IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

IJARSCT

Impact Factor: 7.301

Volume 3, Issue 1, December 2023

Text Language Identification and Translator

Tejas Pinge[1], Prajwal Patil[2] , Mayur Sherki[3], Aditya Nandurkar[4] , Prof. Ravindra Chilbule[5]

Students, Department of Computer Science Engineering[1,2,3,4]

Guide, Department of Computer Science Engineering[5]

Rajiv Gandhi College of Engineering Research and Technology, Chandrapur, Maharashtra, India

tejaspinge232@gmail.com, prajwalpatil2812@gmail.com, sherkimayur@gmail.com, nandurkaraditya8@gmail.com

Abstract: *Language Identification refers to the process of detecting the language(s) of the text in the document based on the script used for writing and observing the diacritics particular to a language. This research area has always fascinated researchers as early as 1970 and till now due to varied applications and increased demands of this field. In this work, I address the problem of detecting language of textual documents. I have introduced a method which is able to detect language of text more efficiently and accurately by determining their respective proportions and finding the greatest of them which represents the language of the text. I have demonstrated the performance comparison of three different approaches which are using n-gram approach (word-wise), using n-gram approach (character-wise) and using a combination of word search and stop words detection. My project currently contains language models for 4 languages. On an average the accuracy of my program is about 96.5%.*

Keywords: Language

I. INTRODUCTION

Language, a system of conventional spoken, manual, or written symbols by means of which human beings, as members of a social group and participants in its culture, express themselves. The functions of language include communication, the expression of identity, play, imaginative expression, and emotional release. Every physiologically and mentally typical person acquires in childhood the ability to make use of a system of communication that comprises a circumscribed set of symbols (e.g., sounds, gestures, or written or typed characters) in both way i.e., as a sender or receiver. By means of these symbols, people are able to impart information, to express feelings and emotions, to influence the activities of others, and to comport themselves with varying degrees of friendliness or hostility toward persons who make use of substantially the same set of symbols. Different systems of communication constitute different languages; the degree of difference needed to establish a different language cannot be stated exactly. Generally, systems of communication are recognized as different languages if they cannot be understood without specific learning by both parties, though the precise limits of mutual intelligibility are hard to draw and belong on a scale rather than on either side of a definite dividing line. So, here comes the role of Language Translation which implies the role and importance of Language Identification. Language Identification typically acts as a pre-processing stage for both human listeners (i.e. call routing to a proper human operator) and machine systems (i.e. multilingual speech processing systems). Language Identification or Language Guessing is studied under Natural Language Processing and is the problem of determining which natural language given content is written in. Computational approaches to this problem view it as a special case of text categorization, solved with various statistical methods.

II. LITERATURE REVIEW

A comprehensive literature review in text language identification should include a diverse range of scholarly articles, research papers, and relevant publications from reputable sources to provide a thorough understanding of the field's current state and future directions.

DOI: 10.48175/IJARSCT-14055

III. PROPOSED METHODS

Statistical Methods:

- **N-gram Language Models:** Utilizing the frequencies of character or word sequences (n-grams) to identify the most probable language based on statistical patterns.
- **Language Profiles:** Creating language profiles or language models based on the frequency distribution of characters, words, or features specific to each language.

Machine Learning Methods:

- **Supervised Learning:** Using labeled datasets to train classifiers (e.g., Support Vector Machines, Naive Bayes, Decision Trees) on text samples from different languages to classify new text.
- **Unsupervised Learning:** Clustering techniques or unsupervised models (e.g., K-means clustering, Gaussian Mixture Models) without labeled data to group similar languages based on text features.

Deep Learning Approaches:

- **Recurrent Neural Networks (RNNs):** Employing RNNs or Long Short-Term Memory (LSTM) networks to capture sequential patterns in text for language identification.
- **Convolutional Neural Networks (CNNs):** Using CNN architectures to extract hierarchical features from text data and identify language based on learned representations.
- **Transformer Models:** Leveraging Transformer-based architectures (e.g., BERT, GPT) that use self-attention mechanisms for contextual understanding and language classification.

Hybrid Approaches:

- **Combining Statistical and Machine Learning Techniques:** Using a combination of statistical models and machine learning algorithms to enhance accuracy and robustness.
- **Ensemble Methods:** Integrating multiple classifiers or models to make collective decisions, leading to improved language identification performance.

Feature Extraction Techniques:

- **Character-Level Features:** Analyzing character n-grams, character frequencies, or statistical properties of characters in the text.
- **Word-Level Features:** Utilizing word n-grams, vocabulary richness, or word length distributions to distinguish between languages.
- **Language-agnostic Features:** Extracting features that are independent of language, such as entropy, punctuation usage, or structural properties of the text.

