SVCE

Sri Venkateswara
College of
Engineering

FEEL TAMIL: ENHANCING SENTIMENT UNDERSTANDING WITH AI AND NLP

SUPERVISOR : SIVAKUMARESWARAN

RASHMIKA R S - 2127220801071

SOWNDARYA S - 2127220801090

VIJAY M - 2127220801108

PROBLEM STATEMENT

Sentiment analysis in Tamil remains a challenging task due to the lack of well-annotated datasets, linguistic complexity, and code-mixed text usage (Tamil-English). Existing sentiment analysis models primarily focus on English and high-resource languages, leading to inaccurate predictions for Tamil text. Furthermore, sarcasm detection in Tamil is an underexplored area, making it difficult to distinguish between genuine opinions and sarcastic remarks.

This project aims to develop a robust Tamil sentiment analysis system that can accurately classify Tamil text into Positive, Negative, Neutral, and Sarcastic categories using state-of-the-art transformer models like MuRIL, IndicBERT, and XLM-RoBERTa.

NEED AND MOTIVATION

With Tamil being spoken by over 75 million people worldwide, there is a growing need for NLP models that can effectively understand and process Tamil text. Businesses, government agencies, and researchers require sentiment analysis tools for applications such as social media monitoring, customer feedback analysis, and opinion mining.

Current Limitations in Tamil Sentiment Analysis

- Limited availability of annotated datasets for sentiment classification.
- High reliance on English-based models, leading to poor understanding of Tamil text.
- Lack of sarcasm detection models tailored for Tamil.

This project aims to bridge this gap by developing a high-accuracy sentiment analysis system for Tamil using deep learning models.

NOVELTY

Multilingual Transformer-Based Approach

- Uses deep learning models (MuRIL, IndicBERT, XLM-RoBERTa) instead of traditional methods like TF-IDF and SVM.
- Fine-tuned specifically for Tamil sentiment analysis.

Sarcasm Detection using Helinivan & English-Sarcasm Detector

- Incorporates a special sarcasm classification layer.
- One of the first Tamil sentiment analysis models to explicitly detect sarcasm.

