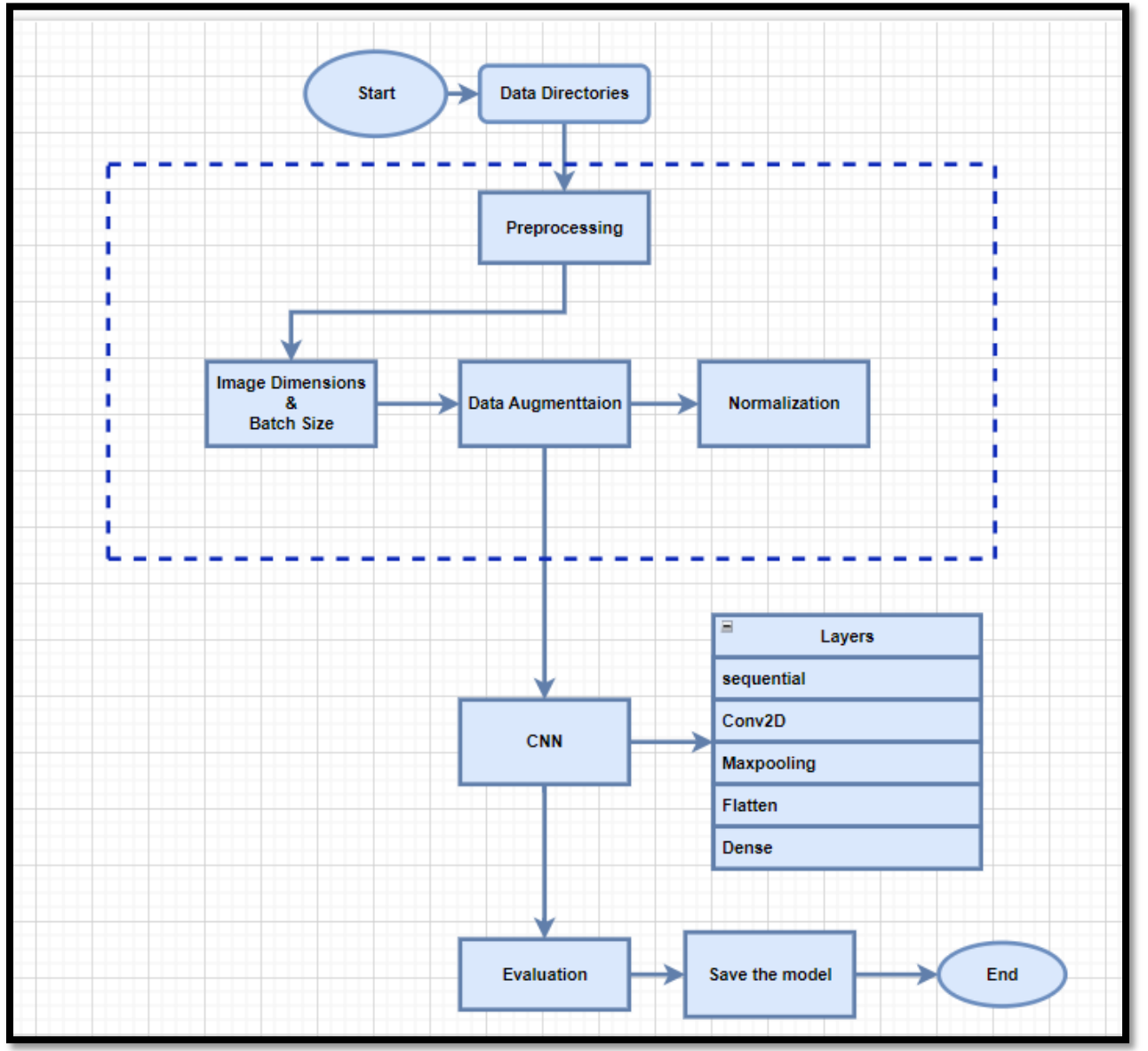


# Cross Language Phonetic Similarity

## 1. Image Classification



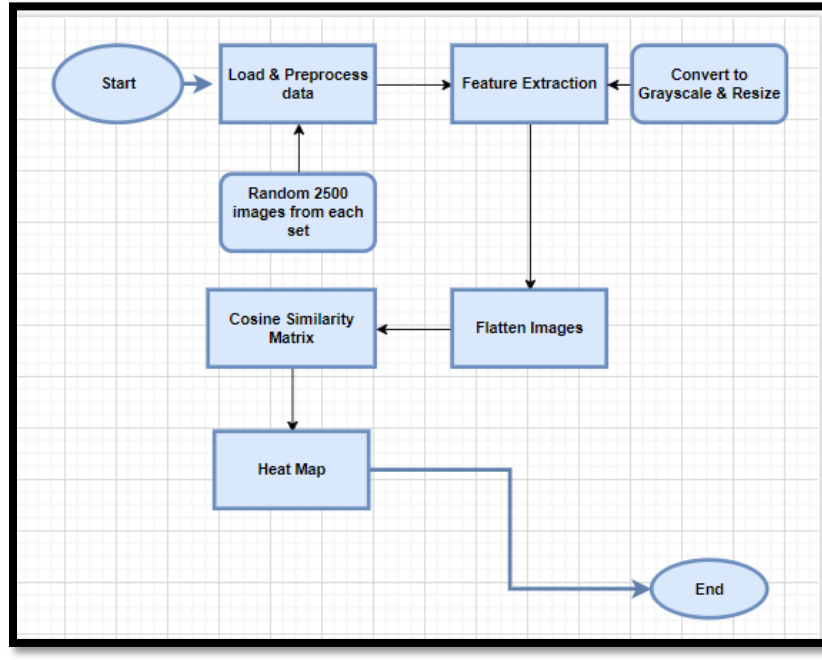
**Epochs:5**

**Oversampling in Dravidian**

- Gujarati and Marathi/Hindi (Devnagri Script)  
Result: overall classification accuracy obtained: 0.9857
- Tamil and Kanada (Dravidian Script)  
