Mini-Project Approval

Project Title : Paraphrasing in Kannada using NLP

Project Guide : Prof. Mamatha H R

Project Team :

Name	SRN
Anagha H M	PES1UG19CS057
Karthik Sairam	PES1UG19CS210

Problem Statement

“Paraphrase identification in Kannada using NLP”

- Paraphrase identification is a natural language processing problem that involves the determination of whether two text segments have the same meaning.
- Our goal is to implement this in the Kannada language.

Literature Survey

Paper 1: Paraphrase plagiarism identification with character-level features

Authors: Srivastava, Shruti & Govilkar, Sharvari

- Details on the approach of Natural Language Processing
- Paraphrase detection techniques
- Similarity Metrics used in NLP
- Need extra details for a language like Kannada, which are not explained here

PES UNIVERSITY

Literature Survey

Paper 2: A Survey on Paraphrase Recognition

Authors: Magnolini, Simone

- Defines Paraphrasing and explains different algorithms/approaches
- Comparison between different approaches for paraphrasing is given
- The Machine Learning approach mentioned specifies Support vector Machine

Literature Survey

Paper 3: An Eccentric Approach for Paraphrase Detection Using Semantic Matching and Support Vector Machine

Authors: P. Vigneshvaran, E. Jayabalan and A. V. Kathiravan

- The approach specified involves the following:

  1) Tokenization

  2) POS Tagging

  3) Token Match

  4) Token Count
- The features extracted are thus fed to the classifier
- SVM classifier is used for the same

Literature Survey

Paper 4: Detection of paraphrases for Devanagari languages using support vector machine

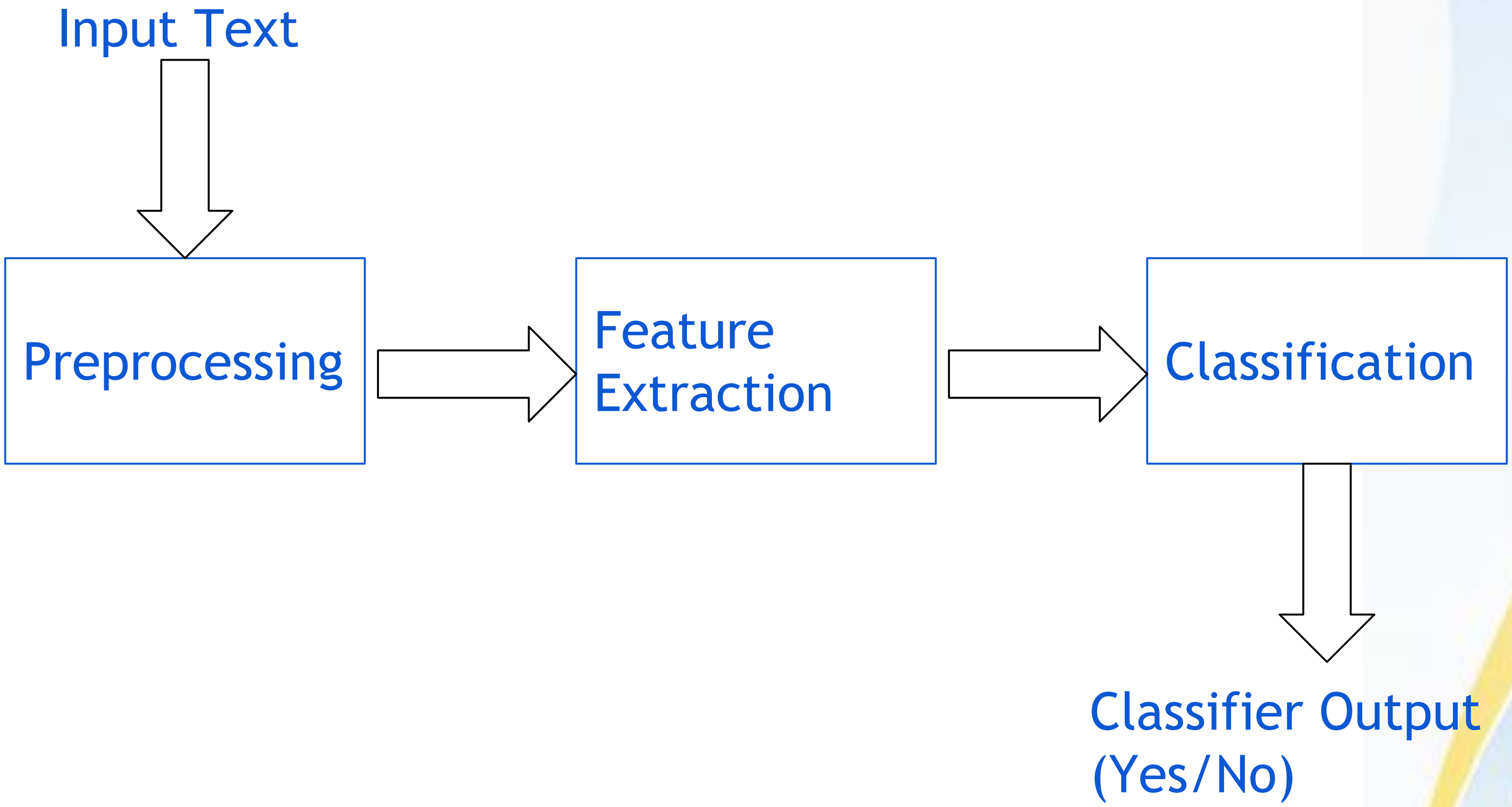
Authors: D. S. Bhole and S. S. Patil

- "For feature extraction, the following are done:
  - Tokenization
  - Stop-word elimination
  - Stemming
  - Synonyms Matching
- Done for Hindi and Marathi, hence an approach in Kannada would require changes in the methodology adopted in this paper.