Lexical and Morphological Analysis:

- **Lexicon-Based Methods:** Employing dictionaries, word lists, or linguistic rules specific to each language for identification purposes.
- **Morphological Analysis:** Considering linguistic features like word morphology, inflections, or grammatical structures characteristic of different languages.

Domain-Specific Techniques:

- **Social Media Text Processing:** Adapting language identification methods to handle informal language, slang, emojis, or code-mixed text prevalent in social media platforms.
- **Multilingual Text Processing:** Developing models capable of handling multilingual environments or code-switching scenarios.

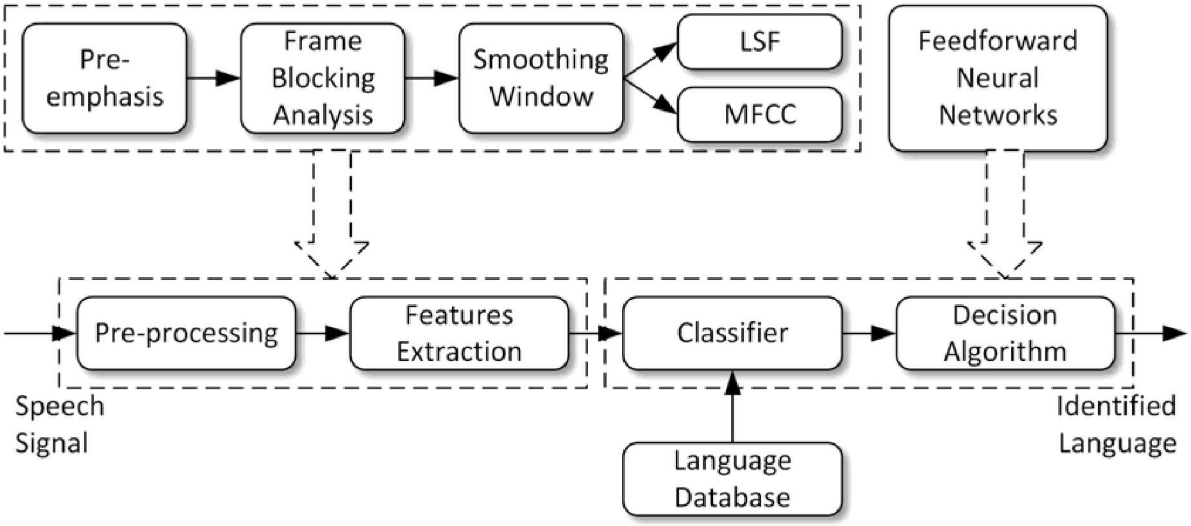
Transfer Learning and Adaptation:

- **Fine-tuning Pre-trained Models:** Leveraging pre-trained language models and adapting them to specific language identification tasks with fine-tuning or transfer learning approaches.

