Sociolinguistically Informed Interpretability: A Case Study on Hinglish Emotion Classification

Kushal Tatariya[1] Heather Lent[2] Johannes Bjerva[2] Miryam de Lhoneux[1]

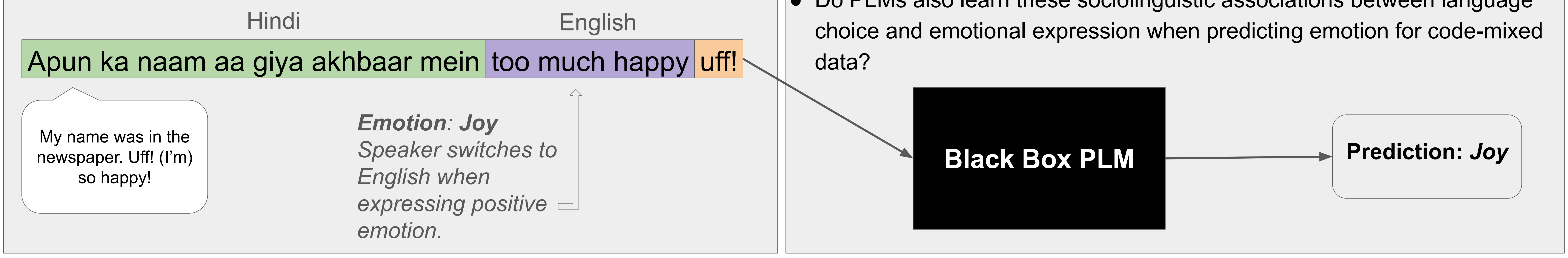
[1] Department of Computer Science, KU Leuven, Belgium
[2] Department of Computer Science, Aalborg University, Denmark

kushaljayesh.tatariya@kuleuven.be

***Disclaimer:** This poster contains some examples of language use that readers may find offensive.*

Code-Mixing and Emotional Expression

- The mixing of two or more languages in the same conversation is called 'code-mixing'.
- Hinglish = mixing of Hindi and English by bilingual speakers
- Studies in sociolinguistics have shown that Hinglish speakers prefer to use English to express positive emotions, and hindi to express negative emotions and swear.


Hindi
English
Apun ka naam aa giya akhbaar mein
too much happy
uff!
My name was in the newspaper. Uff! (I'm) so happy!
Black Box PLM
Prediction: Joy

***Emotion**: Joy*
Speaker switches to English when expressing positive emotion.

Language Models and Interpretability

- Adoption of pre-trained language models (PLMs) has improved performance across the board for emotion classification and code-mixed NLP.
- But, what do these models learn when predicting emotion? They still remain black boxes. This interpretability problem could be approached through the lens of sociolinguistics.
- Do PLMs also learn these sociolinguistic associations between language choice and emotional expression when predicting emotion for code-mixed data?

