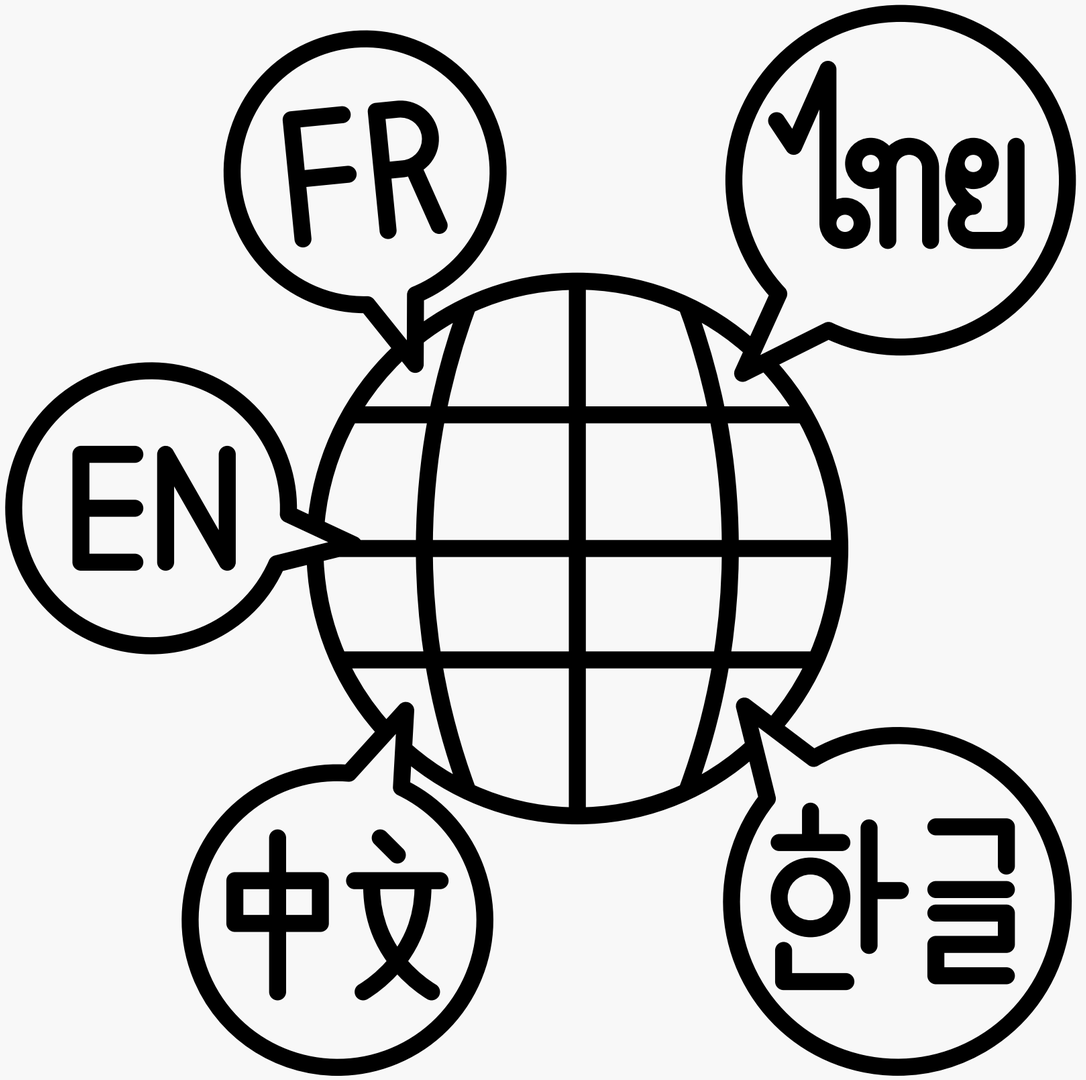
Codemixed text

Team - Redlock

Aung 6411325
Tanat 6410381



What is Codemixed text?

Codemixed text or "code-switching"

It is the phenomenon where two or more languages are **mixed** in the **same utterance, sentence, or discourse.** This can occur at various levels: at the word level, phrase level, or even sentence level.

Example

Sentence:

- "Wah, the chicken rice at that hawker centre very shiok ah!"
  - "Wah" is an exclamation of surprise or admiration.
  - "shiok" is a Malay-derived term that means satisfying or delicious.
  - "ah" is a particle, often used for emphasis or affirmation.

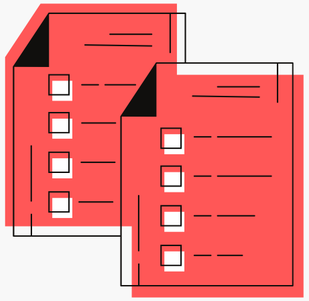
Phrase:

- "kiasu mentality"
  - "kiasu" is a Hokkien-derived term that means 'afraid to lose'. The phrase refers to a fear of missing out or always wanting to be the best and is often used to describe a competitive nature or mindset.

