




Malayalam Parser for Dataset Creation

Design Presentation

Guided by:
Dr. Mary Priya Sebastian

Fathima Jennath
Gautham C Sudheer
Godwin Gino
Mohammed Basil

Contents

- 
- Introduction
 - Problem Definition
 - Objectives
 - Functional Requirements of the Product
 - System Architecture
 - Datasets (if any)
 - UI Design
 - Work Division – Gantt Chart
 - Software/Hardware Requirements
 - Conclusion
 - References

Introduction



- Importance of Natural Language Processing (NLP) in regional languages
- Focus on the specific relevance of Malayalam in the context of NLP applications.
- Challenges associated with the scarcity of annotated datasets.
- Analyze both the syntactic and semantic structures of Malayalam sentences
- Applications such as sentiment analysis, named entity recognition, etc.
- Potential impact on advancing research and applications specific to the Malayalam

Problem Definition



To create a Malayalam Parser for dataset creation, involving data collection, preprocessing, manual annotation, and training using various parsing approaches to address the scarcity of annotated datasets in Malayalam for NLP applications.

Objectives



- Data Collection
 - a) Gather text data from diverse sources in Malayalam language
 - b) Aim for a sufficient volume of data to represent the language's usage patterns adequately
- Data Preprocessing
 - a) Perform tokenization, normalization, and cleaning of the collected data
 - b) Handle any inconsistencies or noise in the data to ensure quality

Objectives



- Manual Annotation
 - a. Annotate a representative subset of the preprocessed data with grammatical and syntactic information
 - b. Employ linguistic experts or proficient annotators to ensure accurate annotations.
- Parser Development
 - a. Train the parser using the annotated dataset to understand Malayalam syntax and semantics

Functional Requirements



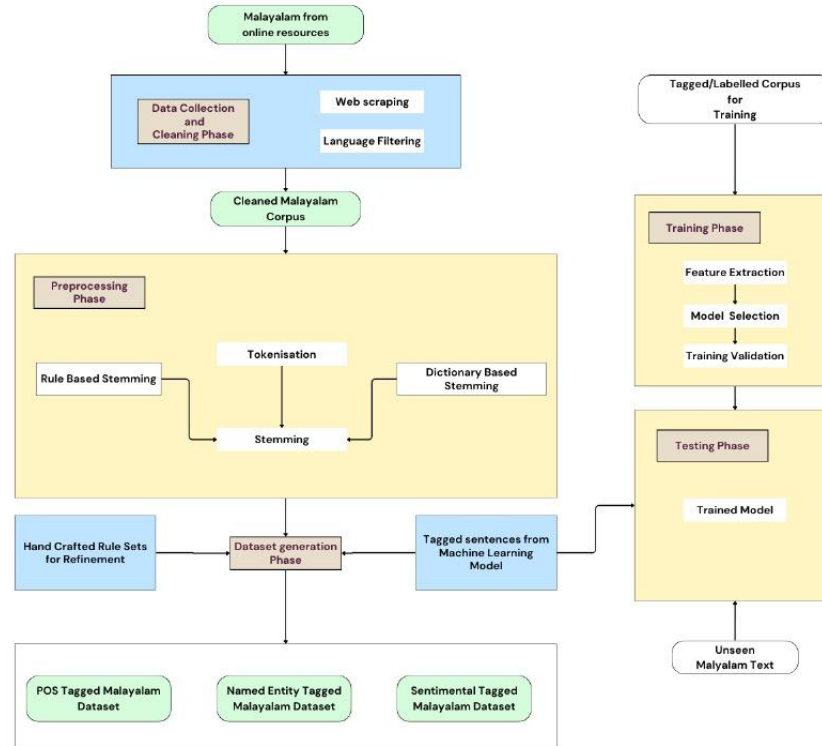
- Parse and analyze Malayalam language text to identify linguistic components such as words, phrases, and sentences.
- Determine grammatical structure, syntax, and semantics of Malayalam sentences to facilitate accurate linguistic analysis.
- Provide functionality for part-of-speech tagging, syntactic parsing, and semantic analysis tailored for the Malayalam language.

