LANGUAGE DETECTION USING MACHINE LEARNING

P. Harshita Krishna Sri [1], R. Tagore [2], P. Anil Kumar [3], M. Sowmya Sri [4], G. Sumanth [5],K. Lakshmi Narayana [6]

[1],[2],[3],[4],[5] Department of Computer Science and Artificial Intelligence (CAI), Sri Vasavi Engineering College, Tadepalligudem, Andhra Pradesh, India.

[6] Professor, Department of Computer Science and Engineering (CSE), Sri Vasavi Engineering College, Tadepalligudem, Andhra Pradesh, India.

ABSTRACT

This paper presents an innovative Machine Learning (ML) model for language detection that combines the power of logistic regression with a multimodal approach. The proposed model is designed to handle three types of inputs: sequential text data, files, and image representations. The proposed model offers a versatile and accurate solution for identifying languages across diverse data modalities. The model architecture employs logistic regression to enhance interpretability and feature extraction from each input modality. Trained on a comprehensive multilingual dataset, the model exhibits robust performance, showcasing its applicability to real-world scenarios. The model's ability to process text, files, and images makes it well-suited for applications in content filtering, cross-modal information retrieval, and multilingual sentiment analysis. This research contributes to the advancement of language detection models by offering a unified solution for handling diverse input types.

INTRODUCTION

Language detection, a critical component of natural language processing (NLP), holds substantial importance across various applications. Its impact extends from tasks like content filtering and sentiment analysis to facilitating cross-modal information retrieval. With the increasing prevalence of diverse data sources, ranging from textual content to multimedia representations, the need for a unified approach capable of handling multiple input modalities has become more pronounced. This paper introduces a novel multimodal language detection model that incorporates logistic regression within a comprehensive machine learning framework. Unlike traditional models focusing solely on sequential text data, our proposed model is designed to seamlessly process three distinct types of inputs: text, files, and images. This integration addresses the challenges posed by mixed-language data sources and provides a more accurate and versatile solution for language detection tasks.

LITERATURE SURVEY

In their work on "Hierarchical Character-Word Models for Language Identification," George Mulcaire, Aaron Jaech, Shobhit Hathi, Mari Ostendor, and Noah A. Smith introduce a model that employs convolutional neural networks (CNNs) to learn and identify languages based on characters. The term "hierarchical" used in their context suggests the utilization of a layered methodology, incorporating multiple levels of analysis. The model likely initiates its language identification process by examining the smallest linguistic units, such as individual characters, and progressively advances to higher-level units like words and phrases. This hierarchical approach is designed to effectively capture a broad range of language features, ranging from fine-grained details to more coarse-grained linguistic characteristics.

In the paper titled "Language Identification from Text Documents" authored by Priyank Mathur, Arkajyoti Misra, and Emrah Budur, the authors outline a comprehensive procedure for developing a language detection model. The process of language identification hinges on the distinct attributes of each language, encompassing elements like vocabulary, grammar, syntax, and character n-grams. These unique features serve as the foundation for constructing language profiles or models. These language profiles essentially act as repositories containing a compilation of language-specific features and statistical data, serving as benchmark models.

The study conducted by Shashank Simha B K, Rahul M, Jyoti R. Munavalli, and Prajwal Anand focuses on the implementation of Dual-Language Detection using Machine Learning. Their research aims to identify languages within the text by leveraging the distinctive characteristics of writing styles and unique diacritics associated with each language. The primary motivation for this dual-language detection approach stems from the need to effectively address the challenges posed by text data that incorporates code-switching, language mixing, or documents containing content in two separate languages. This challenge is particularly prevalent in multilingual and multicultural contexts. The objective is to develop a method that can robustly discern and process such complex linguistic scenarios.

Sowmya Vajjala and Sagnik Banerjee conducted experiments on N-gram and embedding representations for Native Language Identification (NLI) as part of the NLI Shared Task. The study explored various N-gram representations, including word, character, POS, and word-POS mixed representations for the task. Native Language Identification involves determining a person's native language based on their writing or speech, and it has applications in language education, forensic linguistics, and authorship attribution. The research aims to provide insights into effective feature representations for NLI.

SYSTEM STUDY AND ANALYSIS

PROBLEM STATEMENT

Language detection has traditionally focused on sequential text data, often overlooking the challenges posed by mixed-language content and diverse data modalities such as files and images. Existing models, while effective in monolingual text scenarios, struggle to provide accurate language identification when faced with the complexities of contemporary, multimodal data.

