutions.

Before proceeding with this task, Hindi-English code-mixed language needs to be explained in detail. Hindi-English code-mixed language, which is often referred to as *Hinglish*, is a language in which speakers mix Hindi and English in conversations. According to a study by Chand (2016), there are some Hinglish speakers who cannot speak pure Hindi, and even those who speak both Hindi and Hinglish tend to speak Hinglish in conversations with monolingual speakers of Hinglish, causing the number of monolingual speakers of Hinglish to grow as a result. Therefore, the situation in this task is not a small topic but rather an essential one for the future Indian communities.

[1] https://github.com/Hidetsune/SemEval2024_Task10.git



*Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 374–378
June 20-21, 2024 ©2024 Association for Computational Linguistics

Most previous studies aim to recognize emotions in Hinglish by collecting Hinglish sentences and using them directly for their approaches. A study by Vijay et al. (2018) uses n-grams for their solution with Hinglish on Twitter. The work detects emotions with high accuracy, but they state that the accuracy drops by nearly 16% without char n-grams. Another previous work by Sasidhar et al. (2020), which uses the solution by Vijay et al. (2018) as their baseline, collects more data with 12000 Hindi-English code-mixed sentences for training. In addition to these, a study by Wadhawan and Aggarwal (2021) achieves high accuracy with a transformer based BERT model, and another study by Kaur et al. (2019) deals with Hinglish on YouTube comment sections. In all of these works, they process Hindi-English code-mixed sentences without translating them into other languages.

On the other hand, one of the biggest issues with emotion recognition in Hinglish conversations might be shortage of datasets. Although there are many Hinglish sentences expressing emotions on social media platforms, the language they use on them might differ from in real conversations to some extent. Taking the availability of numerous existing datasets in English into account as well, I decided to explore a solution to use English translated data rather than the original Hinglish data. This trial requires machine translation, and this step is combined with classical machine learning method, achieving reasonable accuracy for its simplicity with an additional English dataset concatenated to the translated dataset.

Emotion recognition in Hinglish can be complicated because the language has technically two languages (Hindi+English). However, when simplified into all in English properly, there is much more potential to deal with Hinglish emotion recognition with an easy access to enormous English datasets and established NLP methods that have been mainly used for English. This paper aims to guide a direction to future application of this approach, establishing the basis with easily used datasets and model.

## 3 System overview

The main strategy of my system is a combination of classical machine learning method with simplification of Hindi-English code-mixed sentences into English sentences. Therefore, my methodology is composed of data preparation using machine translation and the classic machine learning process with training. A quick overview of my algorithm is as follows.

1. **Official training data and development data translation**: The official data in json file format are imported and converted into pandas dataframes. The dataframes have episode names, utterances, speaker names and emotions as columns. The utterances are translated into English and saved in a new column.

2. **Additional data concatenation**: Concatenate additional data, which have utterances in English and the emotions, with the translated official data.

3. **Addressing data imbalance**: Separate the concatenated data into each emotion and set a limitation of 3001 utterances (including 3000th counting from 0) per one emotion type.

4. **Model training and prediction**: Train a model with re-concatenated data to predict emotions for unlabeled evaluation data.

For data preparation, I use the official training and development dataset (Kumar et al., 2023), and additional data by Nikam (n.d.) to enhance my model and mitigate data imbalance. Firstly, I import the official datasets in json file format (Kumar et al., 2023), and convert them into pandas dataframes which have episode names, utterances in Hindi-English mixed language (pronunciation forms of Hindi + English), speaker names and emotions as columns. Then, I use Google Translate (Adep, n.d.) to translate all the utterances in Hindi-English mixed language into English, and I add the translated sentences to the dataframe as a new column named "utterances_English". Since episode name, utterances in mixed language and speaker names are of no use at this time, they are dropped from the dataframe to have only "emotion" and "utterances_English" as columns on the dataframes. After that, they are concatenated into one dataframe. There is a huge data imbalance and shortage of training data at this point, so I use additional data by Nikam (n.d.) to mitigate these problems.