Functional Requirements

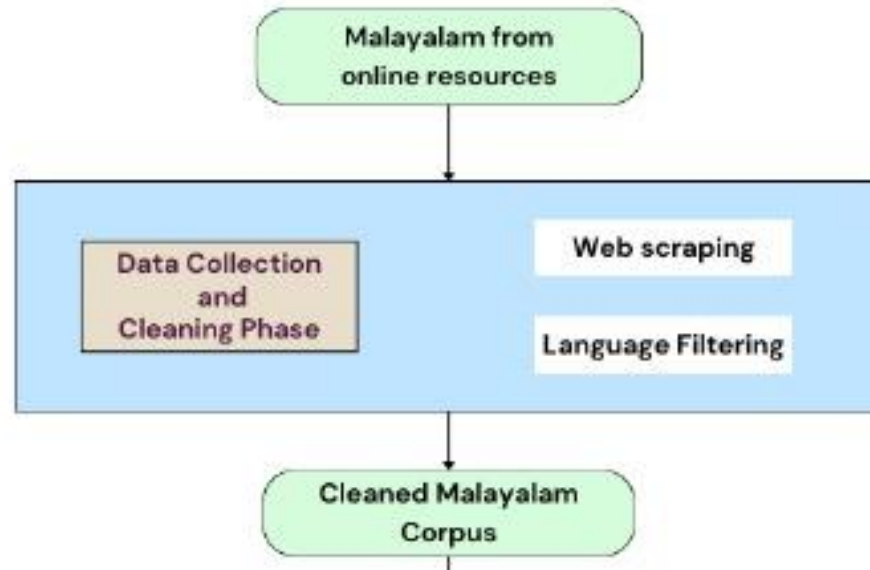


- Support for handling compound words, inflections, and variations in word forms commonly found in Malayalam text.
- Generation of a part-of-speech tagged dataset, named entity dataset, and sentimental tagged dataset, contributing to the advancement of language processing technologies in Malayalam
- Implement a user-friendly interface that allows users to input Malayalam text for analysis

