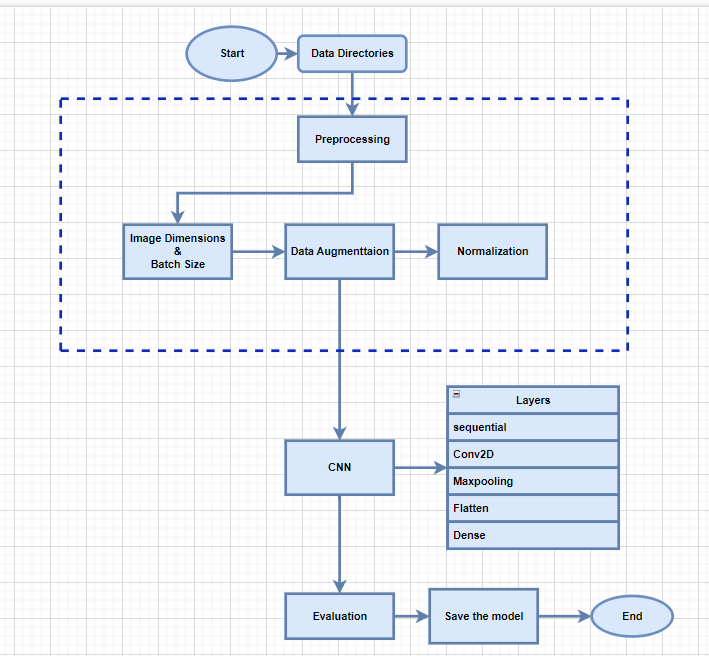
# Cross Language Phonetic Similarity

1. **Image Classification**

****

**Epochs:5**

**Oversampling in Dravadian**

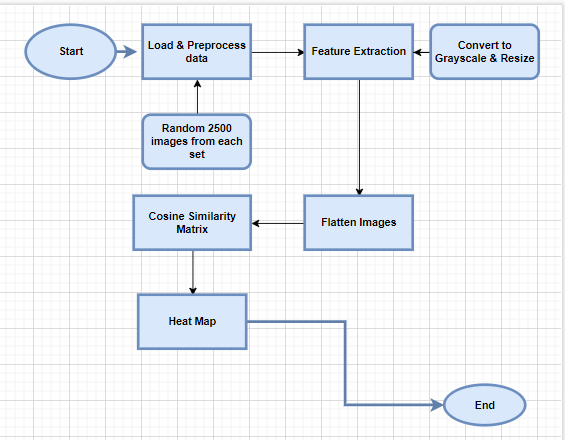
1. Gujarati and Marathi/Hindi (Devnagri Script)

Result: overall classification accuracy obtained: 0.9857

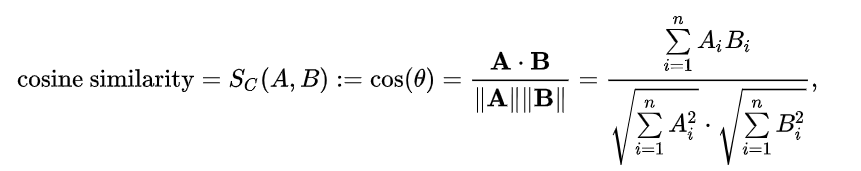
1. Tamil and Kanada (Dravidian Script)

Result: overall classification accuracy obtained: 0.9649

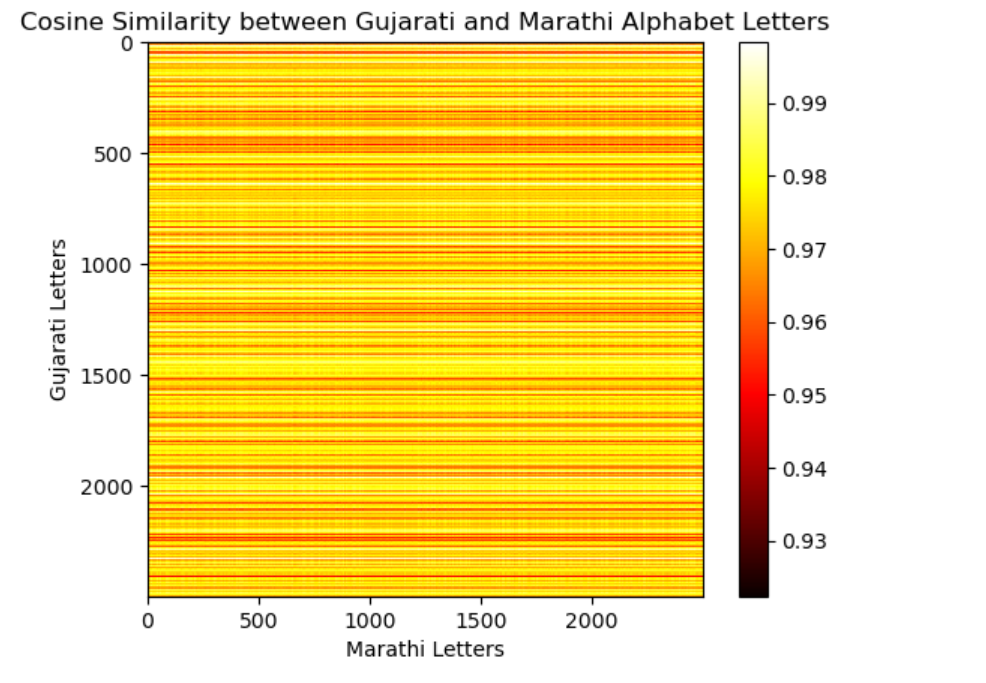
1. **Similarity Trend**
2. **Overall Random Trend**

