

Transfer Learning for Code-Mixed Data: Do Pretraining Languages Matter?

Kushal Tatariya¹ Heather Lent² Miryam de Lhoneux¹

¹ Department of Computer Science, KU Leuven, Belgium

² Department of Computer Science, Aalborg University, Denmark

{kushaljayesh.tatariya, miryam.delhoneux}@kuleuven.be; hcle@cs.aau.dk

Code-Mixing and Sentiment Analysis

- Mixing of 2(+) languages in the same conversation
- Present in informal domains
- Strong relationship between language choice and sentiment

Jaana he padega

but then

ghar jaake

I'll have to cook




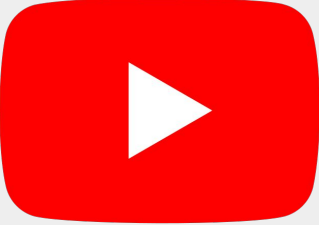
😞😞

Hindi

English

Hindi

English



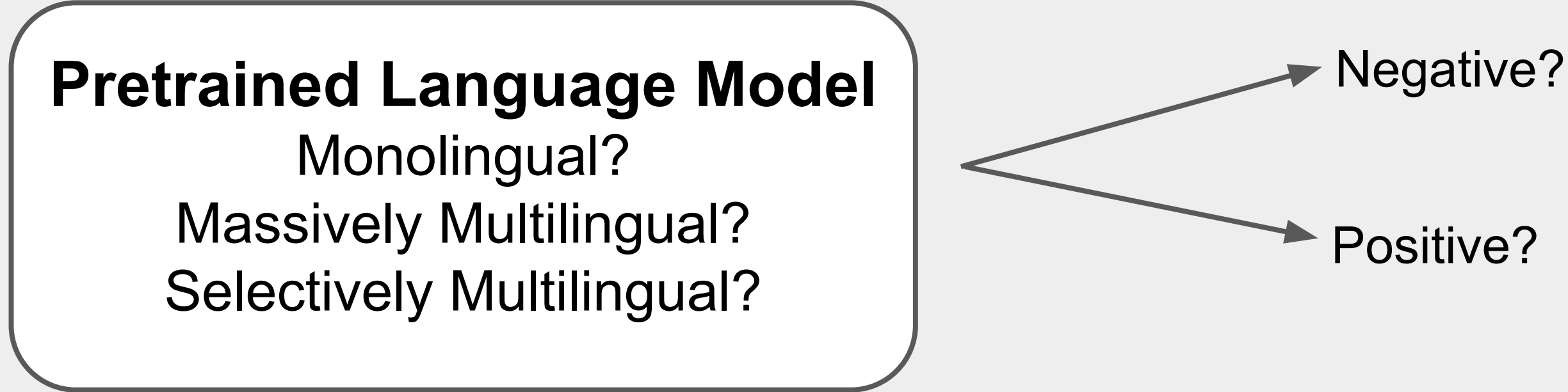
Social media:

- Very informal!
- Lots of code-mixing!
- Lots of sentiment!
- Great source of data!

I have to go. But then I'll have to cook when I get home.

Transfer Learning

There are no PLMs exclusively for code-mixed data.
How far can transfer learning get us?



Caveat:
The curse of multilinguality

Question:
How much do the languages used in the pretraining of PLMs interact with each other to impact their performance on code-mixed data?

Hypothesis

PLMs trained exclusively on data from relevant languages would demonstrate better performance than those that contain other extraneous languages and/or are only trained on one language.

Models

Monolingual	BERT, RoBERTa
Multilingual	mBERT, XLM-R
Indic	IndicBERT, MuRIL
African	AfriBERTa, AfroXLMR
Codemix	HingMBERT

Datasets

Naija	NaijaVader, AfriSenti
Hindi-English	IIITH-CodeMix, SAIL
Tamil-English	DravidianCodeMix (Tamil)
Malayalam-English	MalayalamCodeMix
Kannada-English	DravidianCodeMix (Kannada)

Scenario 1 - In-Language Finetuning

Dataset	Indic-BERT	MuRIL	AfriBERTa	AfroXLMR	mBERT	XLM-R	BERT	RoBERTa	Hing-MBERT	Standard Deviation
AfriSenti	-	-	0.75	0.78	0.76	0.77	0.77	0.76	0.77	0.009
NaijaVader	-	-	0.72	0.74	0.73	0.74	0.74	0.73	0.73	0.007
SAIL	0.62	0.62	-	-	0.60	0.64	0.60	0.61	0.66	0.02
IIITH-CodeMix	0.69	0.70	-	-	0.69	0.71	0.70	0.70	0.74	0.015
Malayalam-English	0.76	0.77	-	-	0.75	0.76	0.76	0.75	0.77	0.007
Tamil-English	0.71	0.70	-	-	0.70	0.71	0.70	0.71	0.71	0.004
Kannada-English	0.71	0.70	-	-	0.70	0.67	0.66	0.70	0.70	0.017