In the next step, I train the SpaCy-v3 model (Kömeçoğlu, 2023) with the prepared data and use it to assign emotions to unlabeled evaluation data. The evaluation data are composed of episode name, utterances in Hindi-English mixed language, speaker names as columns. "utterances_English" column, in which machine translated sentence is assigned to each given utterance, is added to have the model predict emotions behind the utterances.

Participating in this emotion recognition task using the combined strategy enables me to explore the potential of implementing machine translation into a specific situation as this task. The application of machine translation to machine learning makes it possible for a multi-label text classifier to predict emotions behind Hinglish sentences with reasonable accuracy for the entry of this trial. Considering the saved complicated steps needed to handle the data as text in two separate languages, you can see the potential of simplification that my methodology aims at by combing machine translation with machine learning. Future applications for higher accuracy might include more training data in Hindi-English mixed sentences and enhancement of the model with the implementation of machine translation into Hindi written sentences. Participating in this task reveals both strengths and weaknesses of my strategy, guiding directions of future studies to apply machine translation and classic NLP methods to emotion recognition in Hinglish conversations.

## 4 Experimental setup

Before moving on to the actual training of the model, some setups were required to prepare training data. As stated before, the utterances in the provided development data (Kumar et al., 2023) are all in Hindi-English mixed language. Since my participation in this task intends to combine machine translation with machine learning for emotion recognition, I decided to simplify the data by translating them into one language rather than taking the complexity of separating them into the two languages. This simplification by machine translation might have changed the original meanings of the utterances, but the difference might not be significant for recognizing the emotions only. Looking through the Hindi-English code-mixed sentences, I noticed that fairly large parts of the utterances are in pronunciation forms in Hindi and that English is used only partly. For instance, some short words or phrases like "goodbye!" are utterances where English appears only, but sentences like "lekin what about my ghadi? 17000 ki ghadi hai..." are mostly composed of pronunciation forms of Hindi (except "what about my" in this case). Pronunciation forms of Hindi are much more prevalent in many other utterances.

There could be two possible choices in my approach; translate all the mixed language sentences (Hindi+English) into Hindi or into English. It might be a good idea as well to choose the former considering the high prevalence of Hindi, but I chose the latter due to the higher availability of reliable additional data on emotion recognition in English.

As for machine translation, Google Translate (googletrans 3.1.0a-0) was used. The reference (Adep, n.d.) uses it to translate sentences in all actual Hindi characters (no English and in written form of Hindi; not in pronunciation form). On the other had, I undertook an experiment to try it for the official development dataset (Hindi-English code-mixed and pronunciation form for Hindi; not in written form) and founded that it works. Therefore, all the given utterances are translated into English and added to the dataset as a new column.

At the next step, additional data are collected and processed to be concatenated with the translated version of official data. Since there is a huge data imbalance and lack of training data as stated in the previous section, it is obvious that additional data is needed where most of the 8 emotions used in this task (disgust, contempt, anger, neutral, joy, sadness, fear and surprise) are labeled for utterances or sentences. For consistency of additional data, one source was used rather than concatenating multiple sources with different categories of emotions.

Data by Nikam (Nikam, n.d.) were chosen for additional data. The original concatenated data (the official training and validation data) are composed of text in English and 8 emotions (disgust, anger, neutral, joy, sadness, fear, contempt and surprise) for each utterance. The only two differences in categories of emotion are that the additional data does not have "contempt" while the official one have, and that the additional data have "shame" while the official one does not. Details are as shown on Table1. The concatenated version of official data and the additional data are concatenated separately for each emotion, and the number of utterances is adjusted after that. Conducting some experimental trial on development phase, I decided to limit the data so that it has 3001 utterances (3000th rows counting from 0) at maximum for each emotion type, and all the utterances that exceed the number (from 3002th rows) are dropped from the training dataset.

## 5 Results

In the evaluation phase, my trained model worked with accuracy of 0.39 (weighted F1 score calculated by the organizers' system ), and the ranking was 17th out of 39 participants. The result is not as good as to use for a practical use as of now, but it definitely shows a basic ability of my approach to apply machine translation to Hindi-English code-mixed language.