****

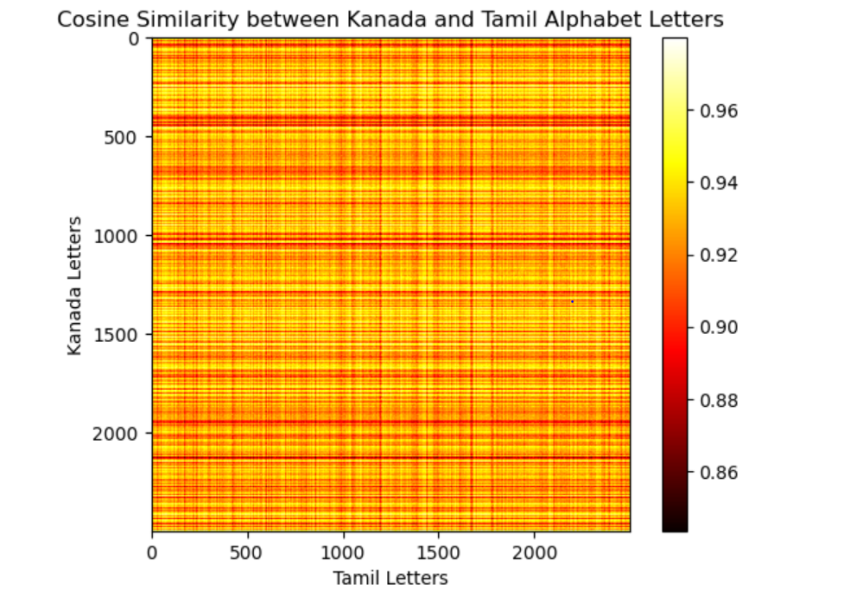
Cosine Similarity :



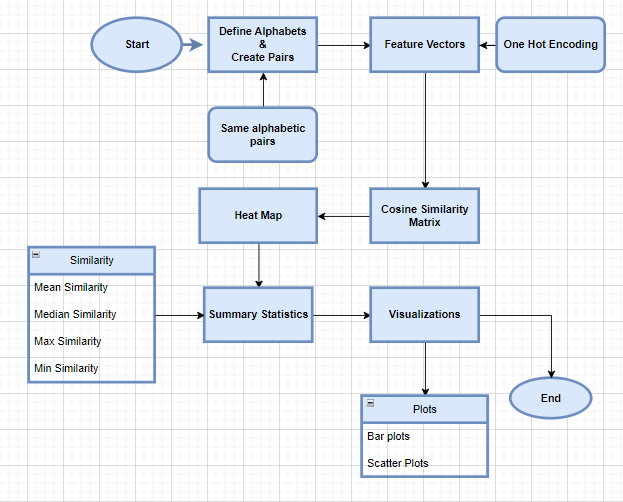
1. Gujarati and Marathi/Hindi



1. Tamil and Kanada

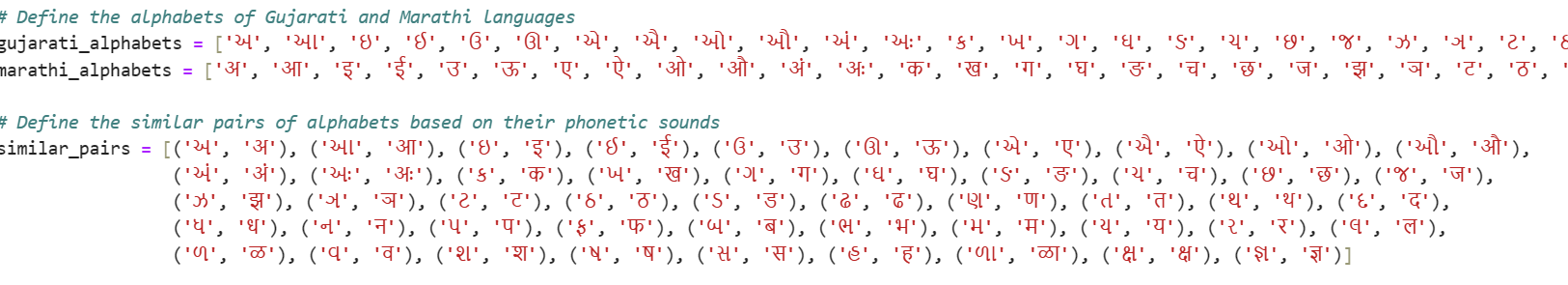


1. **Letter Wise Similarity**

****

Letter wise similarity code snippets and results.