Result: overall classification accuracy obtained: 0.9649

## 2. Similarity Trend

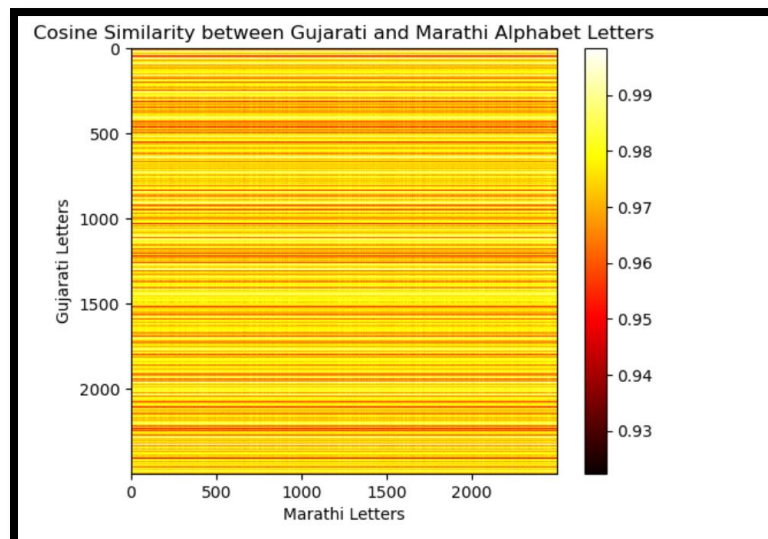
### a. Overall Random Trend



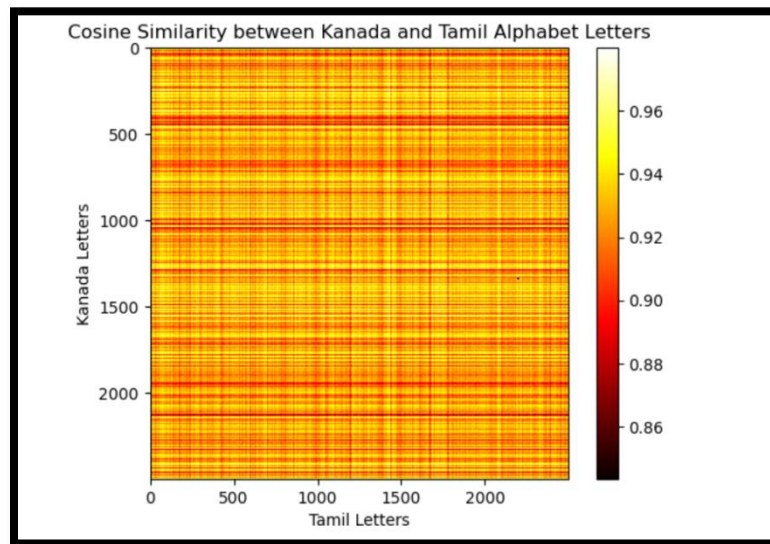
Cosine Similarity :

$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \cdot \sqrt{\sum_{i=1}^n B_i^2}},$$