Literature Survey

- We have completed the literature survey.
- The above mentioned papers are the main papers we are going to base our project on.
- Our literature survey consists of around 14 papers each having different approach on paraphrasing identification.
- It also consists of survey papers that summarize the approaches and lists their accuracy.

Proposed Solution



Input Text

Preprocessing

Feature Extraction

Classification

Classifier Output (Yes/No)

Technologies / Methodologies

Proposed platform and libraries:

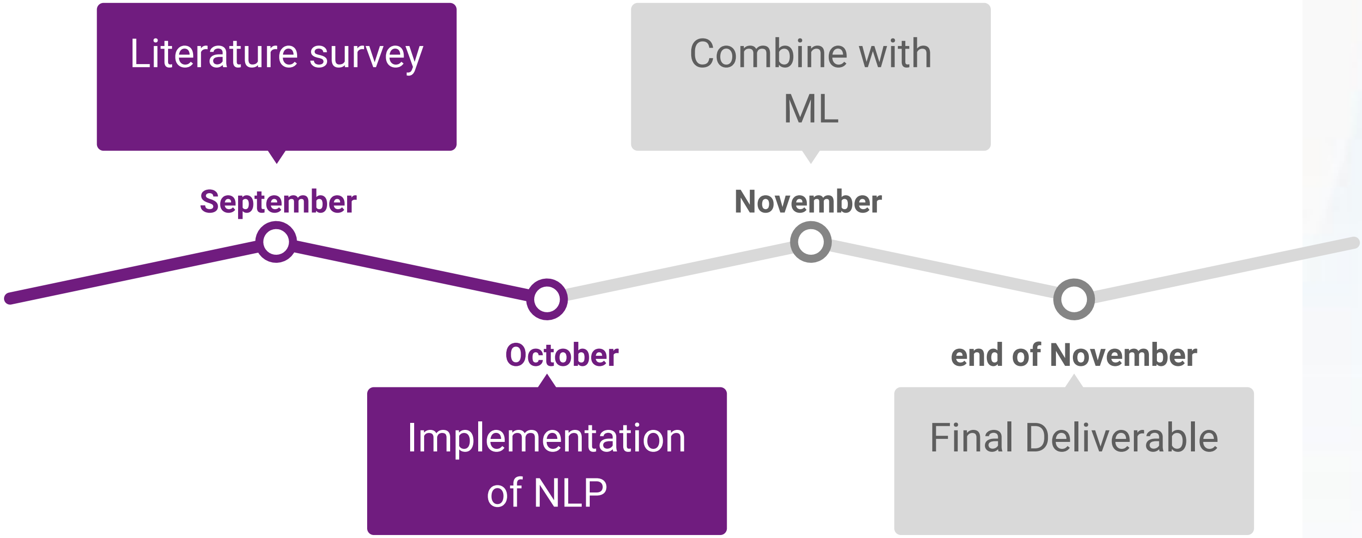
- Python 3.8
- Pandas
- Numpy
- scikit-learn (for classification)
- NLTK (Natural Language Toolkit)

Proposed methods:

- Token-related processes
- Tagging
- Classification

The Software Engineering method proposed is the Incremental approach.

Project Timelines & Plan



Literature survey

September

October

Implementation of NLP

Combine with ML

November

end of November

Final Deliverable

Bibliography

Paper 1: Srivastava, Shruti & Govilkar, Sharvari. (2017). A Survey on Paraphrase Detection Techniques for Indian Regional Languages. International Journal of Computer Applications. 163. 42-47. 10.5120/ijca2017913757.

Paper 2: Magnolini, Simone. (2014). A survey on paraphrase recognition. CEUR Workshop Proceedings. 1334. 33-41.

Paper 3: P. Vigneshvaran, E. Jayabalan and A. V. Kathiravan, "An Eccentric Approach for Paraphrase Detection Using Semantic Matching and Support Vector Machine," 2014 International Conference on Intelligent Computing Applications, 2014, pp. 431-434, doi: 10.1109/ICICA.2014.94.

Bibliography

Paper 4: D. S. Bhole and S. S. Patil, "Detection of paraphrases for Devanagari languages using support vector machine," 2018 International Conference on Communication information and Computing Technology (ICCICT), 2018, pp. 1-5, doi: 10.1109/ICCICT.2018.8325880.

Thank You