There are certain weaknesses in my algorithm as the result shows. One of the biggest weaknesses is the lack of dataset due to inevitable data imbalance. Concatenating the official data (Kumar et al., 2023) with additional data (Nikam, n.d.), there are only 1004 utterances associated with disgust for example while there are many more utterances than 3001 for anger. Limiting the number of utterances up to 1001 for each emotion yielded low accuracy of 0.27 in the development phase, so I chose to accept data imbalance to some extent with the limitation of 3001 utterances for each emotion so that the model is trained better. Larger data imbalance was inevitable to make use of more data, yielding lower scores in development phase, so I had no choice but to decide on around that limitation. However, it cannot be said that less than around 3000 training utterances per emotion are enough for accurate emotion recognition, which is one of the main weaknesses of my solution in this task.

In addition to that, the process of machine translation might have changed the original meaning of the utterances, which might have lowered the quality of training. I cannot look deeper into this possible issue because I am not a Hindi speaker, but the accuracy might go up with a better machine translator.

Another weakness of my solution is that the additional data I used (Nikam, n.d.) are provably not from actual conversations. There are many "@" followed by what I suppose are usernames, so the entire additional data (Nikam, n.d.) are probably from social media platforms. Since this task deals with text version of conversations, situations of the datasets are probably unmatched with each other.

On the other hand, the strength of my approach is the linguistic simplicity that enables the model to have its potential to utilize established NLP techniques and numerous data that have been developed for English. Since English is used widely around the world, there are tons of other data sources that can be implemented into my model. It goes without saying that there are already many existing data in which emotions are labeled with sentences, I can also write down daily conversations in English into text and label each utterance with an emotion for higher accuracy since this task deals with data from conversations.

## 6 Conclusions

To summarize, I firstly simplified utterances in Hindi-English code-mixed language by translating them into English. Machine translation (Adep, n.d.) is used in this step, and additional data (Nikam, n.d.) are concatenated to mitigate data imbalance and enhance the model. Utterances of evaluation data (Kumar et al., 2023) are also translated into English by machine translation in the same way, and the trained model predicted an emotion for each utterance.

Although my solution has a room for improvement, the result shows a basic ability of the simplification by machine translation to recognize emotions behind Hindi-English code-mixed utterances. Given the abundance of data and established NLP techniques in English, my approach, combining machine translation with classical NLP methods, might open the door for addressing the challenges in emotion recognition in Hinglish caused by its linguistic complexity.

## References

Venugopa Adep. n.d. Hindi to English translation using Python. *kaggle*.

Vineeta Chand. 2016. The rise and rise of Hinglish in India.

Gagandeep Kaur, Abhishek Kaushik, and Shubham Sharma. 2019. Cooking is creating emotion: A study on hinglish sentiments of youtube cookery channels using semi-supervised approach. *Big Data and Cognitive Computing*, 3(3).

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

| Dataset | Anger | Contempt | Disgust | Fear | Joy | Neutral | Sadness | Surprise | Shame |
|---|---|---|---|---|---|---|---|---|---|
| Concatenated official data | 937 | 616 | 148 | 602 | 1824 | 4542 | 684 | 507 | 0 |
| Additional data | 4286 | 0 | 856 | 5409 | 11037 | 1811 | 6719 | 4062 | 146 |

Table 1: Datasets and emotion categories

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.

Başak Kömeçoğlu, Buluz. 2023. Emotion classification with SpaCy v3 and comet.

Sanket Nikam. n.d. Emotion detection in text using natural language processing.

T Tulasi Sasidhar, Premjith B, and Soman K P. 2020. Emotion detection in Hinglish(Hindi+English) code-mixed social media text. *Procedia Computer Science*, 171:1346–1352. Third International Conference on Computing and Network Communications (CoCoNet'19).

Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Corpus creation and emotion prediction for Hindi-English code-mixed social media text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 128–135, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Anshul Wadhawan and Akshita Aggarwal. 2021. Towards emotion recognition in hindi-english code-mixed data: A transformer based approach. *arXiv preprint arXiv:2102.09943*.