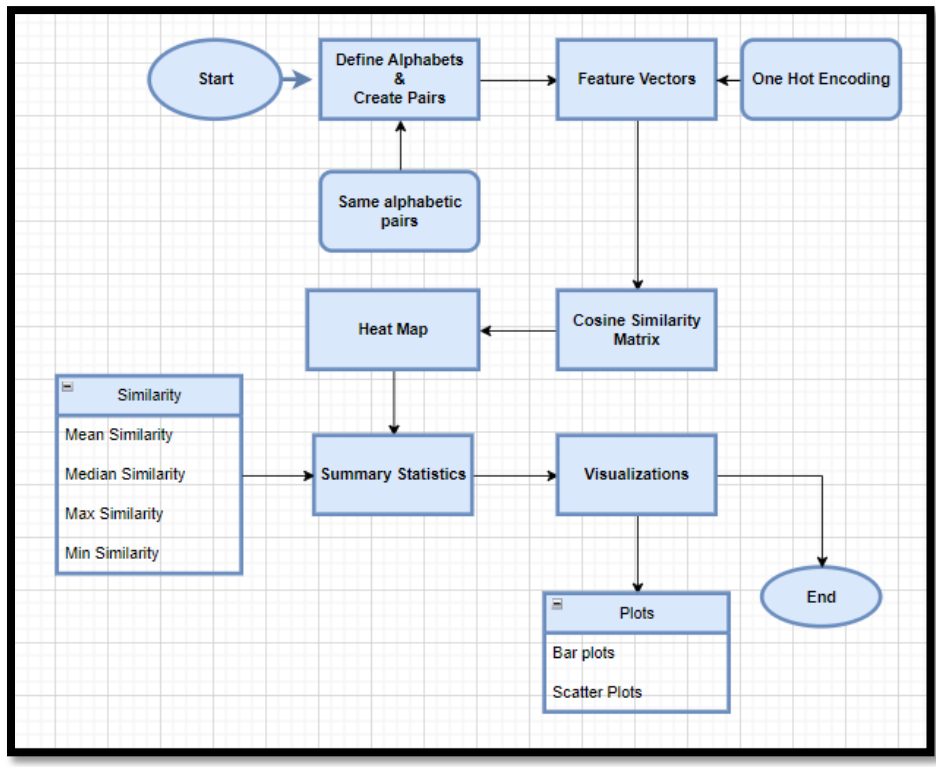
### i. Gujarati and Marathi/Hindi



ii. Tamil and Kanada



b. Letter Wise Similarity



Letter wise similarity code snippets and results.

i. Gujarati and Marathi/Hindi

```
# Define the alphabets of Gujarati and Marathi Languages
gujarati_alphabets = ['અ', 'આ', 'ઇ', 'ઈ', 'ઉ', 'ઊ', 'એ', 'ઐ', 'ઓ', 'ઔ', 'અં', 'અઃ', 'ક', 'ખ', 'ગ', 'ઘ', 'ઙ', 'ચ', 'છ', 'જ', 'ઝ', 'ઞ', 'ટ', 'ઠ', 'ડ', 'ઢ', 'ણ', 'ત', 'થ', 'દ', 'ધ', 'ન', 'પ', 'ફ', 'બ', 'ભ', 'મ', 'ય', 'ર', 'લ', 'વ', 'શ', 'ષ', 'સ', 'હ', 'ળ', 'ક્ષ', 'ઞ્', 'ઙ્', 'ઞ્']
marathi_alphabets = ['अ', 'आ', 'इ', 'ई', 'उ', 'ऊ', 'ए', 'ऐ', 'ओ', 'औ', 'अं', 'अः', 'क', 'ख', 'ग', 'घ', 'ङ', 'च', 'छ', 'ज', 'झ', 'ञ', 'ट', 'ठ', 'ड', 'ढ', 'ण', 'त', 'थ', 'द', 'ध', 'न', 'प', 'फ', 'ब', 'भ', 'म', 'य', 'र', 'ल', 'व', 'श', 'ष', 'स', 'ह', 'ळ', 'क्ष', 'ण्', 'ण्य', 'ञ्']

