# Code-Mixed Probes Show How Pre-Trained Models Generalise On Code-Switched Text

Frances A. Laureano De Leon
Dr. Harish Tayyar Madabushi
Prof. Mark Lee

UNIVERSITY OF BIRMINGHAM

LREC-COLING 2024

UNIVERSITY OF BATH

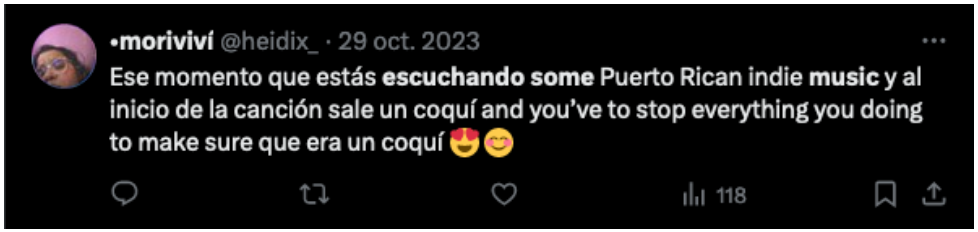
May 1, 2024

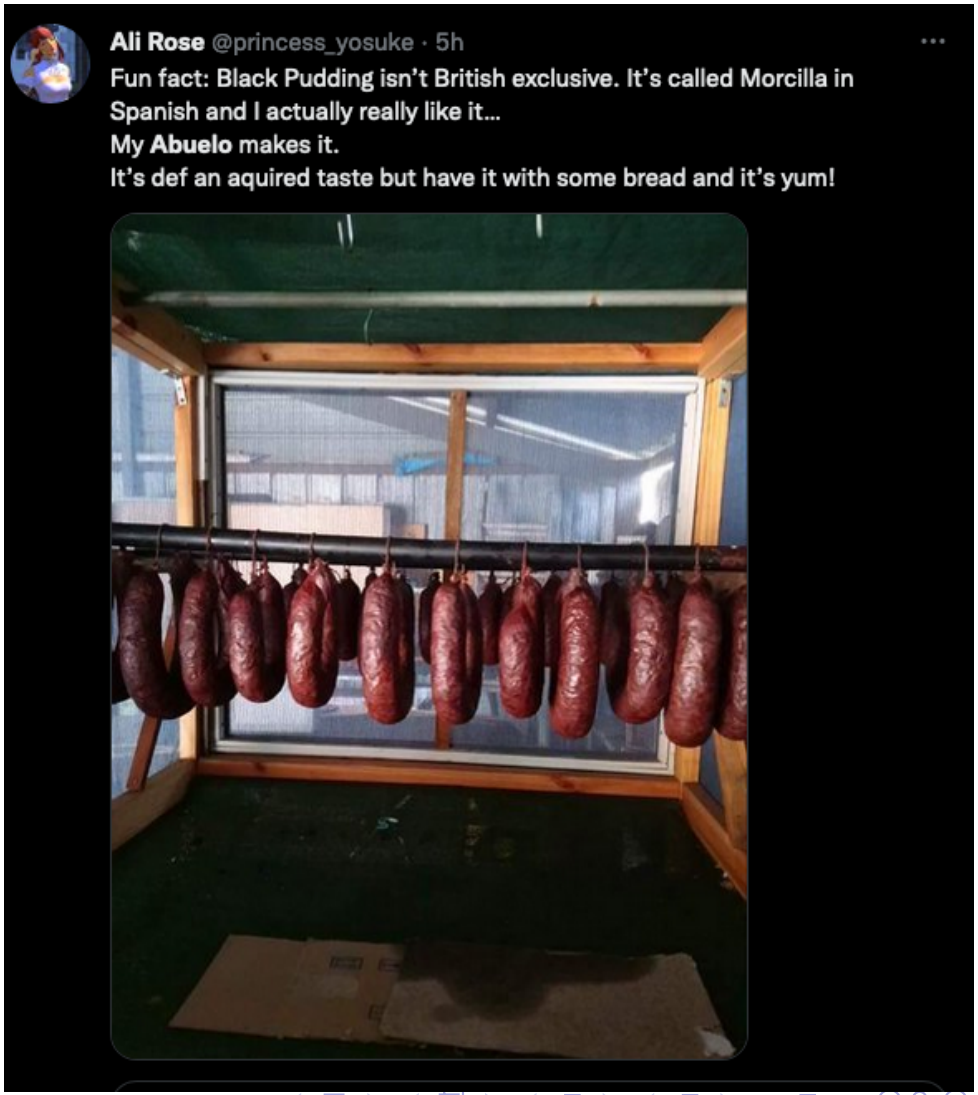
# Overview

# Code-switching

linguistic phenomenon in which multilingual individuals seamlessly alternate between languages.

# Code-switching

# What are we trying to find?

- **Detection:** can models detect code-switched text?
- **Syntax:** structure of code-switched text closer to one source language when compared to another?
- **Semantics:** meaning reps of code-switched text consistent to reps of translation in source languages?

# What tools will we need?

## Probes

- Auxiliary classifier - linear probe trained on detection task
- Structural probe - extract dependency parse
- Semantic probe - train models on STS task and evaluate

## Datasets

- Created a small curated dataset to compare apples to apples using techniques in De Leon, Guéniat, and Madabushi 2020
- SemEval 2020 Task 9 SentiMix, Spanglish dataset
- CALCS 2021 Shared Task: Machine Translation for Code-Switched Data
- Universal Dependencies Ancora and EWT datasets

# Detection

- layer-wise exploration.
- sentence-classification: monolingual vs. CS text
- sentence classification: [CLS] token vs mean pooling
- token classification: LID task