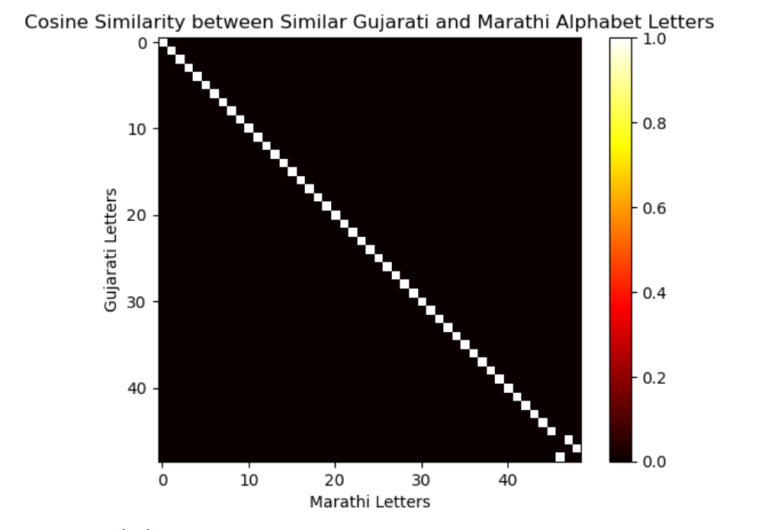
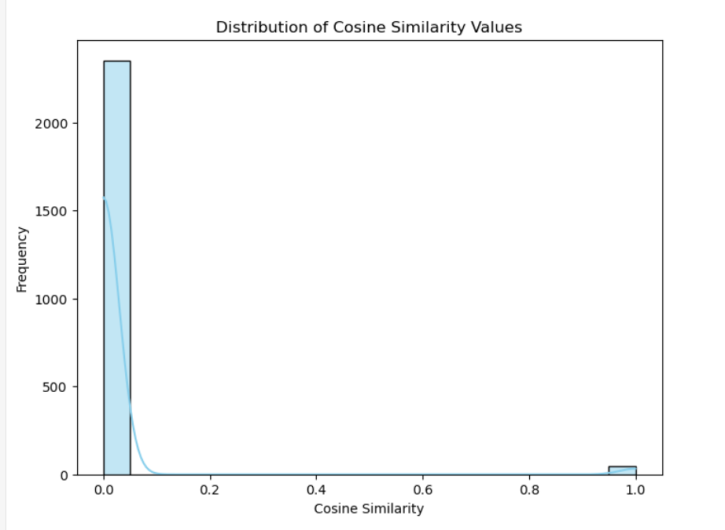
1. Gujarati and Marathi/Hindi