# Define the similar pairs of alphabets based on their phonetic sounds
similar_pairs = [
    ('अ', 'अ'), ('आ', 'आ'), ('इ', 'इ'), ('ई', 'ई'), ('उ', 'उ'), ('ऊ', 'ऊ'), ('ए', 'ए'), ('ऐ', 'ऐ'), ('ओ', 'ओ'), ('औ', 'औ'),
    ('अं', 'अं'), ('अः', 'अः'), ('क', 'क'), ('ख', 'ख'), ('ग', 'ग'), ('घ', 'घ'), ('ङ', 'ङ'), ('च', 'च'), ('छ', 'छ'), ('ज', 'ज'),
    ('झ', 'झ'), ('ञ', 'ञ'), ('ट', 'ट'), ('ठ', 'ठ'), ('ड', 'ड'), ('ढ', 'ढ'), ('ण', 'ण'), ('त', 'त'), ('थ', 'थ'), ('द', 'द'),
    ('ध', 'ध'), ('न', 'न'), ('प', 'प'), ('फ', 'फ'), ('ब', 'ब'), ('भ', 'भ'), ('म', 'म'), ('य', 'य'), ('र', 'र'), ('ल', 'ल'),
    ('व', 'व'), ('श', 'श'), ('ष', 'ष'), ('स', 'स'), ('ह', 'ह'), ('ळ', 'ळ'), ('क्ष', 'क्ष'), ('ण', 'ण')]

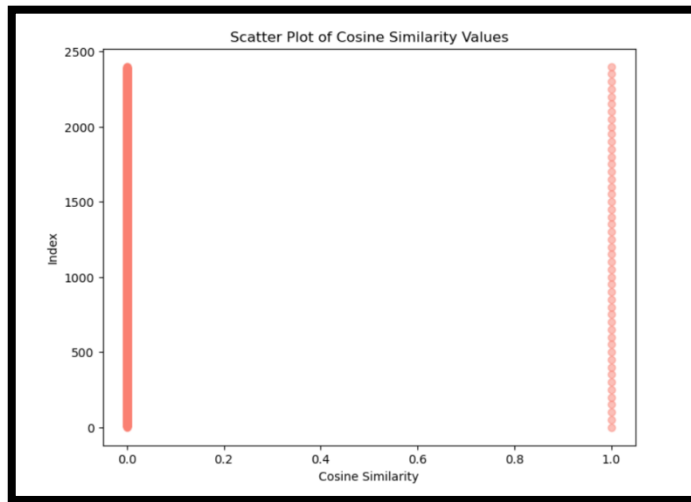
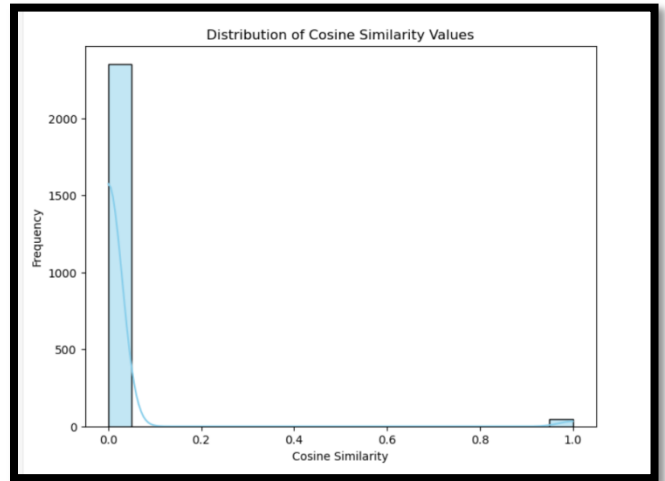
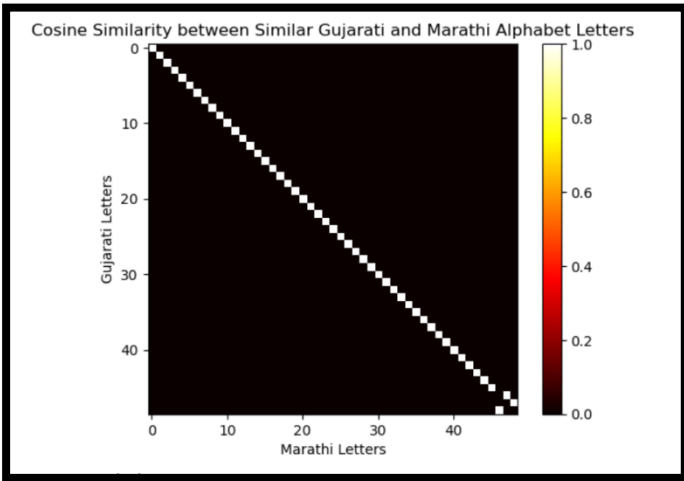
```