# Example of linear probe

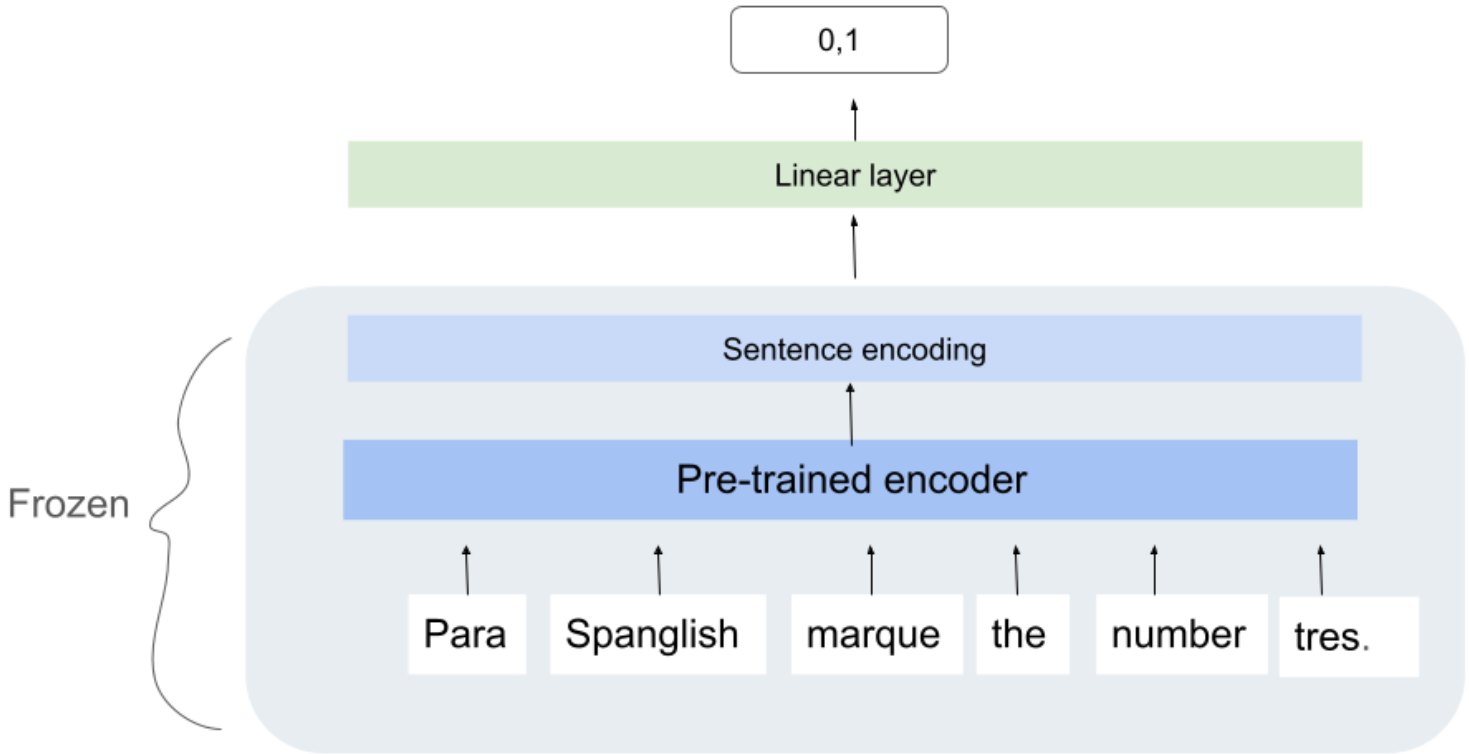



Figure: Linear probe for sentence classification

# Detection

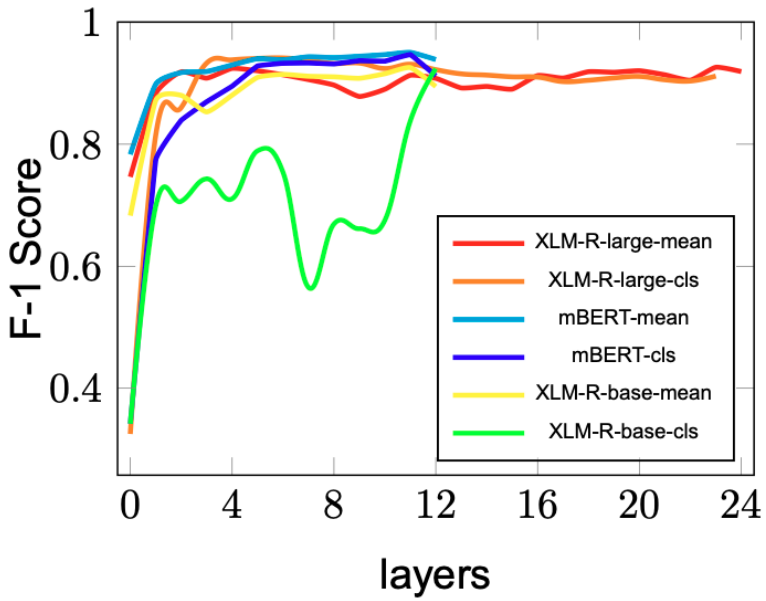



