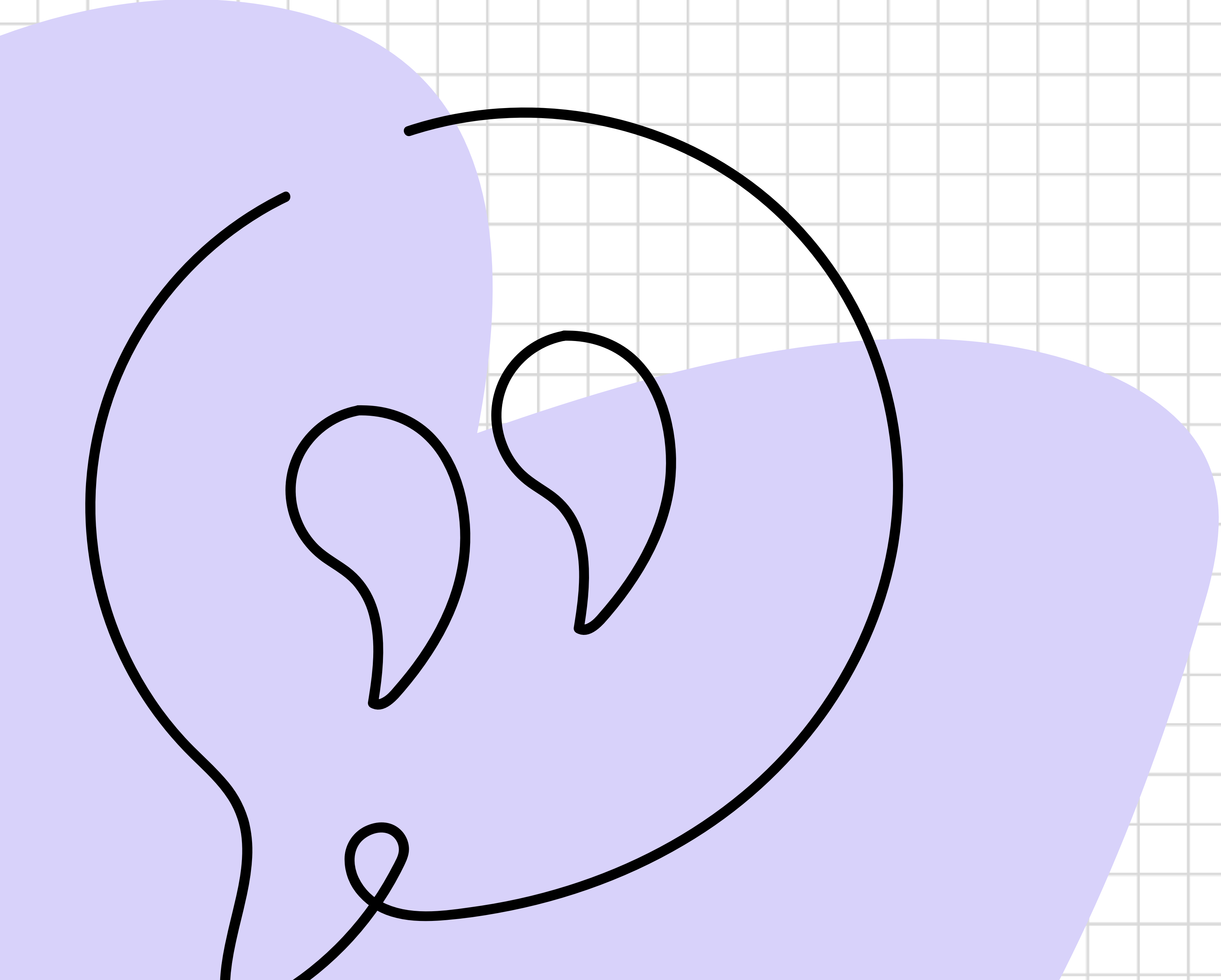
SENTIMENT ANALYSIS FOR ENGLISH <-> THAI CODE-MIXED TEXTS

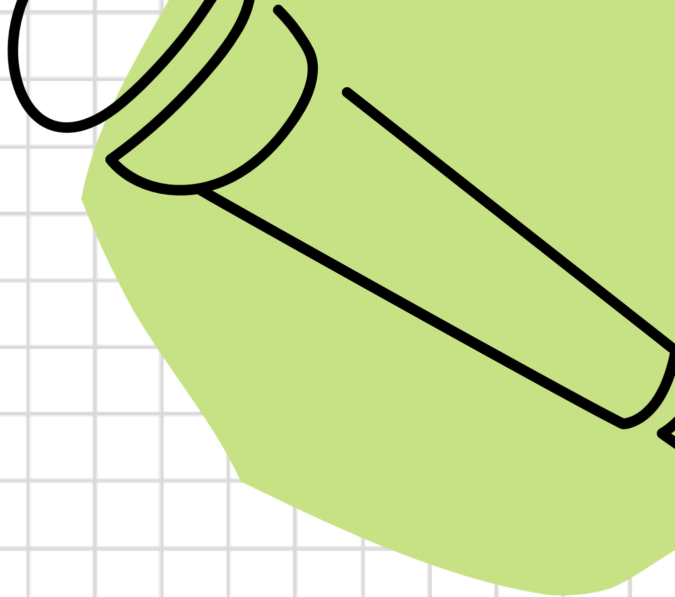
MEMBER

1.	Thit Lwin Win Thant	6540122
2.	Kaung Khant Lin	6540131
3.	Thust Thongsricharoen	6714508

CONTENTS



1. Project Overview
2. Objectives & Goals
3. Methodology
4. Technical Deep Dive
5. Expected Outcomes
6. Reference



OVERVIEW

- **The Context:** Real-world social media and comment streams often feature heavy language mixing, slang, and intentional misspellings.
- **The Problem:** Standard models often struggle with the "wild west" of bilingual internet slang and informal scripts.
- **The Solution:** A lightweight, practical system combining heuristic preprocessing tools with transformer-based fine-tuning.

