Group Name - LinguisticProcessors

Hindi Poem and Era Classifier

Chandrabhushan	-	200101027
Hareesh	-	200101071
Pramodh Billa	-	200101025
Sathvika	-	200101048

Overview

01 Problem Formulation

02 Proposed Goal

03 Achieved Goal

04 Contribution of each member

05 Details of implementation

06 Key papers

07 Lessons learnt

01

PROBLEM FORMULATION

Problem Formulation

- Much work done in English poems.

- **Our project** →An attempt to classify a given poem on the basis of Era and Poet for Hindi poems.

02

PROPOSED GOALS

Proposed Goals

- To create a database of Hindi poems using web crawling.
- Employ different models and various vectorization methods to improve the accuracy of classification.

03

ACHIEVED GOALS

Achieved Goals

- Created a dataset of around **40000 poems** from various websites like **kavitakosh** and **hindwi.**
- Tested various models to get a best accuracy of **94.72%** for **era** and **34.12%** for **poets.**

04

CONTRIBUTION OF ALL TEAM MEMBERS

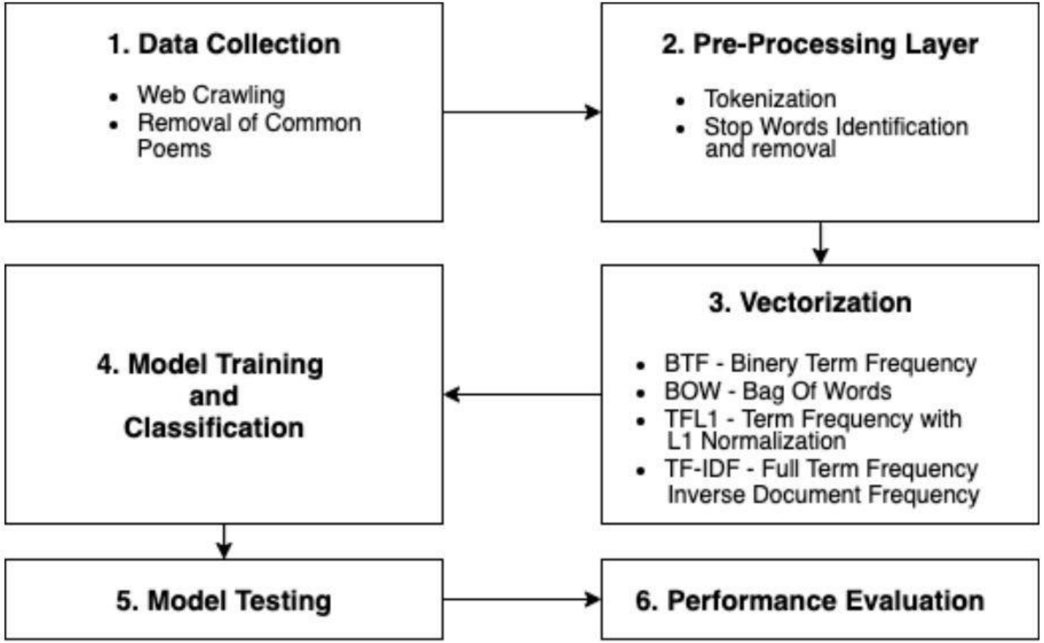
Contribution of Team members

- Dataset Creation → Pramodh and Sathvika
- Merging and Pre-processing → Hareesh
- Vectorization →Pramodh, Sathvika and Hareesh
- Model Training →
  - Cosine Similarity → Chandrabhushan
  - Logistic Regression → Chandrabhushan

05

IMPLEMENTATION / ANALYSIS

Workflow of the project



1. Data Collection

- Web Crawling
- Removal of Common Poems

2. Pre-Processing Layer

- Tokenization
- Stop Words Identification and removal

3. Vectorization

- BTF - Binery Term Frequency
- BOW - Bag Of Words
- TFL1 - Term Frequency with L1 Normalization
- TF-IDF - Full Term Frequency Inverse Document Frequency

4. Model Training and Classification

5. Model Testing

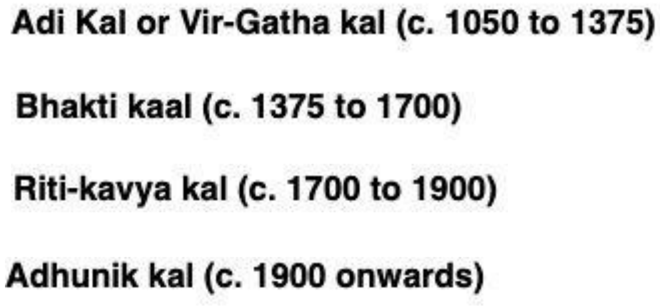
6. Performance Evaluation

Dataset Creation and Merging

- Used **Scrapy** and **BeautifulSoup** to scrap poems from **kavitakosh** and **hindwi** respectively to create a dataset of 40,000 poems.
- **Efforts** in understanding and using the libraries to build the web crawler.

Dataset Creation and Merging (Cont.)

- Dataset may contain many duplicate poems → **Merging**
- Used **year** in which poem was written for classification to create the final corpus. **Split** the poems in **9:1 ratio** appropriately in the **ratio of eras** to create the training set and test set respectively.