IV. DATASET DRISCRIPSION

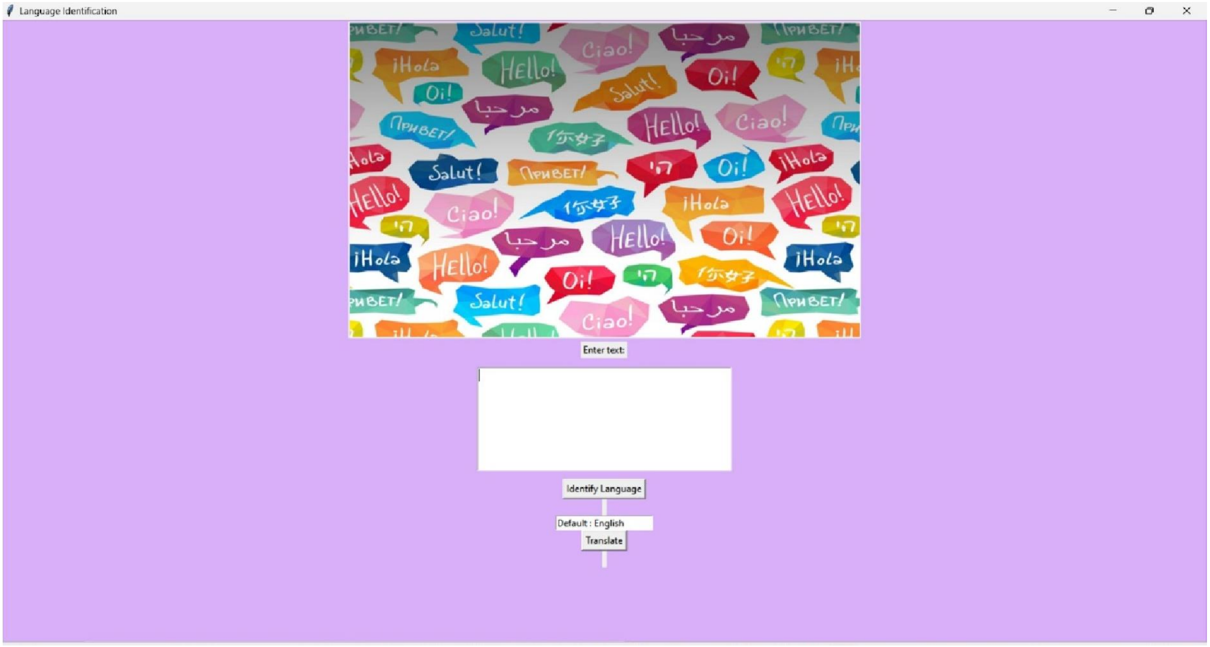
- **Tatoeba:** Tatoeba is a vast multilingual corpus containing sentences in multiple languages. It provides parallel text data in various languages, making it suitable for training and evaluating language identification models.
- **European Parliament Proceedings Parallel Corpus:** This corpus consists of parallel text from the proceedings of the European Parliament in multiple languages. It contains aligned sentences across languages, allowing for language identification tasks.
- **Wikipedia Dumps:** Wikipedia dumps in different languages offer a wealth of textual data. Researchers often use samples or subsets of Wikipedia articles in various languages to create language identification datasets.
- **UDHR (Universal Declaration of Human Rights):** The UDHR has been translated into hundres of languages. Extracting sentences or paragraphs from UDHR documents in different languages forms a diverse language identification dataset.
- **JRC-Acquis Multilingual Parallel Corpus:** This dataset contains legislative texts from the European Union in multiple languages. It provides aligned texts in different languages, suitable for language identification tasks.
- **CLTK (Classical Language Toolkit) Corpora:** For historical or ancient languages, CLTK provides corpora containing texts from languages like Latin, Ancient Greek, Sanskrit, etc., aiding in language identification research for these languages.
- **OSCAR (Open Super-large Crawled ALMAnaCH coRpus):** A large multilingual corpus derived from web pages, covering a wide range of languages. It provides diverse and unstructured text data for language identification tasks.
- **Panlex Database:** Panlex is a vast lexical translation database with translations between thousands of languages. Researchers leverage this resource to create multilingual datasets for language identification.
- **Twitter Multilingual Corpus:** Collections of tweets in various languages, often used to create datasets for social media-specific language identification tasks.
- **Common Crawl Corpus:** An extensive dataset collected by web crawlers, containing text data from a broad spectrum of websites in multiple languages.