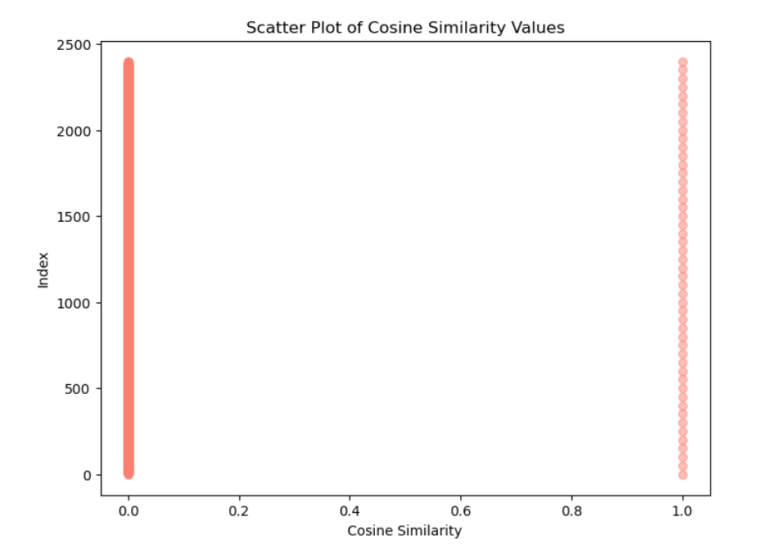
Mean Similarity: 0.02040816326530612

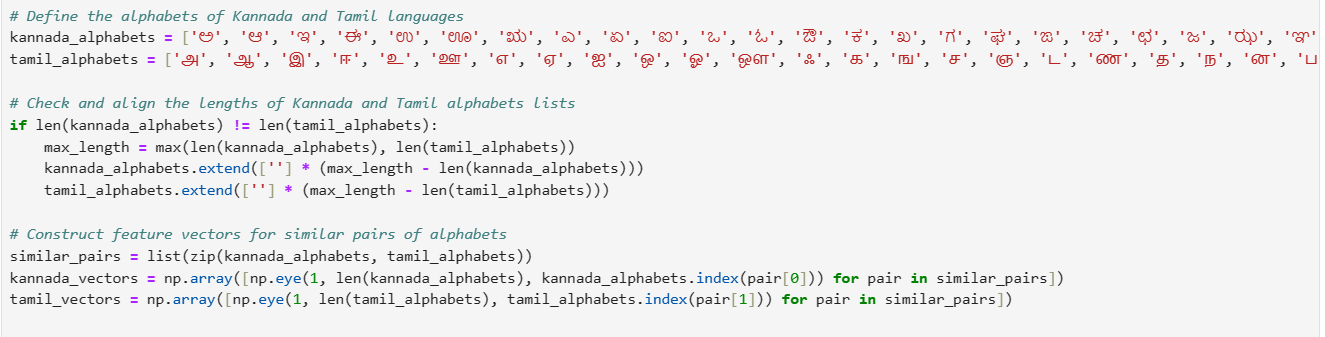
Median Similarity: 0.0

Maximum Similarity: 1.0

Minimum Similarity: 0.0





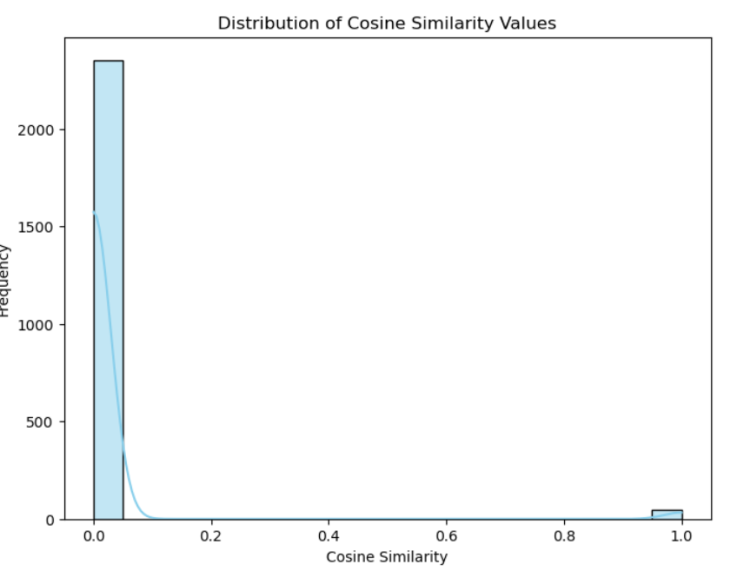
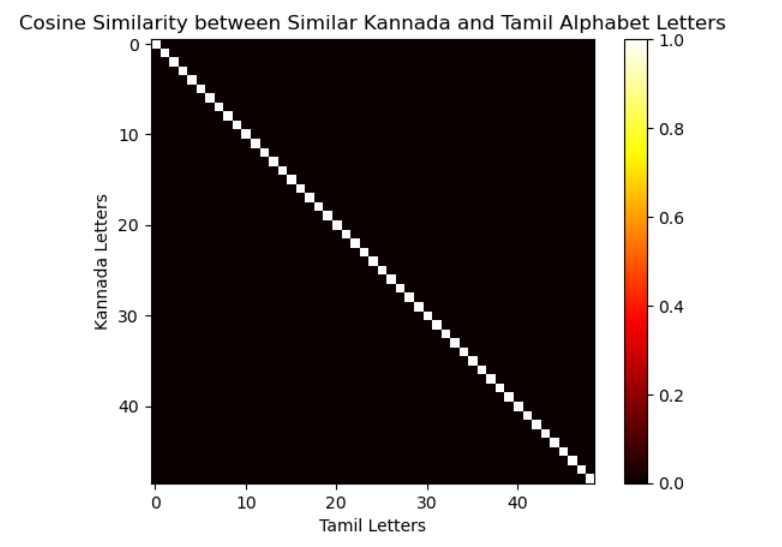
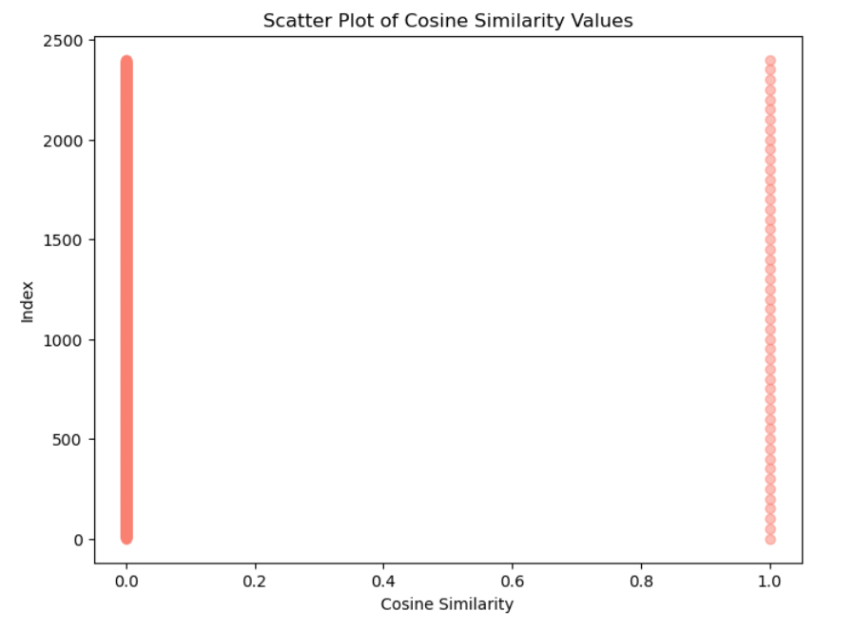
1. Tamil and Kanada