Mean Similarity: 0.02040816326530612

Median Similarity: 0.0

Maximum Similarity: 1.0

Minimum Similarity: 0.0



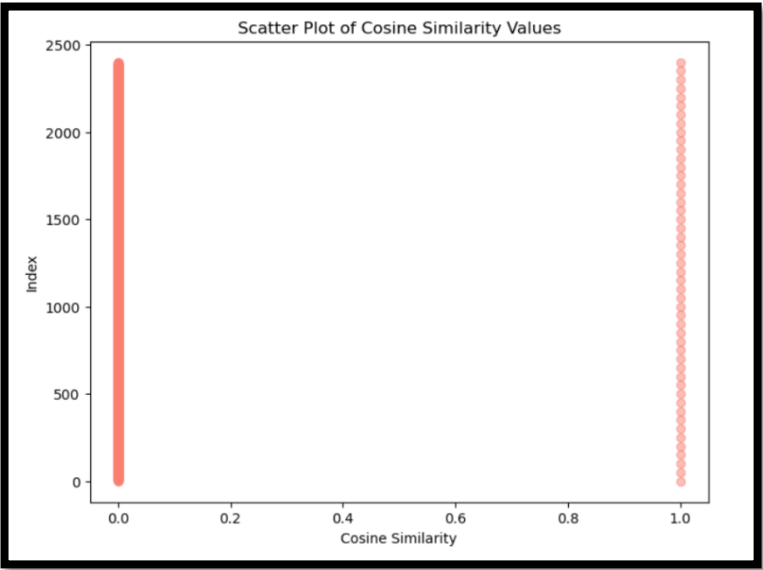
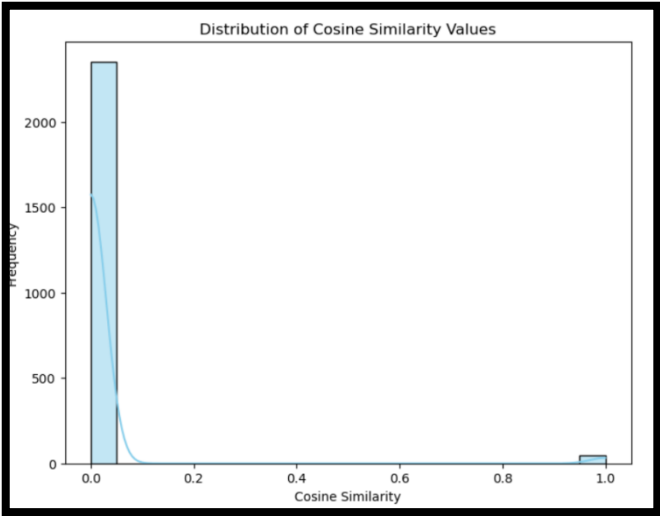
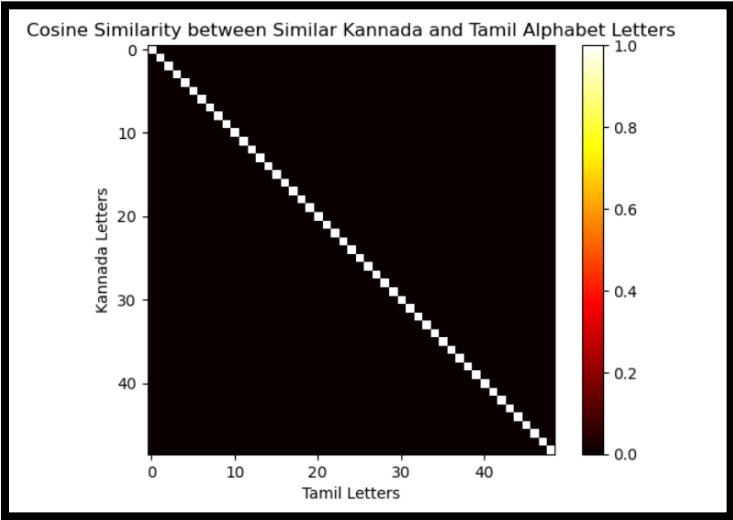
ii. Tamil and Kanada

```
# Define the alphabets of Kannada and Tamil Languages
kannada_alphabets = ['ಅ', 'ಆ', 'ಇ', 'ಈ', 'ಉ', 'ಊ', 'ಋ', 'ಎ', 'ಐ', 'ಒ', 'ಓ', 'ಔ', 'ಕ', 'ಖ', 'ಗ', 'ಘ', 'ಙ', 'ಚ', 'ಛ', 'ಜ', 'ಝ', 'ಞ', 'ಟ', 'ಠ', 'ಡ', 'ಢ', 'ತ', 'ಥ', 'ದ', 'ಧ', 'ನ', 'ಪ']
tamil_alphabets = ['அ', 'ஆ', 'இ', 'ஈ', 'உ', 'ஊ', 'எ', 'ஏ', 'ஐ', 'ஓ', 'ஔ', 'க', 'கா', 'ச', 'சா', 'ட', 'டா', 'ண', 'த', 'ந', 'ன', 'ப']

# Check and align the Lengths of Kannada and Tamil alphabets Lists
if len(kannada_alphabets) != len(tamil_alphabets):
    max_length = max(len(kannada_alphabets), len(tamil_alphabets))
    kannada_alphabets.extend([''] * (max_length - len(kannada_alphabets)))
    tamil_alphabets.extend([''] * (max_length - len(tamil_alphabets)))

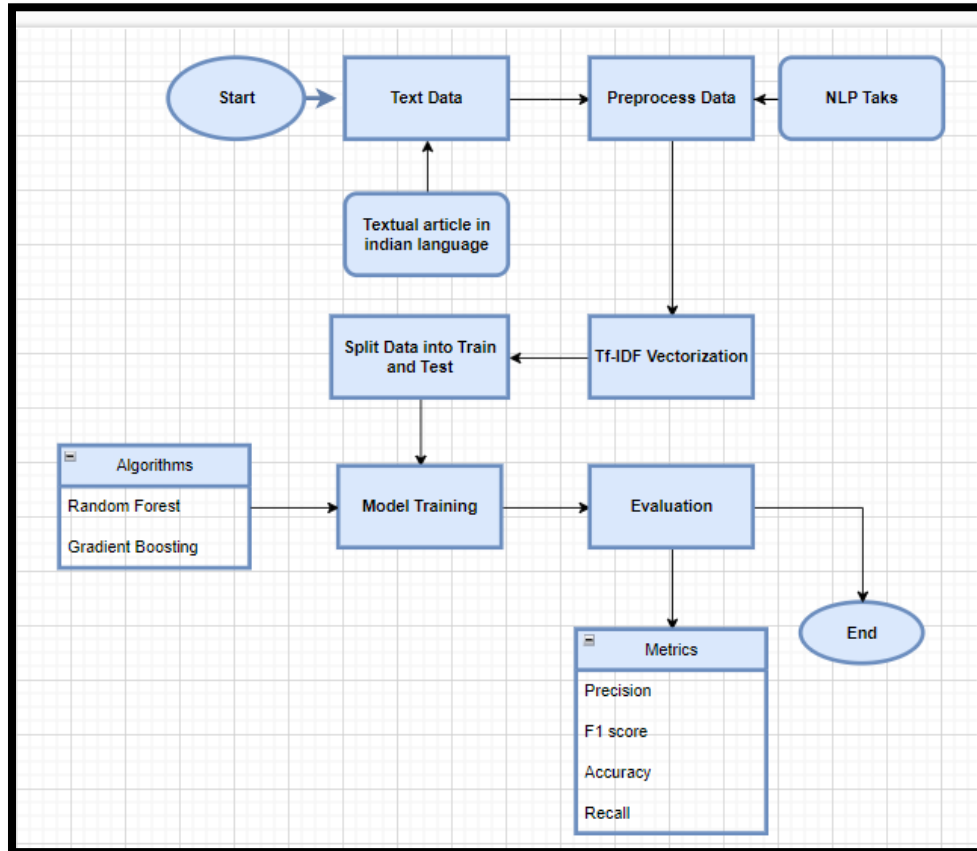
# Construct feature vectors for similar pairs of alphabets
similar_pairs = list(zip(kannada_alphabets, tamil_alphabets))
kannada_vectors = np.array([np.eye(1, len(kannada_alphabets), kannada_alphabets.index(pair[0])) for pair in similar_pairs])
tamil_vectors = np.array([np.eye(1, len(tamil_alphabets), tamil_alphabets.index(pair[1])) for pair in similar_pairs])
```