PROBLEMS IN THE EXISTING SYSTEM

Language detection models might struggle with short texts or texts with a lot of noise, as they might lack enough context to confidently determine the language.

Language usage and vocabulary can vary widely across different domains and contexts. Existing models trained in specific domains may struggle when applied to different contexts and may face challenges with struggling to confidently identify languages due to limited context.

Many traditional language detection models are primarily designed for sequential text data. They struggle to adapt to scenarios where data is not limited to textual content but includes files and images.

Existing models cannot often seamlessly process and integrate multiple types of inputs, such as text, files, and images, into a unified framework. This becomes a significant drawback in real-world applications where data is inherently multimodal.

PROPOSED SYSTEM

The proposed system is designed to accurately identify the language of text and content within images and files. A logistic regression model is trained on the text features to predict the language of textual content, files, and image features.

Our model is capable of even handling short texts or texts with noise, this feature separates our model from others. We accept three different types of input formats i.e., text files and images.

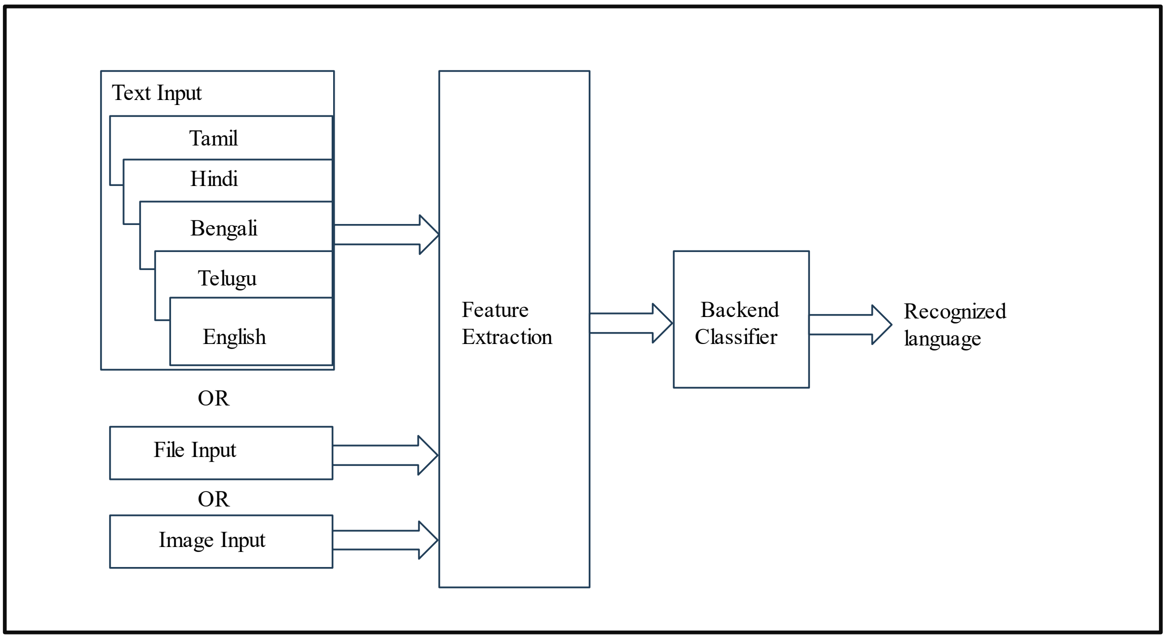


Fig-1: Proposed Model

Input

The user inputs text on the web page, which is sent to the backend. The text goes through cleaning before being analyzed by the language detection model. The detected language is then displayed back on the web page.

Feature Extraction

Feature extraction in language processing incorporates the use of n-grams, which are sequences of N-contiguous words or characters within a text. This technique involves capturing patterns and structures in language by examining consecutive sets of words or characters in a given textual context.

Backend Classifier

The backend classifier is the core of the language detection process. It takes the features generated from the text and uses a trained machine-learning model to predict the language.

Recognized Language

The output of the backend classifier is the recognized language. This is the language that the model believes is most likely to be the language of the input text.

METHODOLOGY

Data Collection and Preparation

- Collect a diverse dataset that contains a variety of language sentences with corresponding labels.
- Data should be cleaned and pre-processed to reduce noise and guarantee consistency.

Model Architecture

- Develop a multimodal model architecture that incorporates logistic regression.
- The pre-processed dataset is used to train the language detection model.
- Optimize the model to get maximum efficiency.
- Create a web interface to integrate the model so that is easy to use for smooth communication.