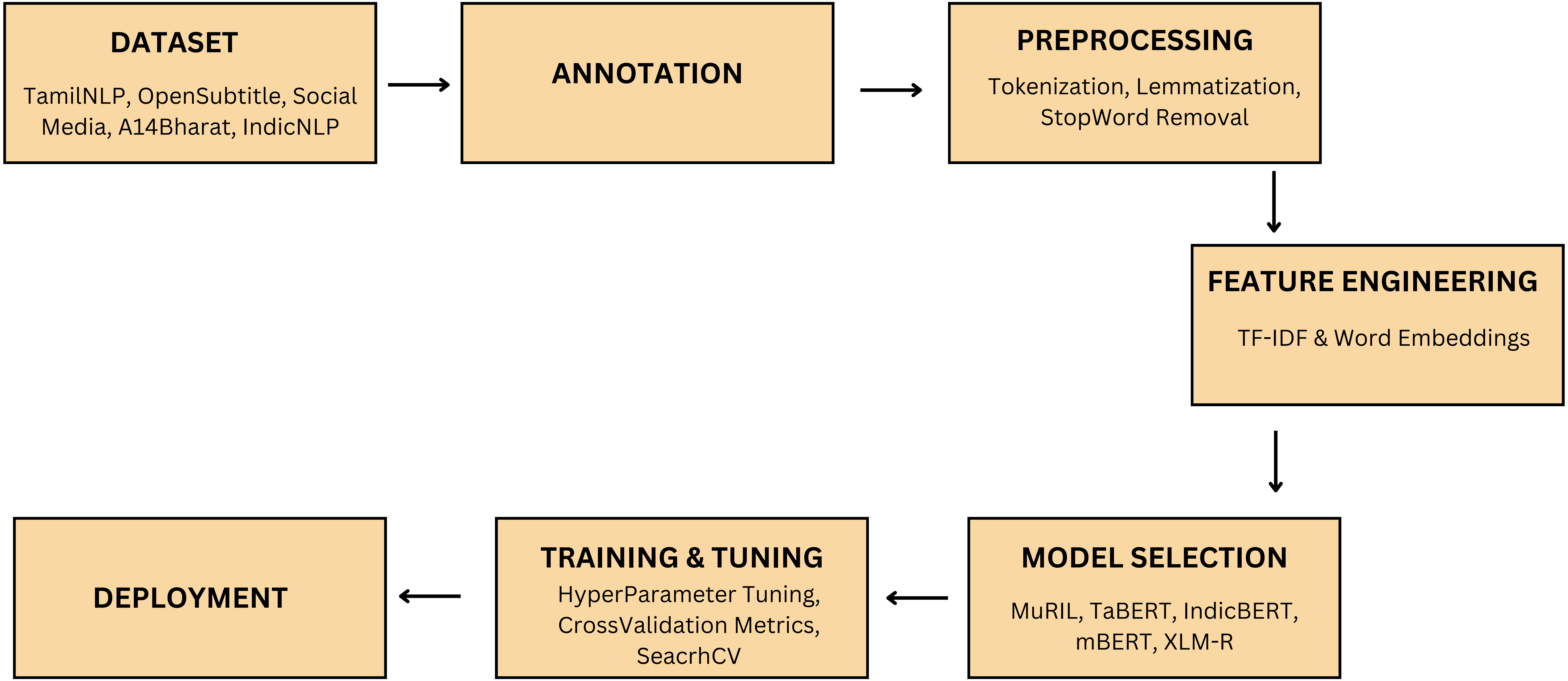
Real-World Deployment

- Deployed as an API using FastAPI & Docker.
- Enables real-world applications like social media monitoring, chatbot sentiment detection, and feedback analysis.

Benchmarking & Optimization

- Evaluates multiple models (MuRIL, TaBERT, IndicBERT, XLM-RoBERTa).
- Uses GridSearchCV and hyperparameter tuning to select the best-performing model