Figure: Mean F-1 Scores across layers for the sentence classification task for each of the PLMs studied.

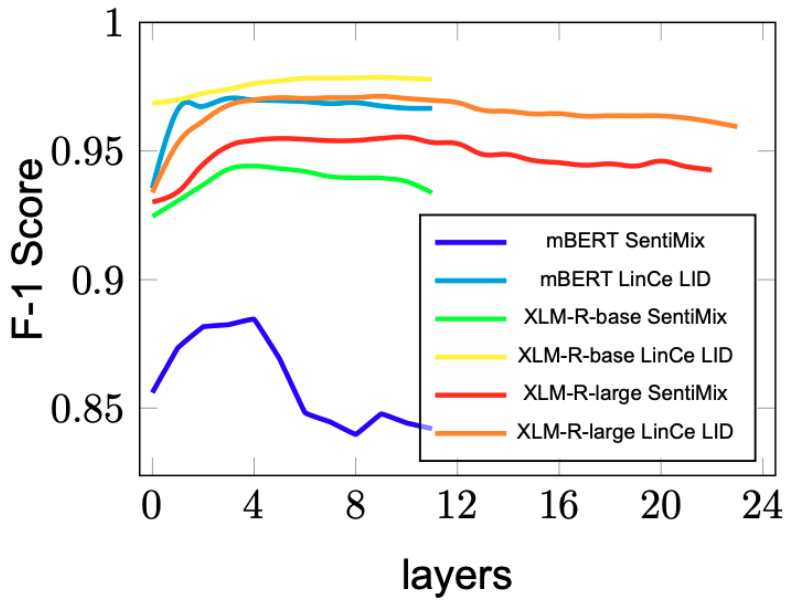



Figure: LID model mean F-1 Scores across layers for the probe classifiers.