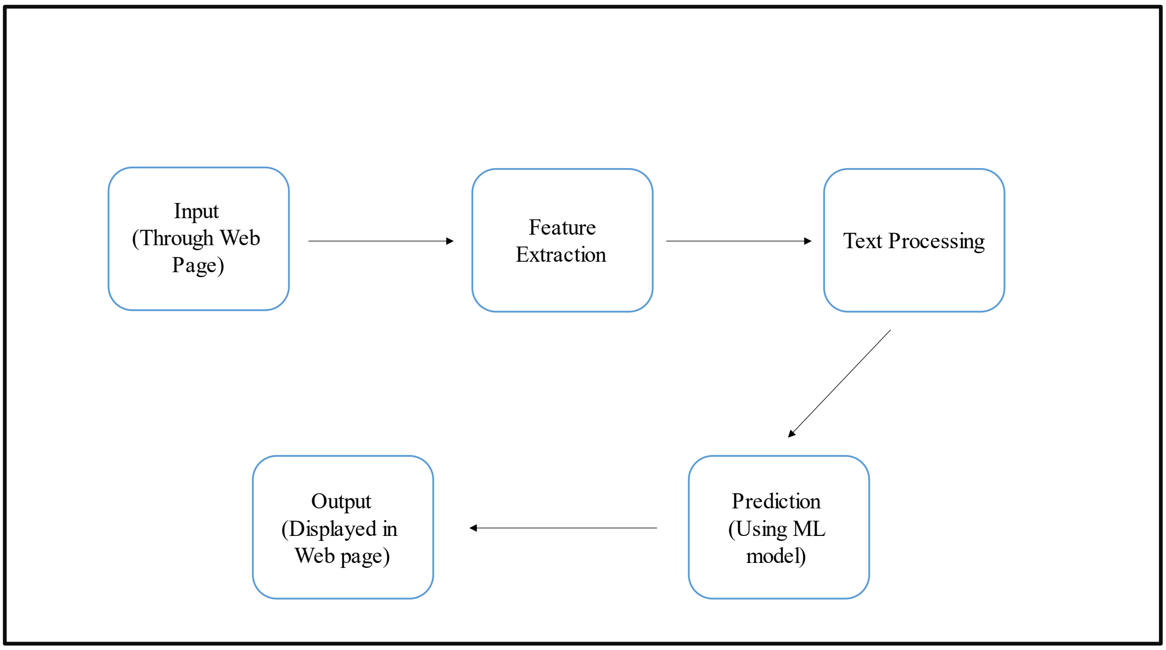


Fig-2: System Architecture

Training

- Split the dataset into training and validation sets.
- Train the chosen model using the pre-processed data.
- Monitor the model's performance on the validation dataset to prevent overfitting.

System input

- Provide a user-friendly input method that enables text input from users to identify languages.
- Provide users with precise input by implementing clear instructions or advice.

Feature Extraction

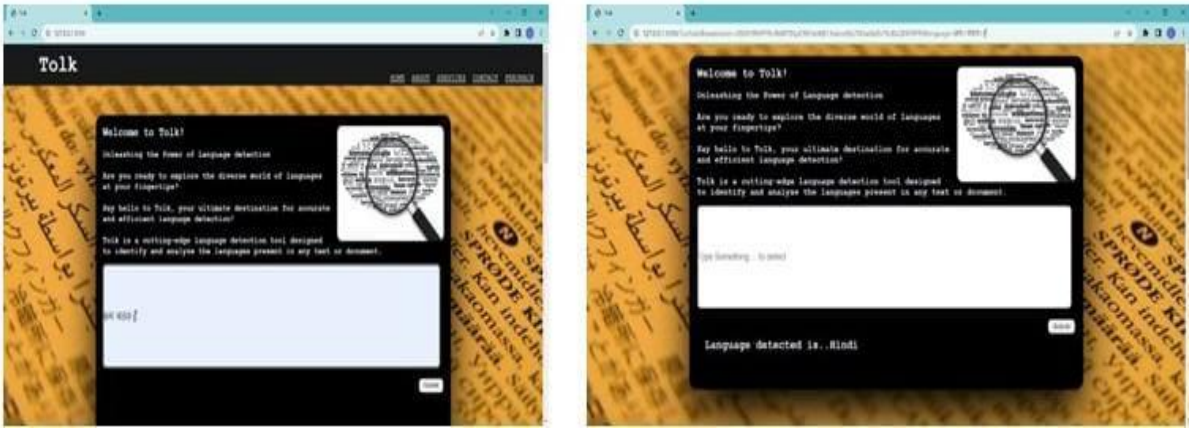
- Feature extraction in language processing incorporates the use of n-grams, which are sequences of N-contiguous words or characters within a text. This technique involves capturing patterns and structures in language by examining consecutive sets of words or characters in a given textual context.
- They can capture language-specific patterns and are commonly used for language detection.