Word:

- "timepass"
  - This is a common term used in India and among Indian communities to describe activities done to pass the time. It combines the English word "time" with the Hindi word "paas" (meaning "pass" or "spend").

Challenges in processing codemixed text.



Inconsistency in Grammatical Structures

Different languages have distinct grammatical rules.

Vocabulary Overlaps

Some words might exist in both languages but with different meanings.

Morphological Ambiguities

The same word might have different morphological structures in different languages.

Scarcity of Annotated Data

There's often a lack of large, annotated datasets for codemixed text, making supervised machine learning tasks challenging.

Additional Complexity

- **Sociolinguistic Factors**

  Codemixing isn't just linguistic but also carries sociocultural and contextual meaning.

- **Phonetic Challenges**

  If the codemixed text is derived from speech, pronunciation variations can introduce another layer of complexity.

- **Named Entity Recognition (NER) Challenges**

  Named entities might be expressed differently in codemixed scenarios.

Language Identification of Code-Mixed Text (Code-mixed LID)

What is Code-mixed LID?

Determining the language in which a piece of text is written

- Identifying and labeling each segment (or token) of the text with its respective language.

- Addressing challenges that arise due to the blending of linguistic structures from multiple languages.

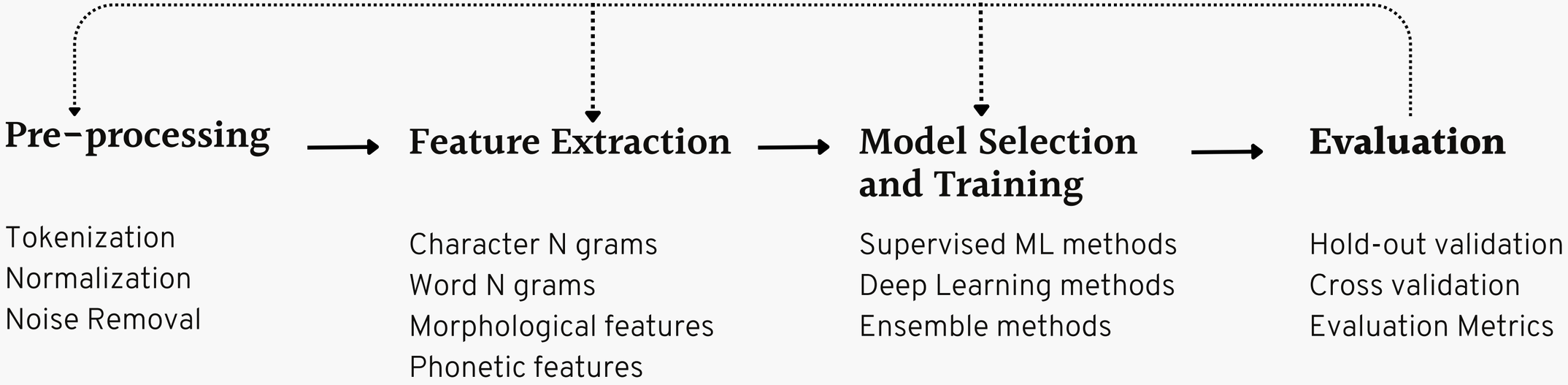
Why do we do LID

Improving NLP Performance: Understanding which language is being used can help apply the correct models or lexicons, leading to better performance in NLP tasks.

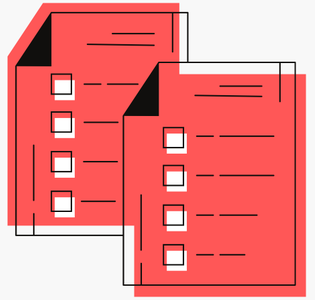
Challenges

word ambiguity.
khana - "eat" in Hindi but "short-time" in Burmese.

Code-mixed LID pipeline



Citations



AHMAD FATHAN HIDAYATULLAH , ATIKA QAZI , DAPHNE TECK CHING LAI , (Member, IEEE), AND ROSYZIE ANNA APONG
A Systematic Review on Language Identification of Code-Mixed Text: Techniques, Data Availability, Challenges, and Framework Development

Gazi Imtiyaz Ahmad, Jimmy Singla, Anis Ali, Aijaz Ahmad Reshi and Anas A. Salameh,
“Machine Learning Techniques for Sentiment Analysis of Code-Mixed and Switched Indian Social Media Text Corpus - A Comprehensive Review” International Journal of Advanced Computer Science and Applications(IJACSA), 13(2), 2022.

Gazi Imtiyaz Ahmad, Jimmy Singla, Anis Ali, Aijaz Ahmad Reshi and Anas A. Salameh,
“Machine Learning Techniques for Sentiment Analysis of Code-Mixed and Switched Indian Social Media Text Corpus - A Comprehensive Review” International Journal of Advanced Computer Science and Applications(IJACSA), 13(2), 2022.