Mean Similarity: 0.02040816326530612  
Median Similarity: 0.0  
Maximum Similarity: 1.0  
Minimum Similarity: 0.0



# Polyglot Comparative Linguistics Analysis

## 1. Text Classification



### i. Hindi, Gujarati and Marathi

#### - Random Forest

Accuracy: 0.9324904105696832

Classification Report:

	precision	recall	f1-score	support
gujarati	1.00	0.97	0.99	12226
hindi	0.88	0.93	0.90	11898
marathi	0.92	0.89	0.91	11071

#### - Gradient Boosting

Accuracy: 0.8552067054979401

Classification Report:

	precision	recall	f1-score	support
gujarati	0.76	1.00	0.86	12226
hindi	0.92	0.80	0.86	11898
marathi	0.96	0.75	0.84	11071

## ii. Tamil and Kanada

### - Random Forest

Accuracy: 0.9906355123193018

Classification Report:

	precision	recall	f1-score	support
kanada	0.98	1.00	0.99	10941
tamil	1.00	0.98	0.99	11057

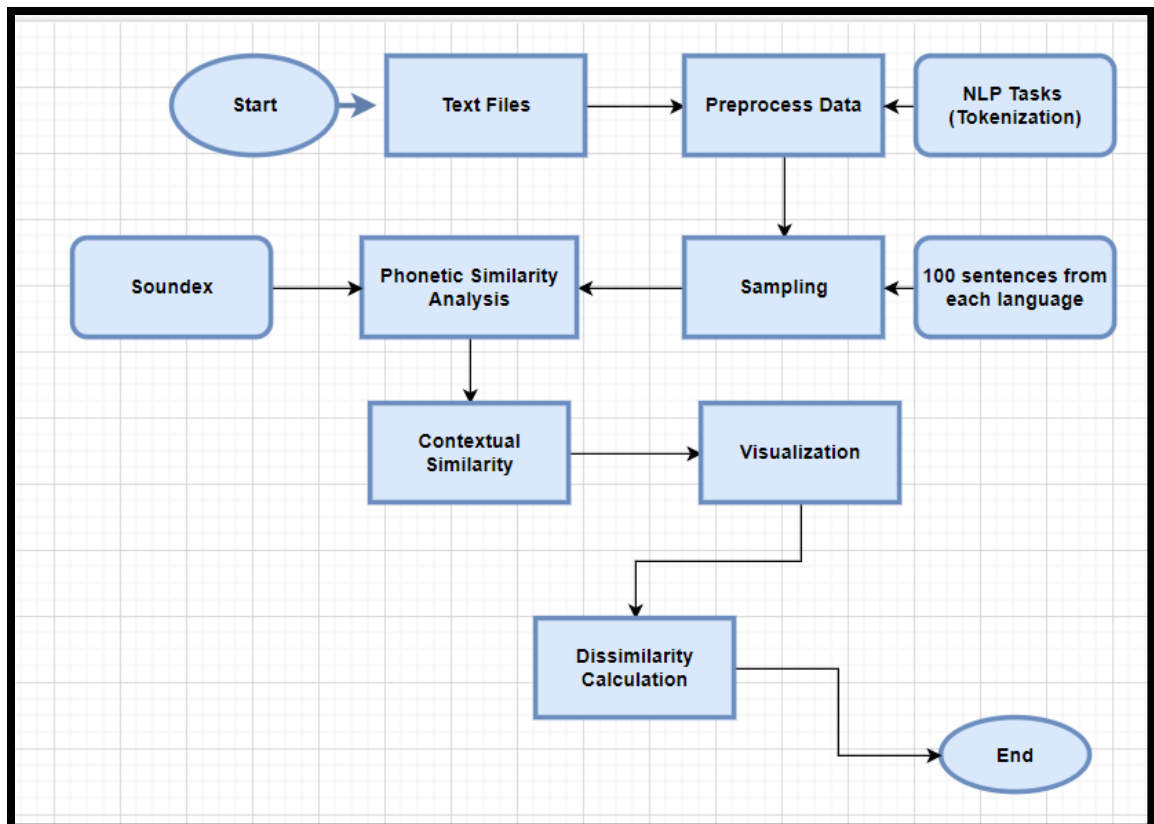
### - Gradient Boosting

Accuracy: 0.9396308755341395

Classification Report:

	precision	recall	f1-score	support
kanada	0.89	1.00	0.94	10941
tamil	1.00	0.88	0.94	11057