Predicting output

- Utilize the learned model to predict the language of the input text.
- Present the anticipated outcome in a manner that is easily understandable.

EXPERIMENTAL RESULTS

The model accepts multi-lingual text input. It demonstrated commendable performance in extracting relevant text from user-given input i.e. by removing all the irrelevant characters and symbols. The model's adaptability to various linguistic styles and its capacity to process domain-specific content is high.



Welcome to Tolk!

Unleashing the Power of Language detection

Are you ready to explore the diverse world of languages at your fingertips?

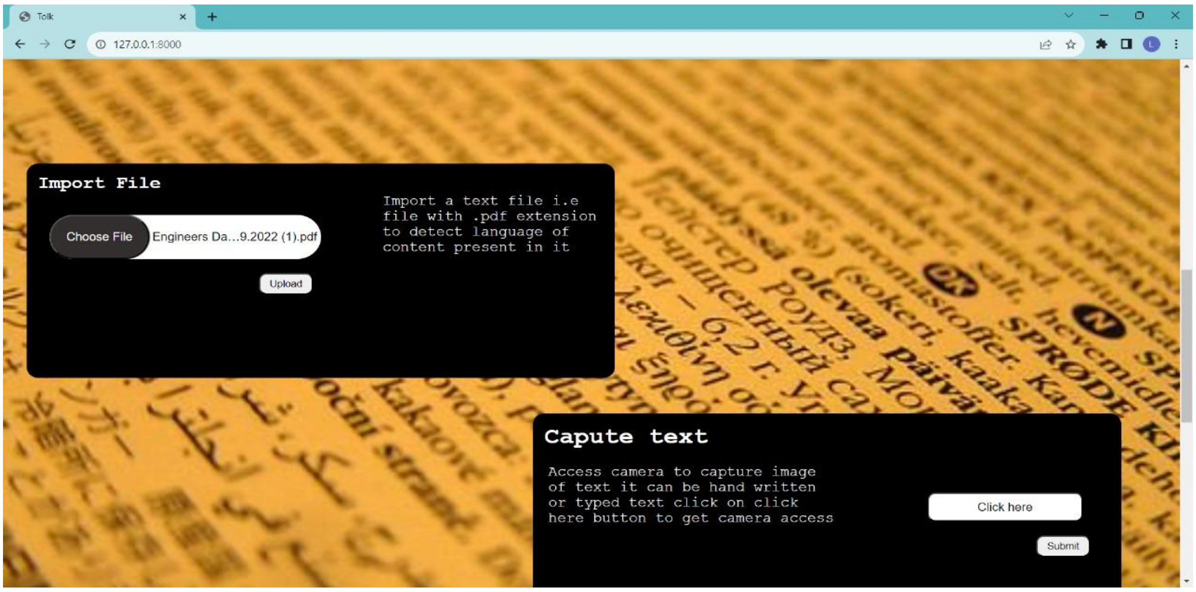
Say hello to Tolk, your ultimate destination for accurate and efficient language detection!

Tolk is a cutting-edge language detection tool designed to identify and analyse the languages present in any text or document.

Language detected is..Hindi

Fig-3: Text input & output

The results from the analysis of file inputs, specifically text files in PDF format, showcase the model's strength in extracting the text from the files of multi-lingual data. The model's adaptability to various linguistic styles and its capacity to process content is a promising solution for applications requiring analysis of text files in PDF format.