# Syntax

- structural probes trained in monolingual es and en
- graph-edit distance (cs vs en), (cs vs es) of dependency parses - no gold labels
- we need parallel corpus, we use our created dataset
- ablation studies using synthetic data derived from collected examples

# Example of syntax probe

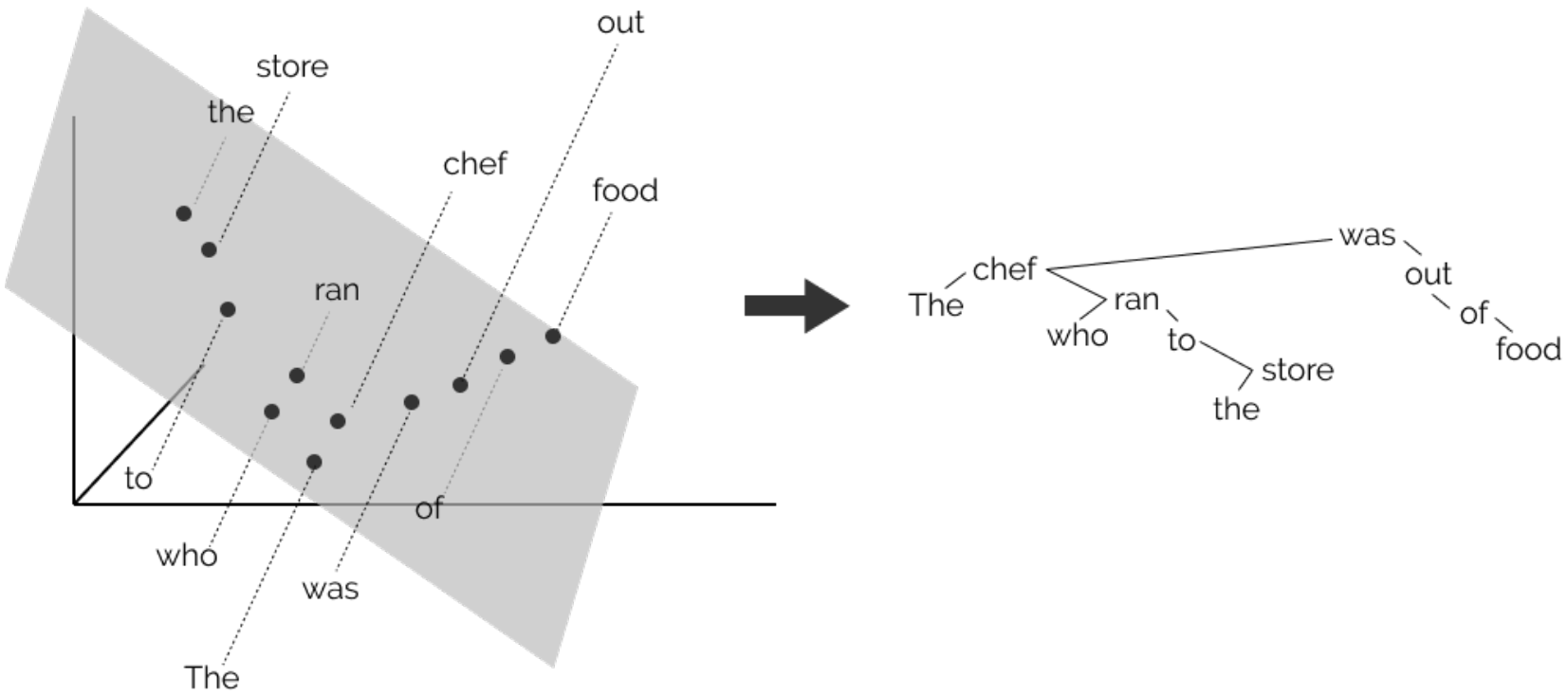
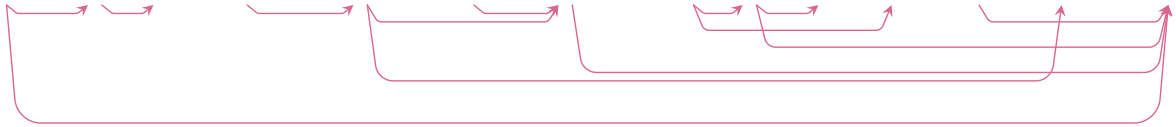