Methodology

1. Fine-tuning

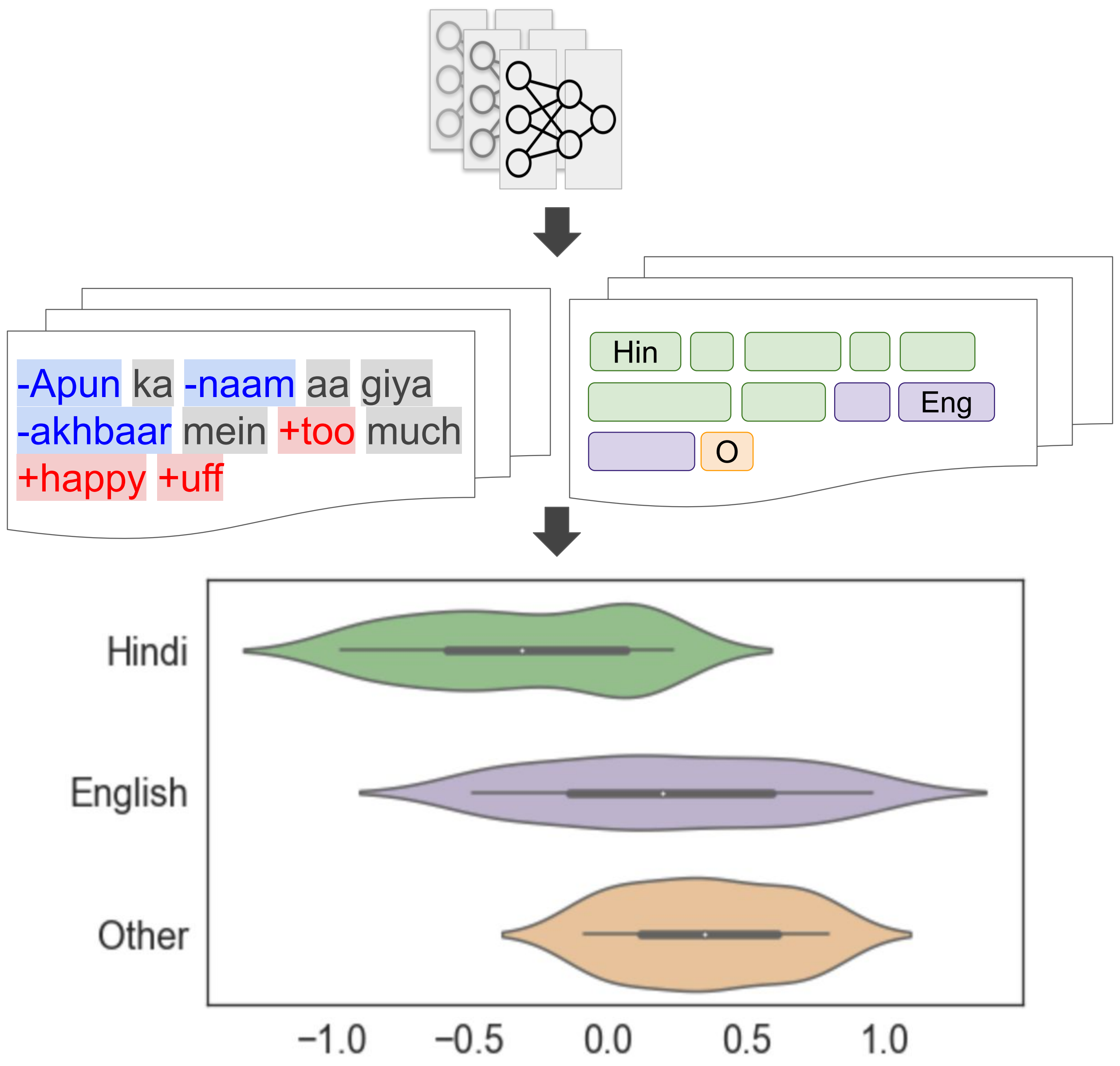
We fine-tuned 3 PLMs (XLM-R, IndicBERT and HingRoBERTa) on a Hinglish emotion classification dataset. The task is to label each example as one of 7 emotion labels: *Joy, Sadness, Fear, Surprise, Disgust, Anger, Others*

2. Token Tagging

We picked 1000 samples from the validation set, maintaining the default distribution across labels. For each sample, we added 2 tags:

- ◆ Token level language identification (English, Hindi, Other)
- ◆ LIME score: LIME is an interpretability metric that assigns a score between -1 and 1 to each token in the sentence. A positive score indicates that the token influenced the model towards the predicted label, and a negative score indicates that the token influenced the model to *not* predict that label.

We examined the distribution of LIME scores across each language ID tag - looking at the frequency with which a token received a positive or a negative LIME score per language, for each model. We validated the significance of our findings with chi-square and 1-way ANOVA.



Results

We observed a dependency between language ID and LIME scores for examples that the models predicted as *joy* (positive emotion) and *anger* (negative emotion).

RQ1: Do English tokens influence models to predict positive emotions?

Yes! English tokens have a significantly higher frequency of being assigned a positive LIME score for *joy,* i.e. influencing the model to predict a positive emotion.

Tweet: @handle Wow dear I am proud of you kiya gali de ho aapne
Lang_ID: other eng eng eng eng eng eng eng hin hin hin hin hin
Translation: Wow, dear, I am proud of you. You have cursed so

HingRoBERTa: @handle Wow dear I am proud of you kiya gali de ho aapne
XLM-R: @handle Wow dear I am proud of you kiya gali de ho aapne
IndicBERT: @handle Wow dear I am proud of you kiya gali de ho aapne

Figure 2: An example from the dataset labelled as *joy*, with the translation and language ID tags. The 3 tokens with the highest LIME scores are marked in blue, and the the 3 tokens with the lowest LIME scores are marked in red.

RQ2: Do Hindi tokens influence models to predict negative emotions?

Yes! Hindi tokens have a significantly higher frequency of being assigned a positive LIME score for *anger,* i.e. influencing the model to predict a negative emotion.

