WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Language Detector

Sunil Thale[1], Prof. S. S. Mandwale[2]

[1]U.G. Student, Department of Computer Science & Engineering, Shreeyash College of Engineering and Technology, Aurangabad, India
[2]Assistant Professor, Department of Computer Science & Engineering, Shreeyash College of Engineering and Technology, Aurangabad, India

ABSTRACT:

Language detection is a fundamental task in natural language processing (NLP) with wide-ranging applications, including text categorization, sentiment analysis, and content recommendation. This abstract provides a concise overview of language detection methods employing machine learning (ML) techniques' primary objective of language detection is to automatically identify the language of a given text. Traditional rule-based approaches often fall short in handling the complexity and diversity of languages present on the internet. In contrast, machine learning offers a promising avenue for effective language identification. For language detection include n-grams, character and word frequency distributions, and statistical measures. These features are fed into ML models such as Support Vector Machines (SVM), Naive Bayes, and neural networks for training. Additionally, ensemble methods and deep learning architectures, such as recurrent neural networks (RNNs) and transformers, have demonstrated superior performance in capturing intricate language patterns.

I. INTRODUCTION

You can clearly identify the languages you know as a human. For instance, I can recognise Hindi and English with ease, but I am not able to recognise all Indian languages because I am Indian. The language identification task can be applied in this situation. With so many users worldwide, Google Translate is one of the most widely used language translators available. In the event that you are unsure of the language you wish to translate, you can also utilise the machine learning model that is included to identify languages.

The first step in accomplishing a number of tasks, such as identifying the source language for machine translation, enhancing search relevancy by tailoring search results to the language of the query, offering a uniform search box for a multilingual dictionary, tagging data stream from Twitter with relevant language, etc., is automatic language detection. Although it is not difficult to classify languages into disjoint groups, it is still difficult to distinguish between dialects and languages that come from the same source in the field of natural language processing.

For such similar languages, standard word-frequency-only classifiers are unable to accurately predict speaker behaviour. Therefore, in order to improve classifier performance, state-of-the-art machine learning technologies must be used to capture the language's structure. In order to create a state-of-the-art language classifier, we used the latest developments in deep neural network-based models, which have demonstrated exceptional performance in numerous natural language processing applications. We used the DSL test dataset to assess our solution against the top performers in the market, and we came in first. +

II. METHODOLOGY

Choose a suitable machine learning algorithm. Common choices include: Naive Bayes: Simple and fast, works well for text classification tasks. Support Vector Machines (SVM): Effective for high-dimensional data like text. Neural Networks: Deep learning models, such as recurrent neural networks (RNNs) or transformers, can also be employed for more complex language models.

III. LITERATURE SURVEY

As a matter of fact, there are more than 15,000 spoken languages in India [1] of which only a handful of them are known to us As observed in literature, there exists a lot of studies on language detection and identification. Different languages like English, Italian, German, Dutch, Japanese and Indian languages like Hindi and Kannada are detected. Either text or handwritten [2]. To detect the handwritten languages techniques like Machine Learning and Deep Learning are used [3-6]. Deep Neural Networks (DNNs) are used for speech signals that automatically identify language at the acoustic frame levels. The designed DNNs architectures are compared with several state-of-the-art acoustic systems which are based on i-vectors, the results when tested against the two datasets i.e., NIST LRE 2009 and Google 5M LID, it was concluded that in most of the cases, the DNNs performed better than the current state-of-art approaches [3]. Class frequencies are used in a centroid-based classification method to determine the language. But, the success rate of centroid-based classification is generally lower when compared to the other methods. Hence, a new and different method which is known as Inverse Class

Frequency (ICF) was developed which increases the quality of centroid values by updating the classical values that provided better successful results and also has lower time complexities when compared with other methods

IV. SYSTEM DEVELOPMENT

• Data Collection:

Gather a diverse dataset containing text samples from various languages.

Ensure the dataset covers a wide range of linguistic characteristics and writing styles.

• Data Preprocessing:

Clean the text data by removing noise, special characters, and irrelevant formatting.

Normalize the text, addressing issues like capitalization, diacritics, and accents.

• Feature Extraction:

Choose appropriate features to represent the text. Common choices include:

Character N-grams (unigrams, bigrams, trigrams).

Word N-grams.

Character and word frequencies.

Language-specific stopwords.

Character-level statistical measures.

Embeddings (e.g., FastText, Word2Vec) for semantic information.

• Labeling and Splitting:

Assign labels to each text sample based on its language.

Split the dataset into training and testing sets.

• Model Selection:

Choose a machine learning algorithm suitable for language detection. Common choices include:

Support Vector Machines (SVM).

Multinomial Naive Bayes.

Random Forest.

Neural Networks (especially for deep learning-based approaches).

• Model Training:

Train the selected model on the labeled training dataset.

Tune hyperparameters for optimal performance (if applicable).

• Model Evaluation:

Evaluate the trained model on the testing dataset using appropriate metrics:

Accuracy, Precision, Recall, F1 Score.

V. RESULTS

Page Interface

Screenshot 1

Screenshot 2

VI. CONCLUSION

Language detection using machine learning is a valuable and versatile tool that addresses the challenges of identifying the language of a given text or document in an automated and efficient manner. The development and deployment of language detectors have been fuelled by advancements in machine learning algorithms, feature engineering techniques, and the availability of diverse datasets. When dealing with a collection of texts published in several languages, it offers the capability of employing background knowledge about the language and using specialist methodologies. Thus, this project focuses on eliminating this problem by using a model designed using Machine Learning algorithms.

VII. REFERENCES

[1] Sengupta, D. and G. Saha, Study on Similarity among Indian Languages Using Language Verification Framework. Advances in Artificial Intelligence, 2015. 2015: p. 325703. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68– 73.

[2] Marco Lui, J.H.L., Timothy Baldwin, Automatic Detection and Language Identification of Multilingual Documents. Transactions of the Association for Computational Linguistics, 2014. 2: p. 27-40

[3] Revay, S., & Teschke, M. (2019). Multiclass language identification using deep learning on spectral images of audio signals. arXiv preprint arXiv:1905.04348.

[4] Jayanthi, N., Harsha, H., Jain, N., & Dhingra, I. S. (2020, February). Language Detection of Text Document Image. In 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN) (pp. 647-653). IEEE

[5] http://hmjournals.com/journal/index.php /IJITC/article/view/1754