WORKFLOW



DATASET

TamilNLP, OpenSubtitle, Social Media, A14Bharat, IndicNLP

ANNOTATION

PREPROCESSING

Tokenization, Lemmatization, StopWord Removal

FEATURE ENGINEERING

TF-IDF & Word Embeddings

MODEL SELECTION

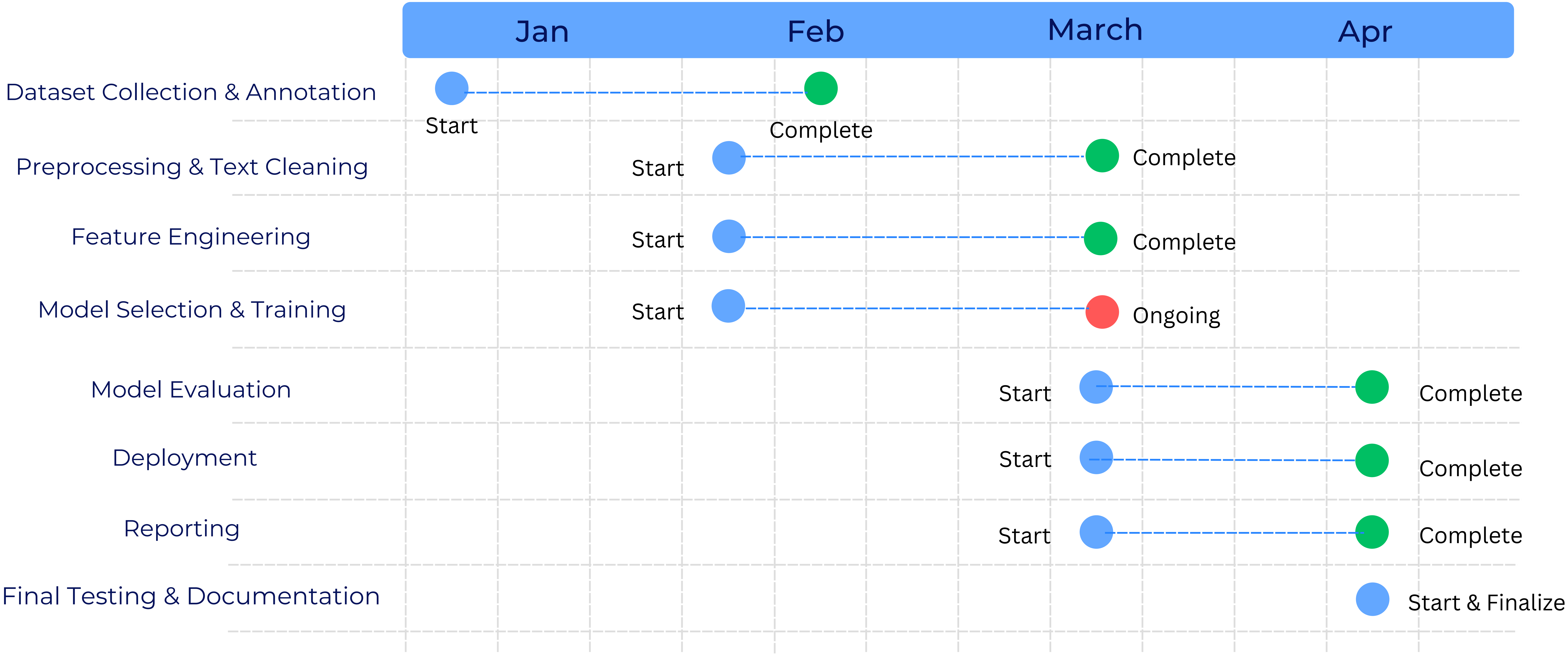
MuRIL, TaBERT, IndicBERT, mBERT, XLM-R

TRAINING & TUNING

HyperParameter Tuning, CrossValidation Metrics, SeacrhCV

DEPLOYMENT

ACTION PLAN – GANTT CHART:



Jan Feb March Apr

Dataset Collection & Annotation: Start – Complete

Preprocessing & Text Cleaning: Start – Complete

Feature Engineering: Start – Complete

Model Selection & Training: Start – Ongoing

Model Evaluation: Start – Complete

Deployment: Start – Complete

Reporting: Start – Complete

Final Testing & Documentation: Start & Finalize

PO's ADDRESSED :

PO1 - Engineering Knowledge

PO2 - Problem Analysis

PO3 - Development of solutions

PO4 - Conduct investigations of complex problems

PO5 - Modern Tool Usage

PO6 - The Engineer and Society

PO8 - Ethics

PO12 - Life-long Learning

PSO's ADDRESSED :

PSO13 - Bussiness Operations & IT Solutions

The Project enables bussiness to analyze customer sentiments, market trends and feedback using NLP, supporting data-driven decision-making

PSO14 - It Infrastructure & Data Security

The project involves handling textual data securely, ensuring privacy, and optimizing computational resources for efficient sentiment analysis.

REFERENCES :

Base Paper: Theedhum Nandrum@Dravidian-CodeMix-FIRE2020: A Sentiment Polarity Classifier for YouTube Comments with Code-switching between Tamil, Malayalam and English

Comments: FIRE 2020, December 16-20, 2020, Hyderabad, India
Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG)
Cite as: arXiv:2010.03189 **[cs.CL]** (or arXiv:2010.03189v2 **[cs.CL]** for this version) https://doi.org/10.48550/arXiv.2010.03189

Reference Papers:

TY - BOOK, AU - Puranik, Karthik, AU - Bharathi, B., AU - Balasubramanian, Senthil Kumar, PY - 2021/11/15
T1 - **IIITT@Dravidian-CodeMix-FIRE2021: Transliterate or translate? Sentiment analysis of code-mixed text in Dravidian languages**
DO - 10.48550/arXiv.2111.07906

TY - JOUR, AU - Chakravarthi, Bharathi, AU - Asoka Chakravarthi, Ruba, AU - Muralidaran, Vigneshwaran, AU - Jose, Navya
PY - 2022/09/01
T1 - **DravidianCodeMix: sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text**
DO - 10.1007/s10579-022-09583-7, JO - Language Resources and Evaluation