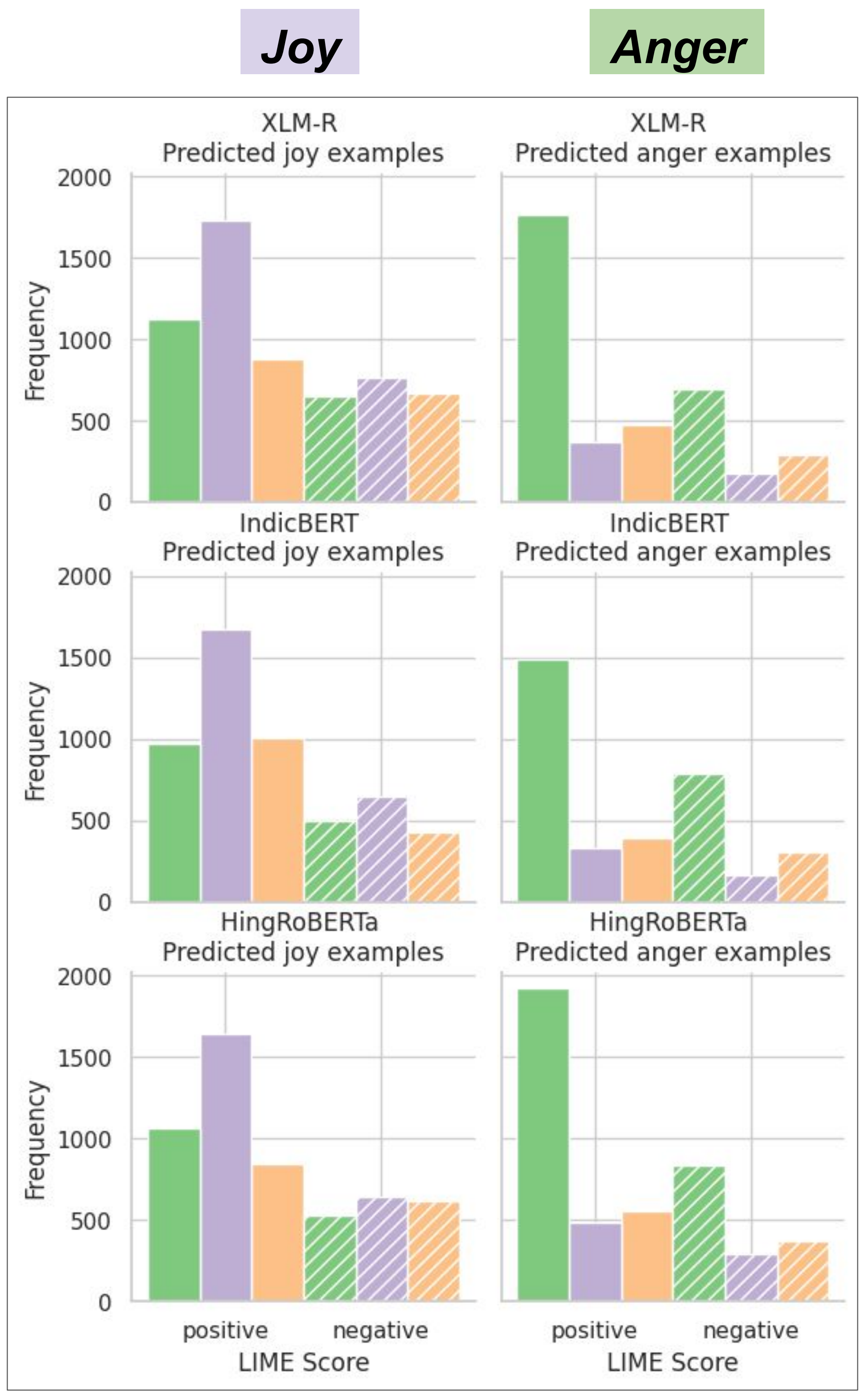


Figure 1: The figure shows the frequency of Hindi (green), English (purple) and Other (orange) tokens to be assigned a negative or positive LIME score for examples predicted *joy* and *anger*.

What role do Hindi swear words play in predicting negative emotions?

Token	Lang_ID	Swear Word?
Fuck	eng	Yes
Chutiye	hin	Yes
Fakeionist	eng	No
Bsdk	hin	Yes
Sadly	eng	No
Bakwas	hin	No
Kutta	hin	Yes
Gaddar	hin	No
Shame	eng	No
Sala	hin	Yes

Table 1: Top 10 tokens with highest LIME scores when predicting negative emotions (*anger, sadness, disgust* and *fear*) for all models. They have been mapped to a canonical form and are in descending order of LIME score.

Four of the words are Hindi swear words. As such, we can see that the models not only learn the negative connotation of the Hindi swear words, but also that these Hindi swear words are the *most* negative of all other tokens, regardless of language.

TL;DR

Sociolinguistic studies suggest Hinglish speakers switch to Hindi to express negative emotions and English for positive ones. Our interpretability analysis using LIME reveals pre-trained language models mimic these patterns when predicting Hinglish emotions—Hindi tokens influence the model to predict *anger*, while English tokens influence prediction of *joy*.