Figure: Structural probe for dependency parses of a sentence. Hewitt and Manning 2019

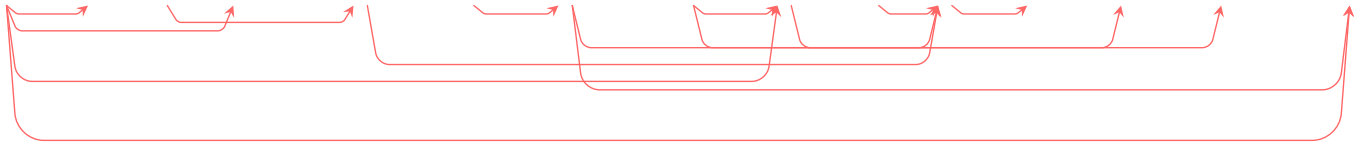
# Example of dependency parse for each language

Pensé que se había muerto bad bunny , do n't ever do that again tuiter .



I thought bad bunny was dead , do n't ever do that again tweeter .

Pensé que se había muerto bad bunny , no vuelvas a hacer eso nunca más , tuiter .



# Syntax results

| lang-pair 1 | lang-pair 2 | Spearman statistic |
|---|---|---|
| cs vs. en | cs vs. es | 0.8308 |
| NP-CS-en vs. en | NP-CS-en vs. es | 0.6876 |
| NP-CS-es vs. en | NP-CS-es vs. es | 0.7564 |
| randCS vs. en | randCS vs. es | 0.6983 |

Table: Spearman rank for correlation between distances of code-mix and monolingual text. Results on real CS data is highlighted.