Mean Similarity: 0.02040816326530612

Median Similarity: 0.0

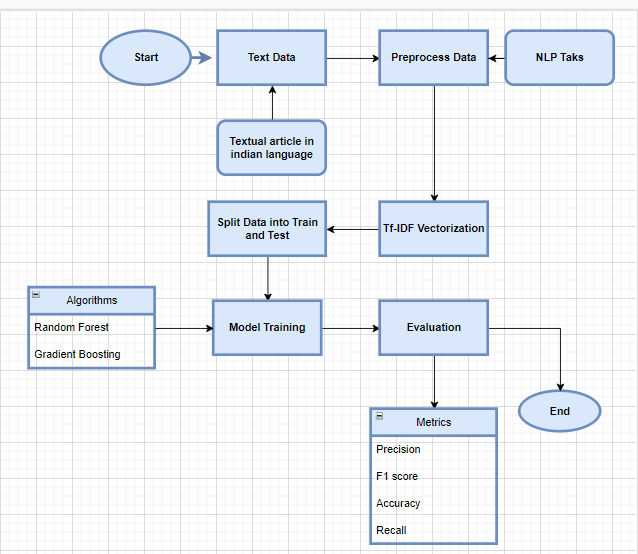
Maximum Similarity: 1.0

Minimum Similarity: 0.0



# Polyglot Comparative Linguistics Analysis

1. **Text Classification**

****

1. **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

1. **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

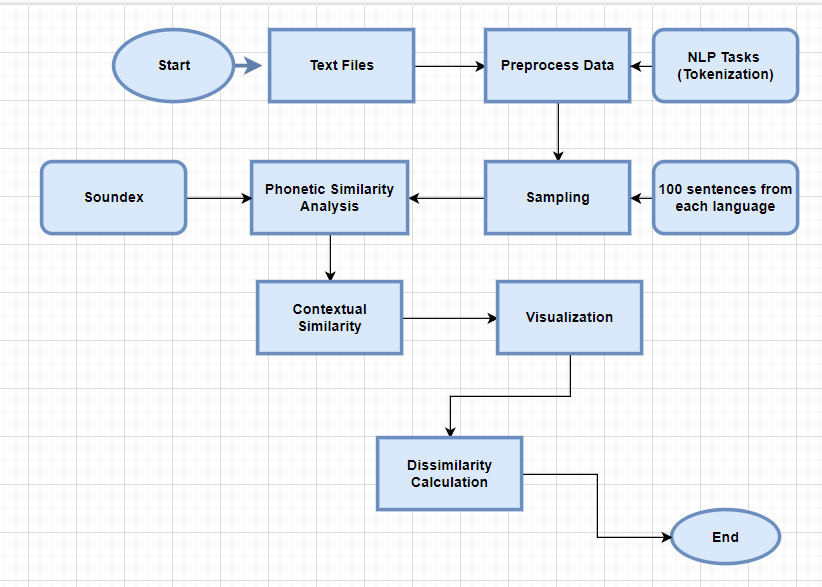
Classification Report:

precision recall f1-score support

kanada 0.89 1.00 0.94 10941

tamil 1.00 0.88 0.94 11057

1. **Textual Similarity**



1. 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

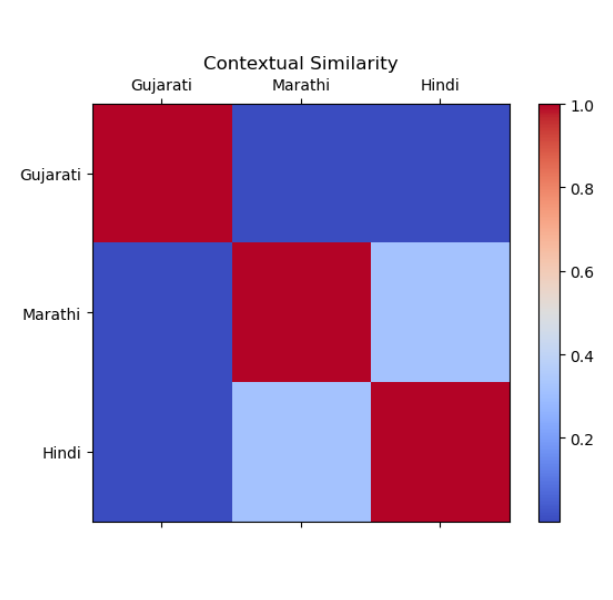
Length of language article: {'Gujarati': 3445837, 'Marathi': 3361362, 'Hindi': 3384385}