## 2. Textual Similarity

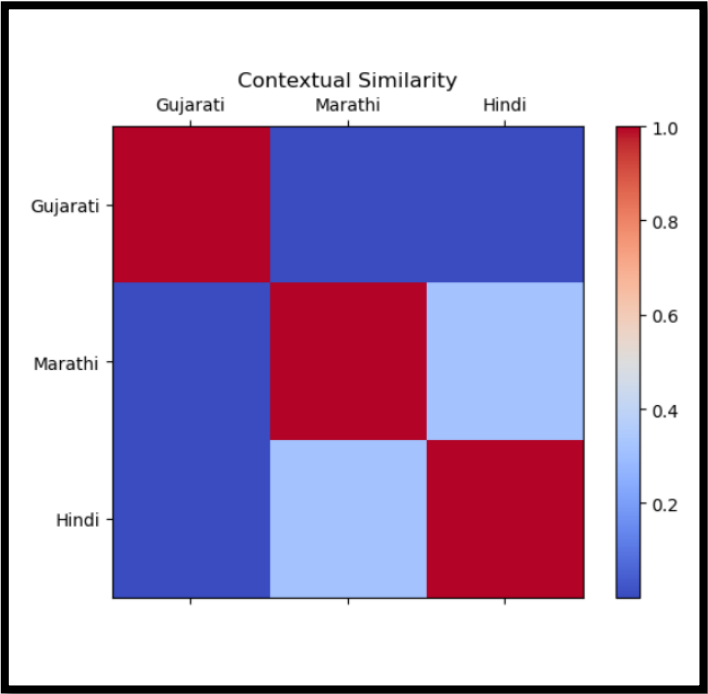
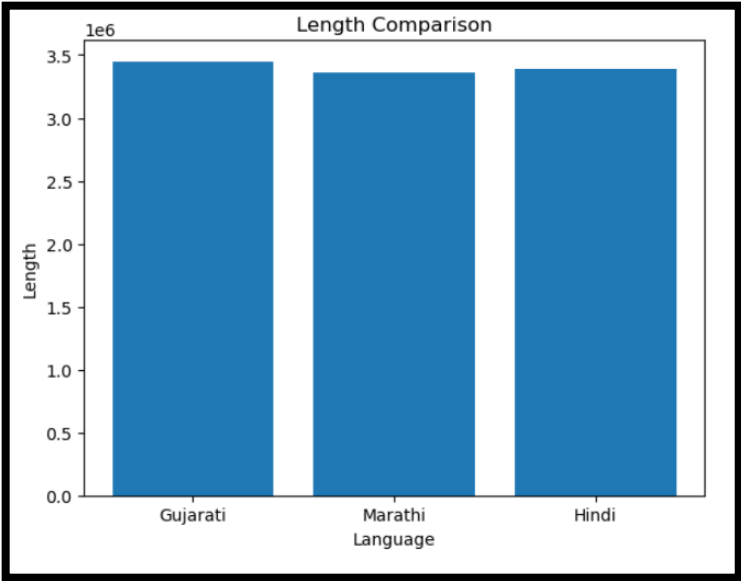
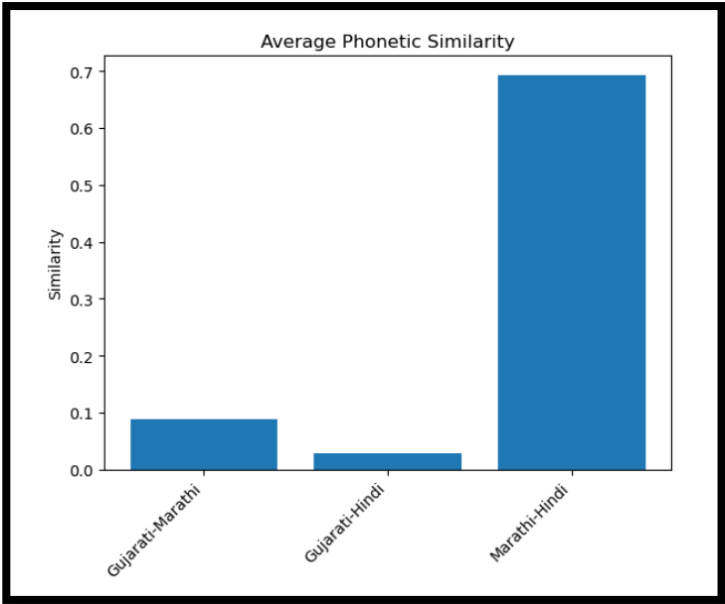


## i. Hindi, Gujarati and Marathi

Average phonetic similarity between Gujarati-Marathi: 0.09

Average phonetic similarity between Gujarati-Hindi: 0.03

Average phonetic similarity between Marathi-Hindi: 0.69  
Length of language article: {'Gujarati': 3445837, 'Marathi': 3361362, 'Hindi': 3384385}  
Dissimilarities below threshold:  
Gujarati-Marathi: 0.00  
Gujarati-Hindi: 0.00  
Marathi-Hindi: 0.32