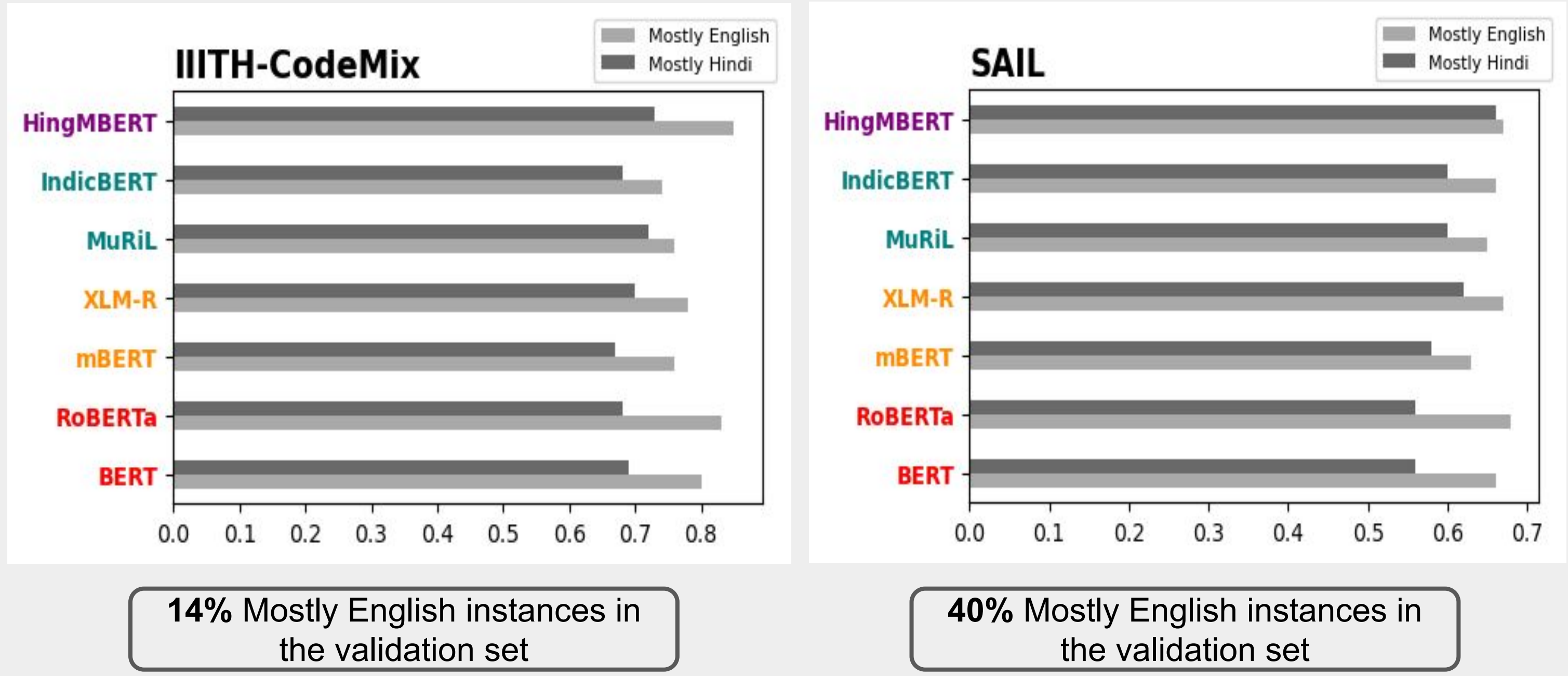
Very little difference seen in performance!

We found the same trend for other tasks: NER, Sarcasm detection, & UDPoS

Analysis

Language Identification and Composition

Each instance in the Hinglish validation sets was labelled 'mostly-English' or 'mostly-Hindi' based on the number of English and Hindi terms in the instance. We looked at whether there were differences in model performances on both types of instances.



All models perform better on 'Mostly-English' than 'Mostly-Hindi' instances, with the pretraining language of the PLM and the language composition of the dataset potentially accounting for how big that difference is. The monolingual models have the biggest difference in both datasets. Models with the least difference are the Indic ones for IIITH-CodeMix and the code-mixed model for SAIL.

Scenario 2: Zero-Shot

We finetuned the monolingual, multilingual, codemix and Indic PLMs on monolingual English and Hindi sentiment analysis datasets and tested the models in a zero-shot setting on the Hinglish datasets.

	SAIL		IIITH-CodeMix	
	Hindi	English	Hindi	English
IndicBERT	0.62	0.61	0.60	0.56
MuRIL	0.64	0.57	0.74	0.43
mBERT	0.57	0.56	0.64	0.47
XLM-R	0.63	0.62	0.70	0.46
BERT	0.61	0.62	0.63	0.57
RoBERTa	0.61	0.66	0.55	0.73
HingMBERT	0.72	0.69	0.78	0.77
Standard Deviation	0.04	0.04	0.07	0.12

Both the pretraining language and the language used in downstream finetuning affect performance.

- RoBERTa suffers drastically from Hindi finetuning, and is the best model with English finetuning (after HingMBERT).
- MuRIL suffers drastically from English finetuning, and is the best model with Hindi finetuning (after HingMBERT).
- The language composition of the dataset also potentially affects how much the score of the best performing model differs from the least performing model.

Conclusion

Do Pretraining Languages Matter?

When finetuning a PLM on code-mixed data: **Not Really!**
The process of finetuning negates the effects of the pretraining languages in the PLMs and generates even performance across the board.

When testing a PLM on code-mixed data zero-shot: **Yes!**
The impact of the pretraining languages is further magnified by the language used during downstream finetuning.