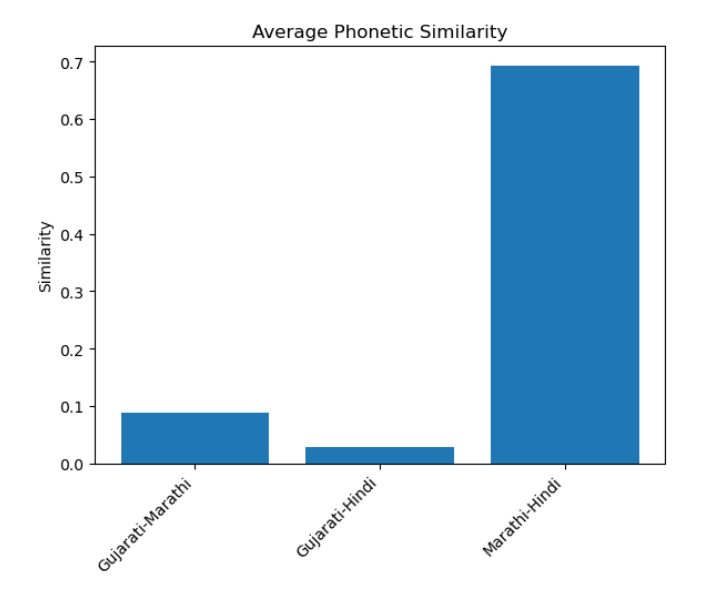
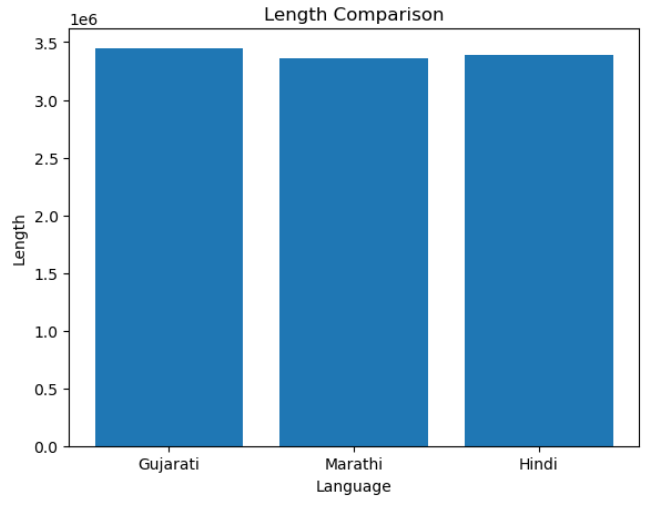
Dissimilarities below threshold:

Gujarati-Marathi: 0.00

Gujarati-Hindi: 0.00

Marathi-Hindi: 0.32





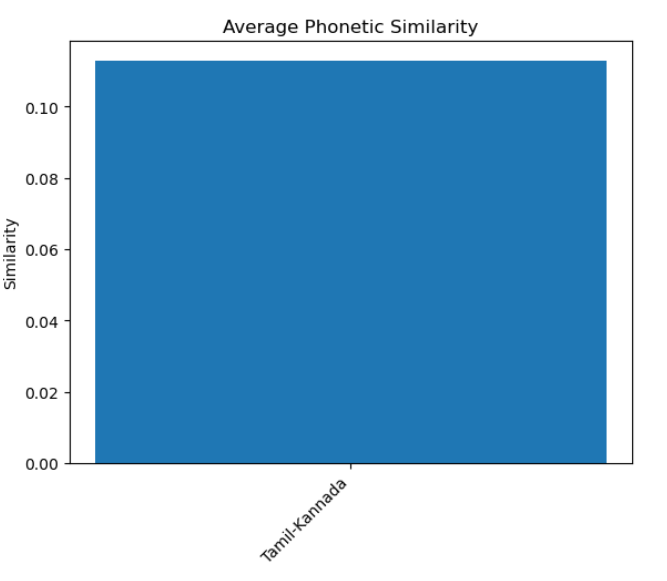
1. Tamil and Kanada

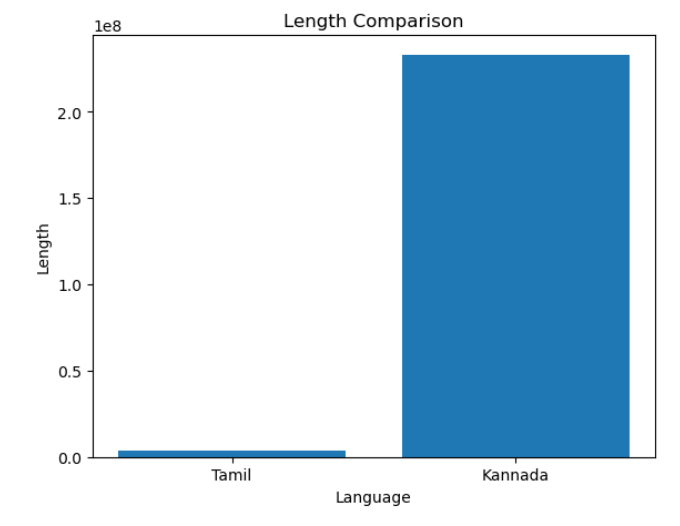
Average phonetic similarity between Tamil-Kannada: 0.11

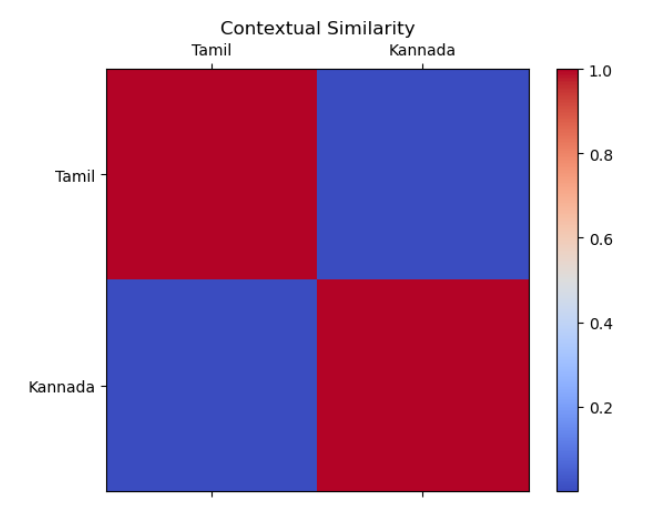
Length and language Article {'Tamil': 3941467, 'Kannada': 232699244}