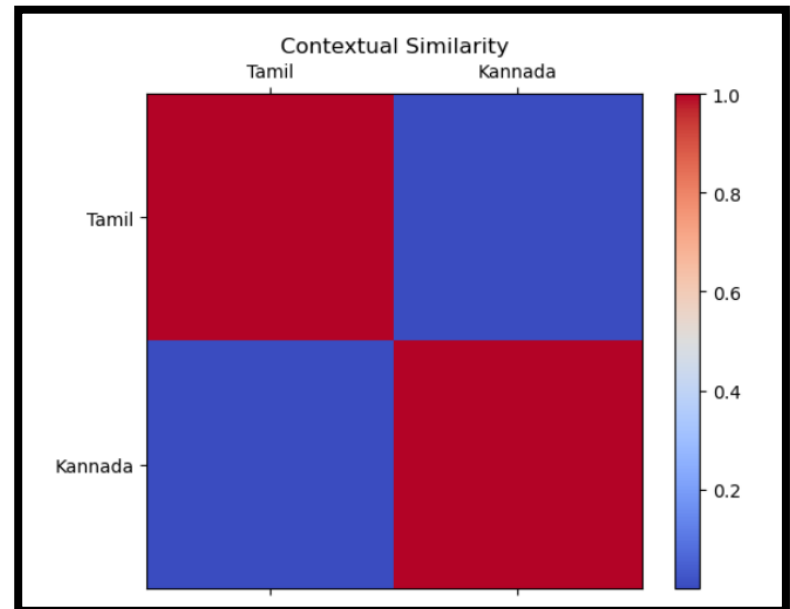
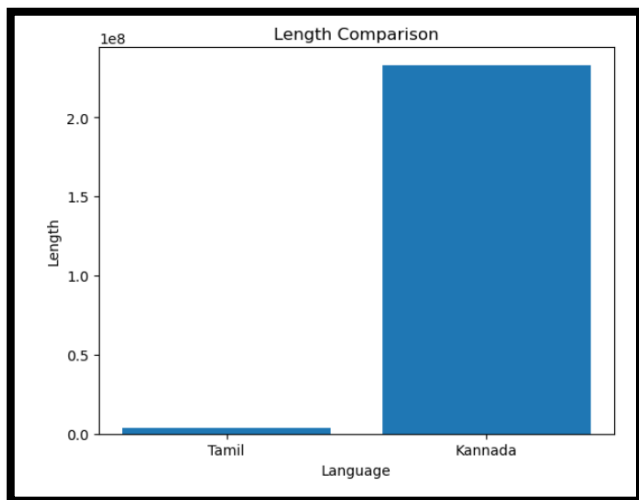
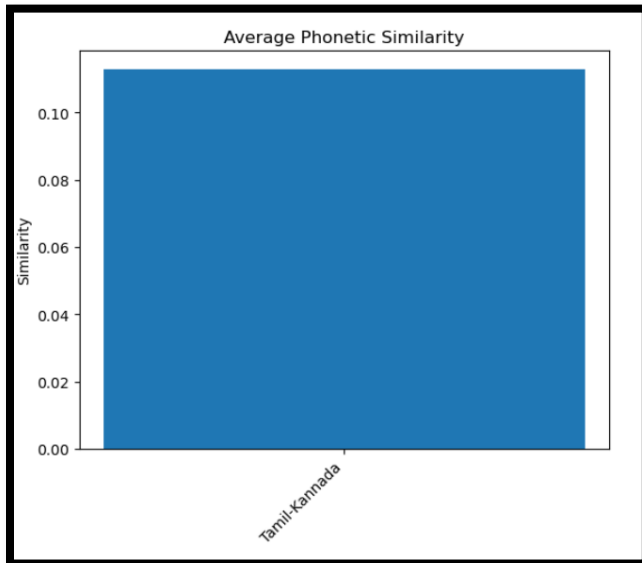
ii. Tamil and Kanada

Average phonetic similarity between Tamil-Kannada: 0.11

Length and language Article {'Tamil': 3941467, 'Kannada': 232699244}

Dissimilarities below threshold:

Tamil-Kannada: 0.00



iii. All 5 languages

Average phonetic similarity between Gujarati-Marathi: 0.09

Average phonetic similarity between Gujarati-Hindi: 0.03

Average phonetic similarity between Gujarati-Tamil: 0.09

Average phonetic similarity between Gujarati-Kannada: 0.09

Average phonetic similarity between Marathi-Hindi: 0.66

Average phonetic similarity between Marathi-Tamil: 0.10

Average phonetic similarity between Marathi-Kannada: 0.10

Average phonetic similarity between Hindi-Tamil: 0.01

Average phonetic similarity between Hindi-Kannada: 0.01

Average phonetic similarity between Tamil-Kannada: 0.11

{'Gujarati': 3445837, 'Marathi': 3361362, 'Hindi': 3384385, 'Tamil': 3941467, 'Kannada': 3548271}

Dissimilarities below threshold:

Gujarati-Marathi: 0.00

Gujarati-Hindi: 0.00

Gujarati-Tamil: 0.00

Gujarati-Kannada: 0.00

Marathi-Hindi: 0.33

Marathi-Tamil: 0.00

Marathi-Kannada: 0.00

Hindi-Tamil: 0.00

Hindi-Kannada: 0.00

Tamil-Kannada: 0.00