V. DAIGRAM IMPLEMENTATION

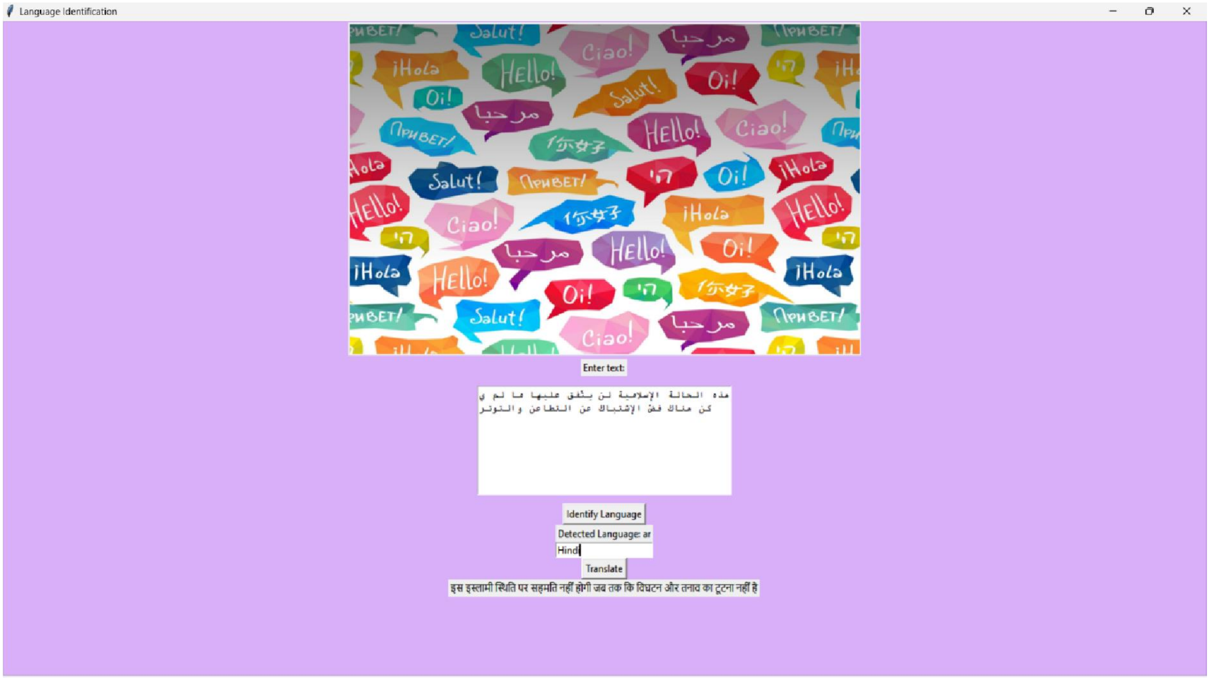


DOI: 10.48175/IJARSCT-14055

Output



Sample GUI



هذه الحالة الإسلامية لن يتفق عليها ما لم ي
كن هناك فض الإشتباك عن التطاعن و التوتر

Detected Language: ar

Hindi

इस इस्लामी स्थिति पर सहमति नहीं होगी जब तक कि विघटन और तनाव का टूटना नहीं है

GUI with output

VI. FUTURE WORK

- **Handling Low-Resource Languages:** Development of models capable of accurately identifying languages with limited data or resources, often termed as low-resource languages. Exploring transfer learning, unsupervised techniques, or data augmentation methods for improved performance in such scenarios.
- **Code-Mixing and Multilingual Text:** Addressing the challenges posed by code-switching and multilingual text environments prevalent in social media, chat platforms, or mixed-language contexts. Creating models capable of accurately identifying languages in code-mixed text or multilingual content.

- **Improving Robustness and Generalization:** Enhancing the robustness of language identification models against noisy, ambiguous, or misspelled text data. Developing models that generalize well across various text genres, domains, and dialectal variations.
- **Domain Adaptation and Cross-Domain Learning:** Investigating techniques for adapting language identification models from one domain to another, ensuring better performance in specific domains or applications without significant retraining.
- **Handling Short Texts and Noisy Data:** Developing models that can effectively identify languages in short texts or snippets commonly found in search queries, microblogs, or chat conversations. Exploring techniques to handle noisy or informal text data prevalent in social media.
- **Enhancing Multilingual Models and Representations:**
- Further advancements in multilingual embeddings, representations, or pre-trained language models (e.g., transformers) to improve language identification capabilities. Fine-tuning these models for language identification tasks across diverse languages.
- **Zero-shot and Few-shot Learning:** Researching zero-shot or few-shot learning techniques for language identification, enabling models to identify languages not encountered during training by leveraging similarities or transferable knowledge
- **Ethical Considerations and Bias Mitigation:** Addressing biases in language identification models related to underrepresented languages, dialects, or minority communities. Ensuring fairness and inclusivity in language identification systems.
- **Integration with Downstream Applications:** Integrating language identification seamlessly with various downstream applications such as machine translation, information retrieval, sentiment analysis, etc., to enhance their performance in multilingual scenarios.
- **Benchmarking and Evaluation Standards:** Establishing standardized benchmarks, evaluation metrics, and test suites for assessing the performance of language identification models across different languages, domains, and data types.

VII. CONCLUSION

Summarize key findings, challenges, advancements, and trends discussed in the literature review.Emphasize the significance of text language identification in today's multilingual and interconnected digital environment.

REFERENCES

[1]. Cavnar, William B., and John M. Trenkle. "N-gram-based text categorization." Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. 1994.

[2]. Dunning, Ted. "Accurate methods for the statistics of surprise and coincidence." Computational Linguistics 19.1 (1993): 61-74

[3]. Cortes, C., and Vapnik, V. "Support-vector networks." Machine Learning 20, 3 (1995): 273–297.

[4]. Baldwin, Timothy, and Su Nam Kim. "Multiword expressions." Language and Linguistics Compass 3.1 (2009): 870-894.

[5]. Lui, Marco, and Timothy Baldwin. "langid.py: An off-the-shelf language identification tool." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2012.

[6]. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[7]. Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the Association for Computational Linguistics 5 (2017): 135-146.

[8]. Devlin, Jacob, et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

DOI: 10.48175/IJARSCT-14055