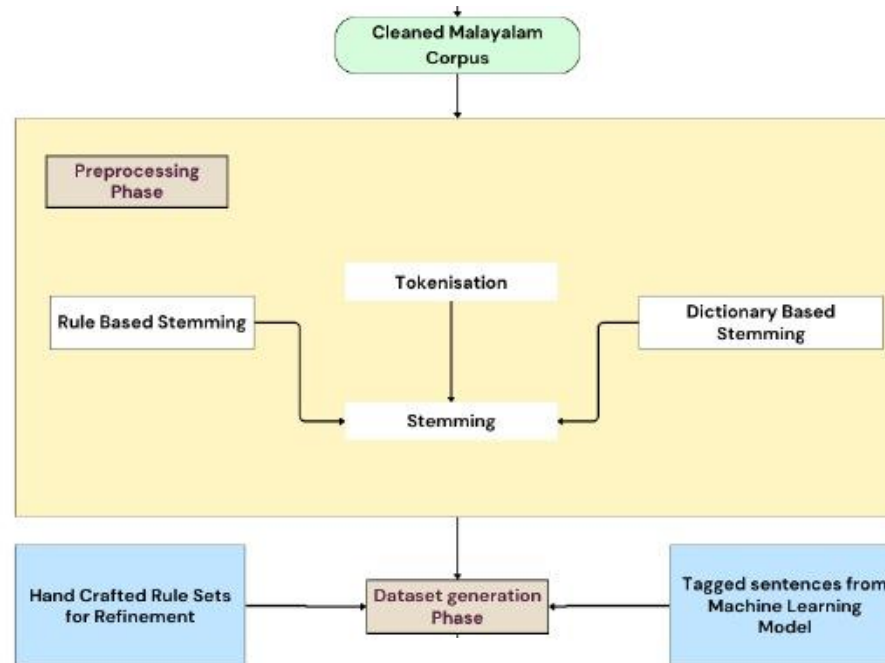
System Design



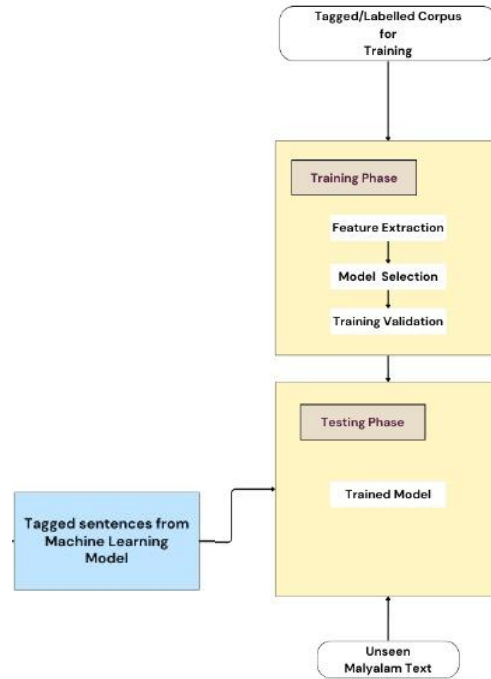
Data Collection and Cleaning Phase



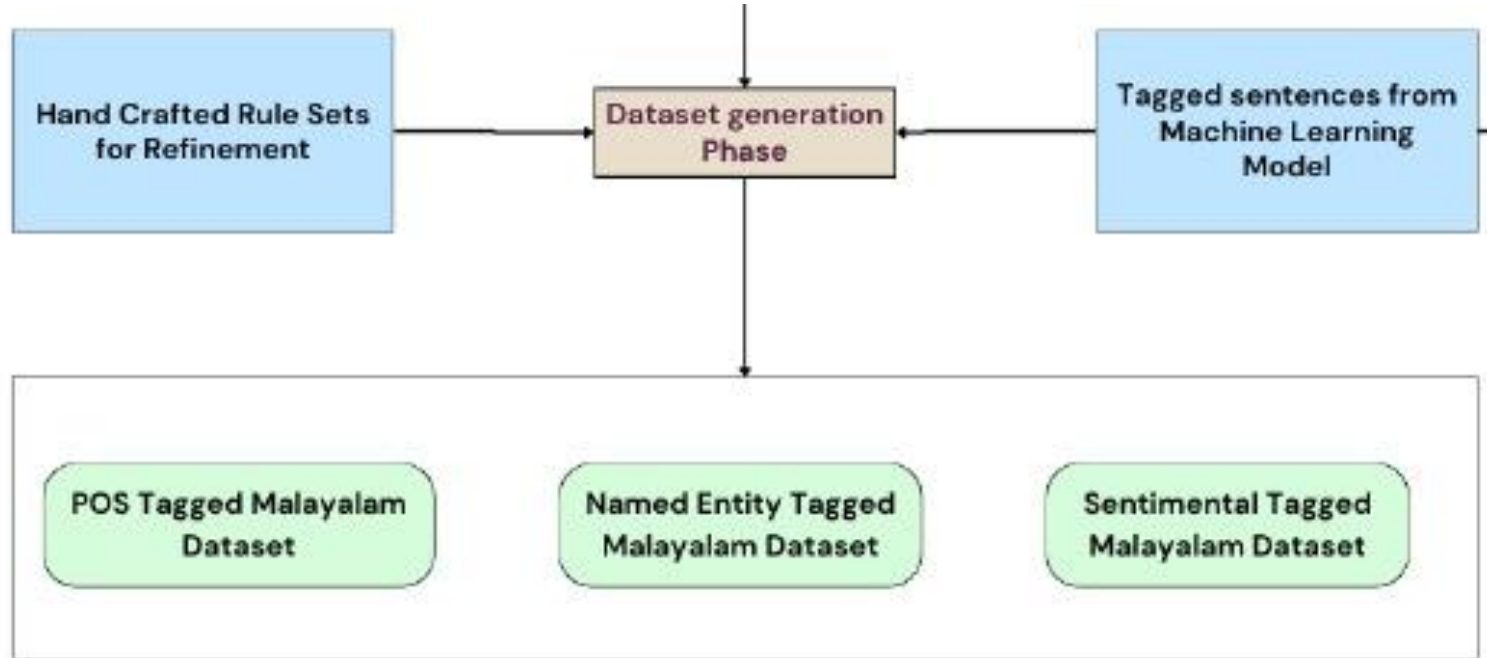
Data Preprocessing



Training and Testing



Output



UI Design

Malayalam Parser

Step into a world of linguistic exploration! Dive into our webpage to uncover existing datasets and transform your input into a mosaic of named entities, POS tags, and sentiment analysis.

Try now, enter text

Submit

Click here to download and use our data sets

Name Entity 

POS Tagged 

Sentimental 



UI Design

Malayalam Parser

Step into a world of linguistic exploration! Dive into our webpage to uncover existing datasets and transform your input into a mosaic of named entities, POS tags, and sentiment analysis.

Try now, enter text

Submit

Results

Name Entity

POS Tagged

Sentimental Analysis

Click here to download and use our data sets

Name Entity 

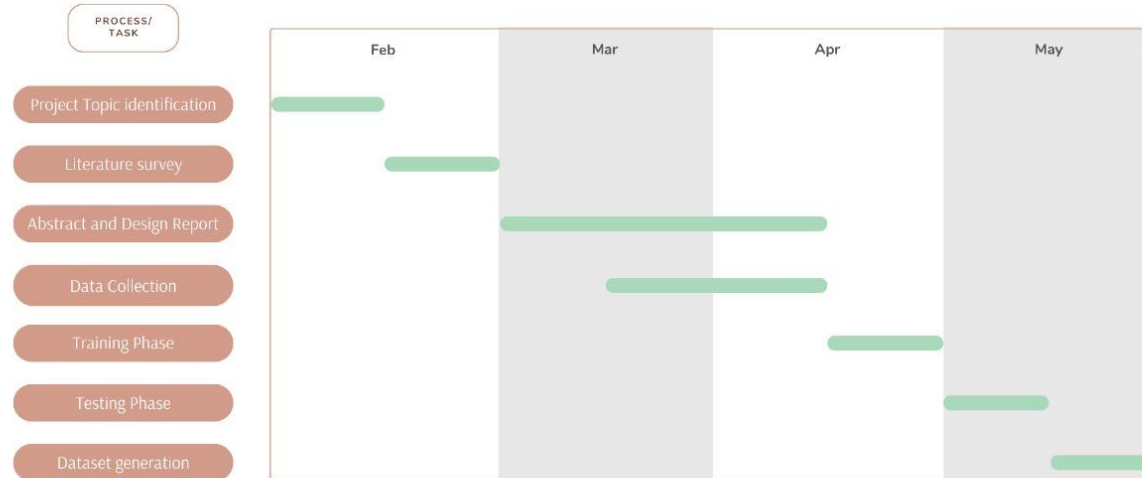
POS Tagged 

Sentimental 



Work Division

Gantt chart



Software / Hardware Requirements



- Windows 10 or later
- MacOS 10.13 High Sierra or later
- Ubuntu 18.04 LTS or later
- A modern processor (e.g., Intel Core i5 or equivalent)
- Sufficient RAM (at least 4GB)
- Available storage space for software installation
- Python (version 3.6 or later)
- Other programming languages and frameworks suitable for NLP development like NLTK, spaCy, scikit-learn, TensorFlow, etc. may be necessary

Conclusion



A comprehensive Malayalam language processing tool facilitating accurate linguistic analysis and dataset generation for NLP applications.

- Parsing and analysis of Malayalam text, enabling identification of linguistic components and determination of grammatical structure, syntax, and semantics
- Generates part-of-speech tagged, named entity, and sentiment-tagged datasets
- Contribute significantly to the advancement of language processing technologies in Malayalam.



Thank you