Dissimilarities below threshold:

Tamil-Kannada: 0.00







1. 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

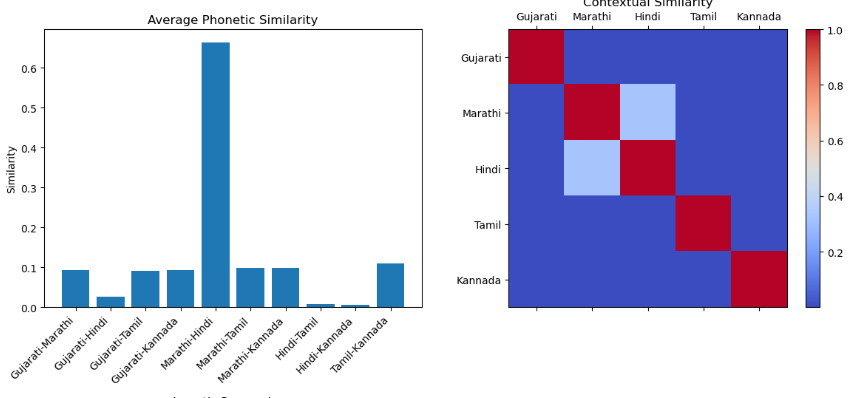
Marathi-Hindi: 0.33

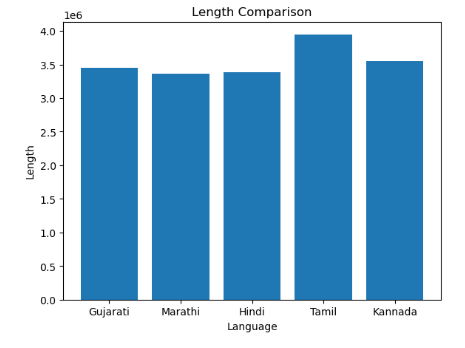
Marathi-Tamil: 0.00

Marathi-Kannada: 0.00

Hindi-Tamil: 0.00

Hindi-Kannada: 0.00

Tamil-Kannada: 0.00



**Significance of the Study**

1. Enhanced Multilingual Text and Image Processing: - The achievement of high classification accuracy in distinguishing between similar-looking characters from different languages, such as Gujarati and Marathi/Hindi, underscores the robustness of the models employed. This is pivotal for the development of more accurate and reliable Optical Character Recognition (OCR) systems tailored for multilingual environments.

2. Phonetic Similarity Analysis: - The meticulous analysis of letter-wise phonetic similarity elucidates the degree of resemblance between characters across different languages. This insight is instrumental in the advancement of more efficient transliteration and translation systems, thereby enhancing the accuracy and efficacy of cross-language communication tools.

3. Advancements in Machine Learning Models: - The application of sophisticated machine learning models, such as Random Forest and Gradient Boosting, for text classification across languages, demonstrates their efficacy in handling complex linguistic data. The high accuracy rates achieved indicate the potential for these models to be effectively utilized in similar tasks within other multilingual contexts.

**Future Applications**

1. Development of Multilingual OCR Systems: - The study's findings can inform the development of advanced OCR systems capable of accurately recognizing and processing text from multiple languages with similar scripts, thereby enhancing digital text processing and archiving capabilities.

2. Cross-Language Information Retrieval: - A deeper understanding of phonetic similarities can facilitate the creation of more effective information retrieval systems, capable of searching and retrieving relevant data across languages with analogous phonetic structures.

3. Improvement of Translation and Transliteration Tools: - Insights derived from letter-wise similarity and phonetic analysis can be integrated into translation and transliteration tools, leading to more accurate and contextually appropriate translations.

4. Advancement in Language Learning Applications: - These findings can be applied to the development of improved language learning applications that leverage phonetic similarities between languages, aiding learners in the acquisition of new languages more effectively.

5. Enhancement of Speech Recognition Systems: - The analysis of phonetic similarities has the potential to enhance the performance of speech recognition systems in multilingual contexts, making them more adaptable and accurate.

6. Contribution to Cultural and Linguistic Research: - This study provides a quantitative basis for understanding the relationships between different languages, which is valuable for linguistic and cultural research. This can aid in the preservation and study of linguistic heritage.

**Importance of the Findings**

1. Technological Advancements: - The high classification accuracy achieved signifies significant advancements in machine learning techniques applied to multilingual text and image classification, setting a new benchmark for future research and development in this field.

2. Linguistic Insights: - Understanding phonetic similarities between languages offers deeper insights into linguistic structures and relationships, which is crucial for both theoretical and applied linguistics.

3. Broad Practical Applications: - The practical implications of these findings are extensive, ranging from the improvement of everyday tools such as OCR and translation services to more specialized applications in linguistic research and education.

4. Interdisciplinary Impact: - This study bridges the gap between computational linguistics and machine learning, demonstrating the potential for interdisciplinary research to address complex problems involving multiple languages.