OBJECTIVES & GOALS

TURNING CODE-MIXED CHAOS INTO CLEAR DATA

- **Classification:** Achieve accurate sentiment labeling (Positive, Negative, Neutral) for mixed-language sentences.
- **Identification:** Perform token-level language identification using fast heuristics rather than heavy, slow models.
- **Robustness:** Implement normalization for slang and misspellings to improve model reliability.
- **Efficiency:** Create a pipeline with minimal manual work by using automated “silver labels” and a high-quality "gold set" for final evaluation.
- **Key Deliverable:** A working sentiment classifier, a demo web app, and a detailed failure analysis report.

METHODOLOGY

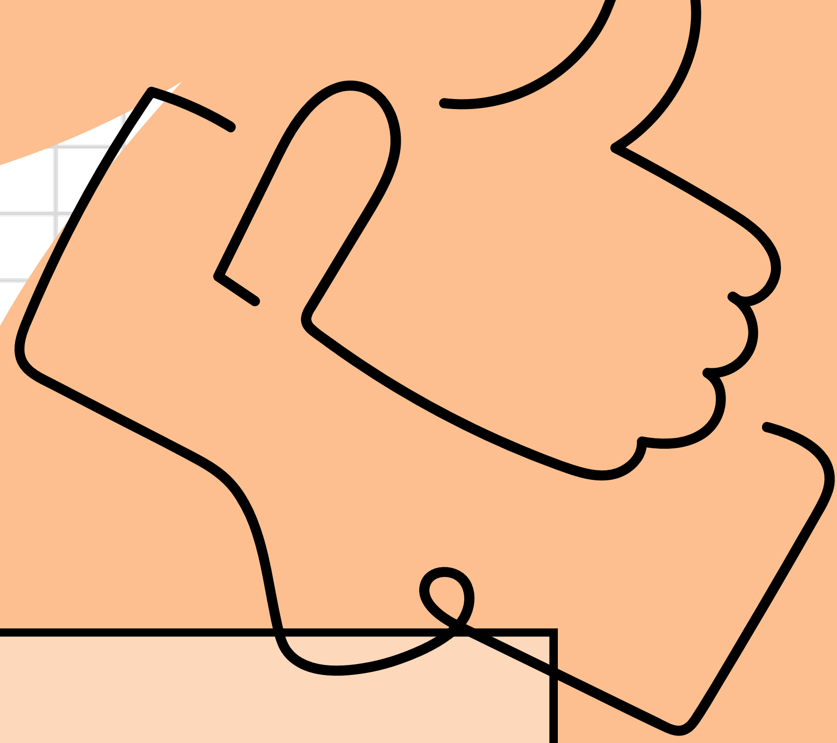
Phase	Key Activities
Data Collections	Scrape tech/gaming comments (9arm, Bay Riffer); integrate Wisesight (Thai) and SST-2 (English).
Using Necessary Tools	Implement PyThaiNLP tokenization and a regex-based Language ID (Unicode ranges).
Fine-tune Modeling	Fine-tune XLM-RoBERTa-base (XLM-R).
Launch	Evaluate using Macro F1 and build a web app demo for interactive testing.

TECHNICAL DEEP DIVE

Preprocessing & Fine-Tuning Strategy

- **Heuristic Normalization:** A slang dictionary (slang_dictionary.json) handles common terms, while Levenshtein distance provides fuzzy matching for variations.
- **Language ID:** Token-level tagging uses Unicode ranges (Thai script -> TH, Latin -> EN) to keep the pipeline lightweight.
- **Model Selection:** XLM-R was chosen for its superior performance in Thai contexts compared to standard mBERT.
- **Evaluation:** Focus includes a confusion matrix to analyze performance when English exceeds 50% of the sentence.

EXPECTED OUTCOMES



Practicality Meets Performance

- **The "Silver" Dataset:** A large silver-labeled corpus supplemented by a manual 200-sentence "Gold Set" for precision testing.
- **The Toolkit:** A full suite including the tokenizer, regex-based LID, and normalization scripts.
- **The App:** An interactive web interface providing sentiment scores.
- **The Verdict:** A pipeline that balances speed, accuracy, and interpretability—ideal for both academic evaluation and real-world experimentation.

REFERENCES

- https://github.com/PyThaiNLP/wisesight-sentiment
- https://huggingface.co/datasets/stanfordnlp/sst2
- https://huggingface.co/FacebookAI/xlm-roberta-base
- https://huggingface.co/google-bert/bert-base-multilingual-cased
- https://github.com/sagorbrur/codeswitch

THANK YOU