# Semantics

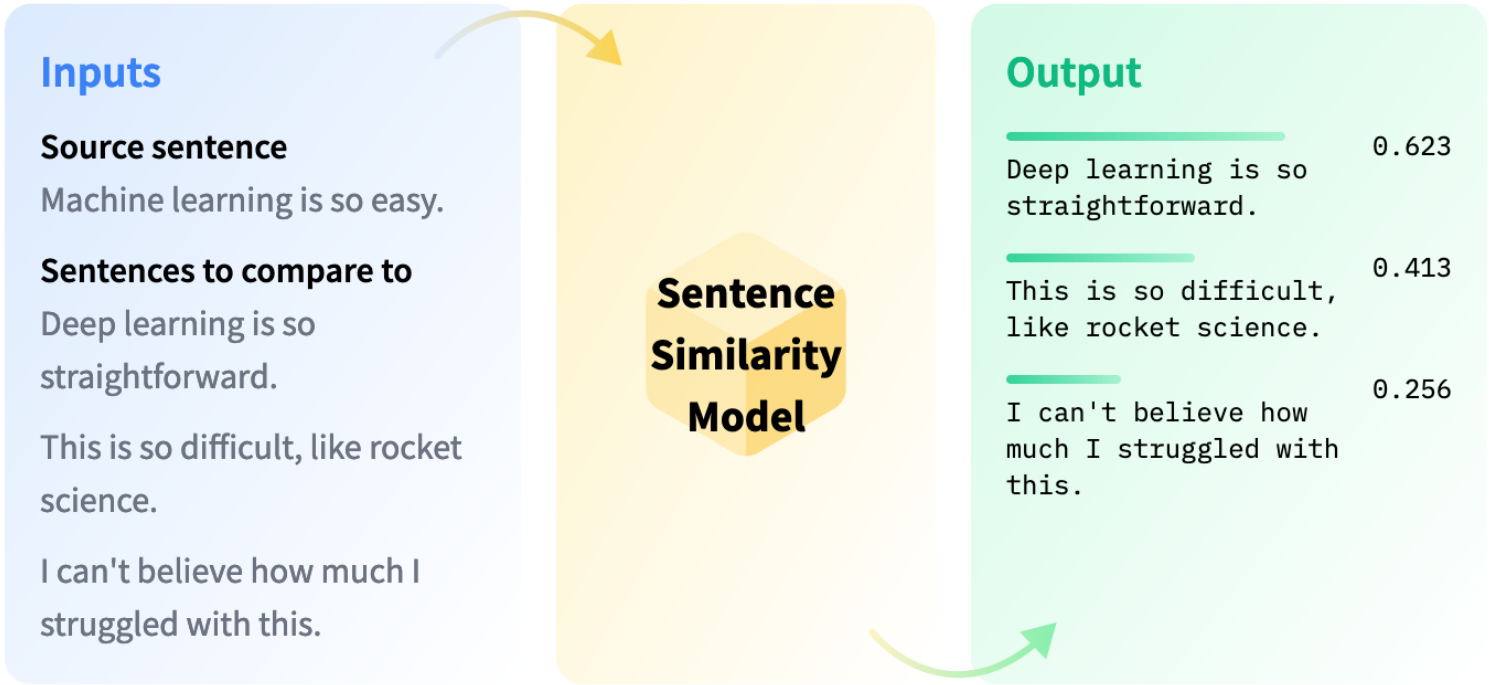



Figure: Example of STS task from `https://huggingface.co/tasks/sentence-similarity`

# Semantics

- testing *consistency* in representation of monolingual vs CS text
- fine-tune models on STS task using monolingual benchmarks
- STS of $(i_{es}, j_{es})$ and $(i_{en}, j_{en})$ similar to $(i_{cs}, j_{cs})$

$$sim(S_i^{l_1}, S_j^{l_2}) = sim(S_i^{cs}, S_j^{l}) \quad (1)$$

# Semantics Results

| l-pair-1 | l-pair-2 | cosine spearman | | |
|---|---|---|---|---|
| | | mBERT | XLM-R-base | XLM-R-large |
| en-en | cs-cs | **0.8503** | 0.8208 | 0.8256 |
| es-es | cs-cs | **0.7892** | 0.7655 | 0.7799 |
| en-es | cs-en | 0.8695 | 0.8656 | **0.8704** |
| en-es | cs-es | **0.7266** | 0.6947 | 0.7200 |

Table: Spearman rank statistic for the cosine similarity between language pair 1 (l-pair-1) and language pair 2 (l-pair-2).

# What do these results mean?

experiments show...

- models are generalising to handle CS text even when not explicitly trained to handle CS text.
- models may be representing CS text in their own way, not necessarily aligning with popular CS linguistic theories.
- seem to capture syntactic structure and semantic meaning in CS text.
- ablation studies show that performance degrades when using synthetic CS text, naturalistic CS matters?

# Going forward

- use other language pair - does it work because Spanglish? Hinglish?
- expand to using generative, decoder only models
- generate synthetic CS text from more state-of-the-art models (GPT-4)
- expanding to exploring bias in language vs models

# Thank you

Thank you for listening!

# Bibliography


- De Leon, Frances Adriana Laureano, Florimond Guéniat, and Harish Tayyar Madabushi (2020). "CS-Embed at SemEval-2020 Task 9: The effectiveness of code-switched word embeddings for sentiment analysis". In: *arXiv preprint arXiv:2006.04597*.
- Hewitt, John and Christopher D Manning (2019). "A Structural Probe for Finding Syntax in Word Representations". In: *NAACL*.