In conclusion, this study not only propels the field of multilingual text and image processing forward but also opens numerous avenues for future research and application. Its importance lies in both the technological advancements it brings and the deeper understanding of linguistic relationships it provides.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Study** | **Techniques** | **Application** | **Future Work** | **Languages** | **Dataset** |
| **[1]** | **Transformer-based embedding models** | **Recognition of hate speech in Dravidian languages** | **Understanding misclassification** | **Kanada Malyalam, Tamil** | **Dravidian CodeMix: Kannada-English, Malayalam-English, Tamil-English. From offensive language identification tasks.** |
| **[2]** | **Wikipedia text corpora, Unified multilingual model** | **Uniform ASR for South Indian languages** | **Exploration of language coverage and model effectiveness** | **Kannada, Telugu, Sanskrit, Malayalam, Tulu** | **Speech data: Telugu, Kannada, Malayalam, Sanskrit. Includes training, development, testing subsets. Sanskrit has OOD subset.** |
| **[3]** | **SVM, Random Forest, m-BERT** | **Hate speech detection in Indian languages** | **Improvement of classifier performance** | **Hindi, Dravadian** | **Organizers' datasets: HASOC, TRAC, Dravidianlangtech. Fine-grained labels available for detailed hate speech info.** |
| **[4]** | **mT5 transformer, Asian Language Treebank database** | **Multilingual Neural Machine Translation System** | **Tackling translation difficulties** | **Hindi, Bengali, English, Bhojpuri, Sindhi, Magadhi, Marathi, URDU** | **Modified Asian Language Treebank dataset (Hindi corpus).** |
| **[5]** | **LSTM, attention techniques** | **Inter-language translation for Indian languages** | **Further research in machine translation** | **Hindi, Odia, Malayalam, Indian Sign Language** | **Review paper.** |
| **[6]** | **Markov Chain, n-gram models** | **Analysis of Indus Valley civilization's writing system** | **Understanding ancient communication techniques** | **Indus Script** | **Texts from ICIT database in CSV files. Indus Corpus.** |
| **[7]** | **Weiner filter, HOG, ANN** | **Handwritten digit recognition in South Indian languages** | **Integration of ANN and HOG for accuracy** | **Malayalam, Kannada, Telugu, Devanagiri, Hindi** | **Handwritten digit images: Malayalam, Kannada, Telugu, Devanagiri, Hindi. Binary images (130x66 pixels) with various writing styles and numeral sizes.** |
| **[8]** | **NMT, RBMT, SMT** | **Machine translation for Indian languages** | **Refinement of translation techniques** | **Hindi, Punjabi, Bengali, Kannad, Marathi, Telugu** | **Survey paper.** |
| **[9]** | **Language-independent transcribers, RNN, SVM** | **Automatic Spoken Language Identification (LID)** | **Improvement of LID accuracy** | **Language transcribers** | **Studio Record dataset. 8 transcribers for Indian languages.** |
| **[10]** | **Tokenization, semantic triplets, WordNet, TF-IDF** | **Linguistic semantic structure for Indian languages** | **Further refinement of semantic graphs** | **English, European, Indian** | **NA.** |
| **[11]** | **Lexicon-based analysis, machine learning** | **Sentiment analysis in code-mixed text** | **Development of sentiment analysis resources** | **Code-mixed languages** | **Review paper. English-Hindi, English-Bengali, English-Tamil, English-Urdu.** |
| **[12]** | **Statistical and linguistic approaches, topic modeling** | **Text summarization for Indian languages** | **Improvement of summarization effectiveness** | **Hindi, Gujarati, Marathi, Tamil, Malayalam, Odia, Punjabi, Assamese, Tamil** | **Review paper.** |
| **[13]** | **Hybrid CTC/attention, Multilingual Transformer** | **Speech recognition for low-resource Indian languages** | **Enhancement of speech recognition accuracy** | **Gujarati, Tamil, Telugu** | **Microsoft, SpeechOcean.com challenge: Wave files + UTF-8 transcriptions.** |
| **[14]** | **CTC loss, CNN, TDNN** | **Multilingual and code-switching ASR** | **Integration of LM and noise reduction for ASR improvements** | **Gujarati, Hindi, Odia, Marathi, Telugu, Tamil, Bengali** | **Microsoft Research Open Data, Hindi stories, local dialects.** |
| **[15]** | **Hidden Markov Model (HMM)** | **Part-of-Speech tagging for Indian languages** | **Utilization of larger corpus and tagsets** | **Marathi, Hindi, Assamese, Bengali, Telugu, Konkani, Manipuri** | **Katkari dataset.** |
| **[16]** | **Deep neural networks (Recursive Auto Encoder, CNN, LSTM, Bidirectional GRU)** | **Automated creation and analysis of natural language** | **Integration of deep learning into Indian languages** | **Hindi, Malayalam, Punjabi, Assamese** | **ILCI Phase II Malayalam dataset.** |
| **[17]** | **LSTM, Encoder-Decoder model** | **Machine translation between Indian languages** | **Improvement of translation accuracy** | **Marathi to Gujarati** | **Indicnlp.ai4bharat.org/samanantar/.** |
| **[18]** | **Multilingual TTS systems** | **Multilingual voice training for Indian languages** | **Designing Indic speech synthesis systems** | **Indo-Aryan, Dravidian** | **Indic TTS database. Bengali, Gujarati, Hindi, Odia, Rajasthani (Devanagari script), Kannada, Malayalam, Tamil, Telugu.** |