Adi Kal or Vir-Gatha kal (c. 1050 to 1375)

Bhakti kaal (c. 1375 to 1700)

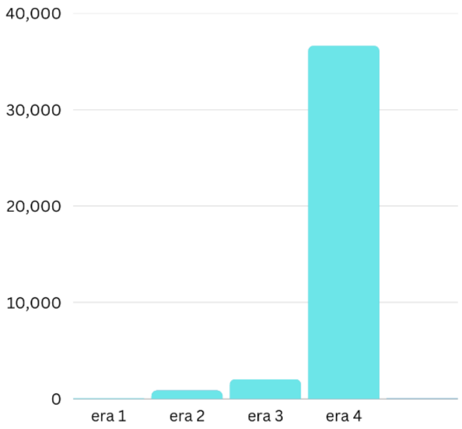
Riti-kavya kal (c. 1700 to 1900)

Adhunik kal (c. 1900 onwards)

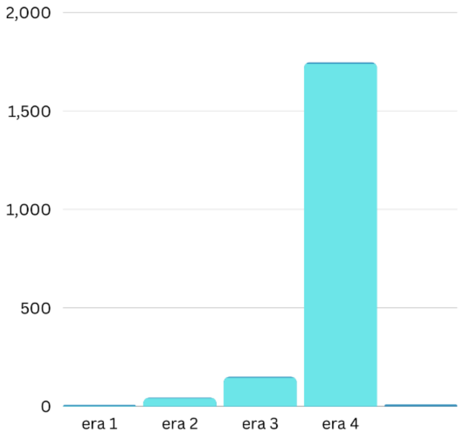
Various classes of poems

Dataset Creation and Merging (Cont.)

Dataset Classification



Poems in each era

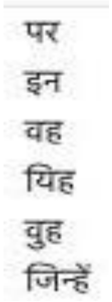


Poets in each era

Pre–Processing

- **Pre-processed** the data to remove reduce noise in data:
  - Numbers.
  - Punctuations.
  - White spaces.
- Used **tokenization** to split the poems into words. (Used **iNLTK**)
- Removed stop words using 3 different sets of words for optimization.

Stop words



पर
इन
वह
यिह
वुह
जिन्हें

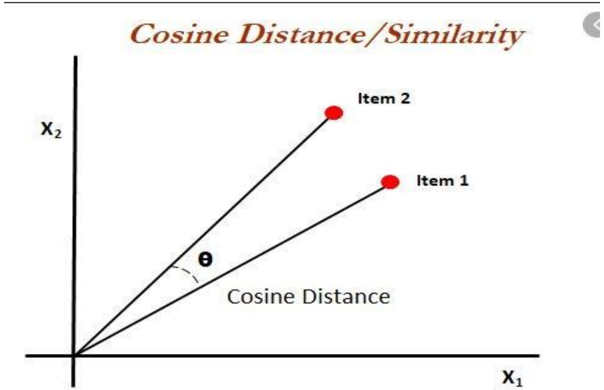
Vectorization

- Bag of Words
- Binary Term frequency
- L1 normalized Term frequency
- L2 normalized tf-idf

- These techniques provides long and sparse vectors

Cosine Similarity

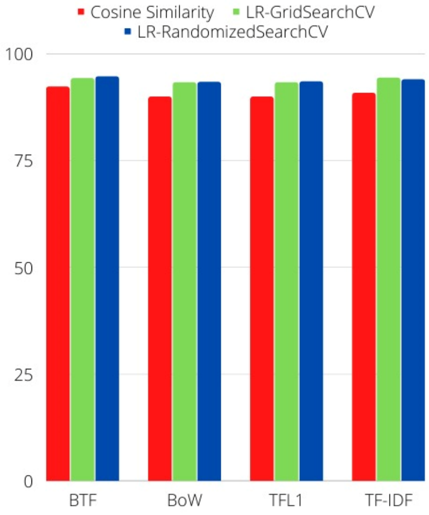
- Used **sklearn** to perform cosine similarity.
- For all testing data poems, found the poem in training data with smallest angle and returned its era and author name.
- **Accuracy** → **92.31%**and **14.78%** for era and poet respectively.
- Poet prediction → Why bad?
  - Maybe multiple authors have similar writing styles.
  - Same author has multiple writing styles.



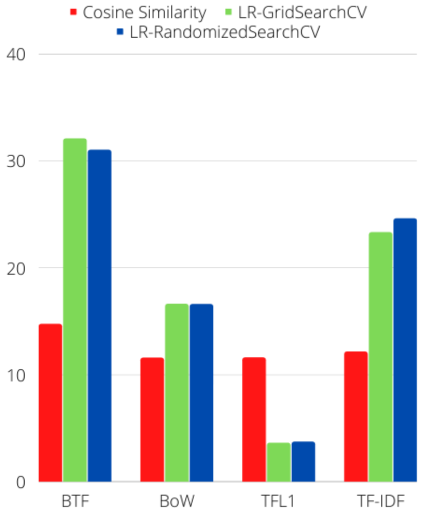
Logistic Regression

- Used **scikit-learn** library, to implement logistic regression with L2 regularization penalty.
- Tried two different hyper parameters tuning :-
  - GridSearchCV
  - RandomizedSearchCV
- **Accuracy -> 94.72%** and **34.12%** for era and poet prediction.

Final Results



Poem accuracy



Era accuracy

06

REFERENCES / KEY PAPERS

- **Title :** Automatic poetry classification using natural language processing.
  - Year - 2019
  - Journal - University of Ottawa
  - Author(s) - Vaibhav Kesarwani
- **Title :** Automated Analysis of Bangla Poetry for Classification and Poet Identification.
  - Year - 2015
  - Journal - IITB-Monash Research Academy
  - Author(s) - Geetanjali Rakshit, Anupam Ghosh, Pushpak Bhattacharyya, Gholamreza Haffari

07

LESSONS LEARNT

LESSONS LEARNT

01

Web Crawling

Implemented web crawlers from scratch

02

Vectorization

Implemented various vectorization methods

03

Models

Training and testing models & making them working together

THANK YOU