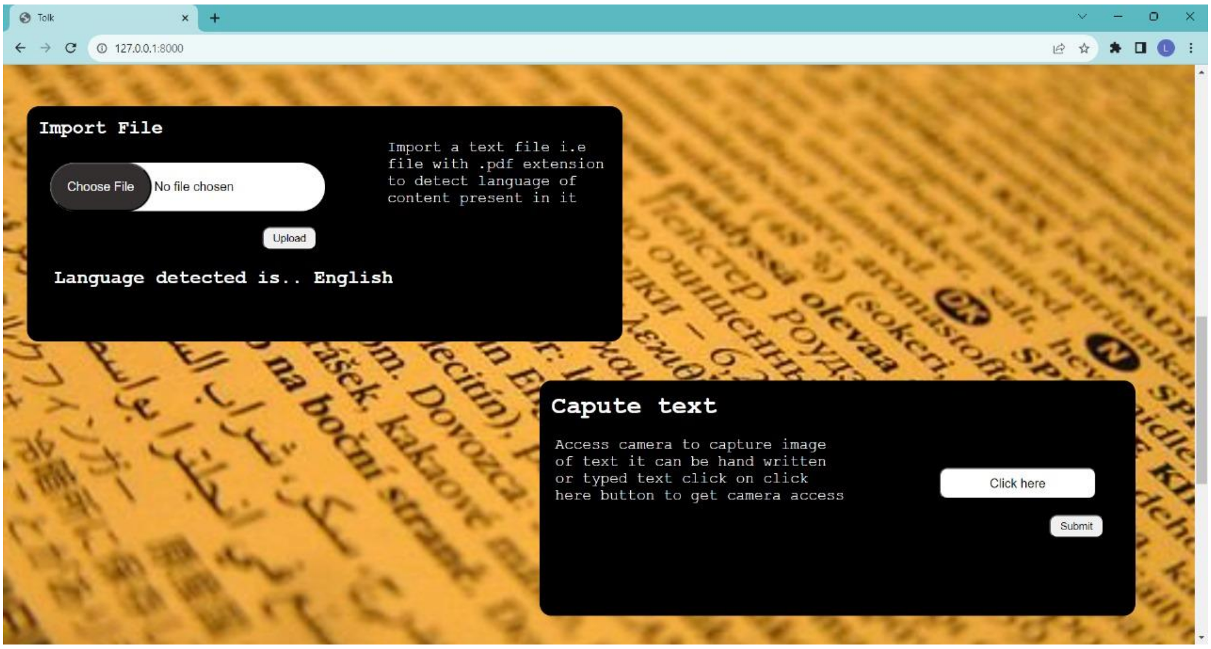
Import File

Import a text file i.e file with .pdf extension to detect language of content present in it

Capute text

Access camera to capture image of text it can be hand written or typed text click on click here button to get camera access

Fig-4: File input



Import File

Import a text file i.e file with .pdf extension to detect language of content present in it

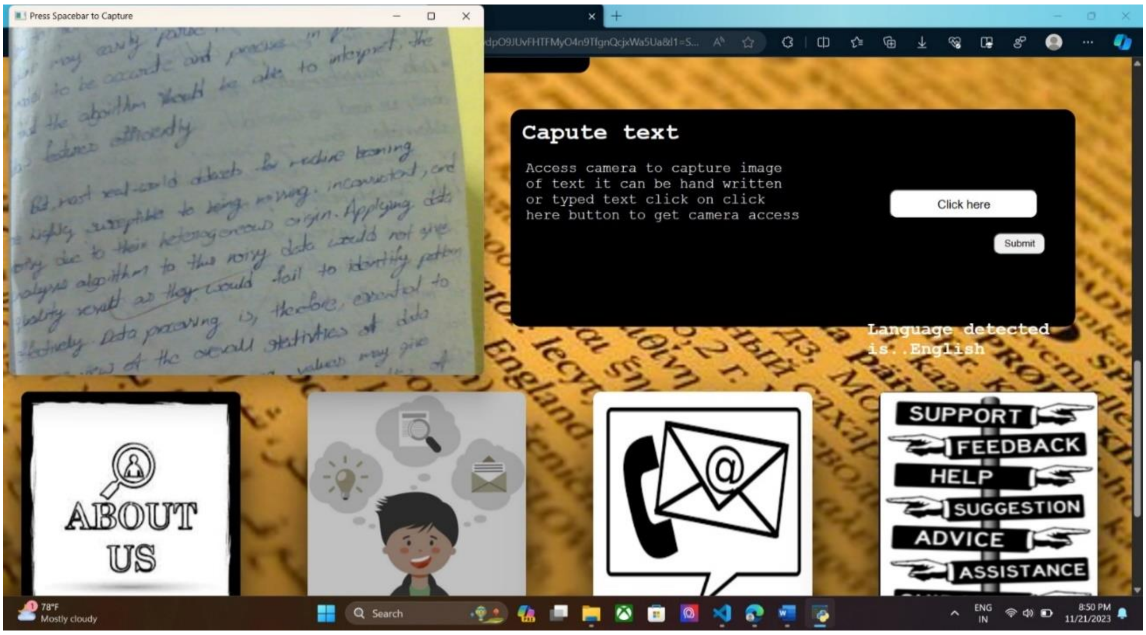
Language detected is.. English

Capute text

Access camera to capture image of text it can be hand written or typed text click on click here button to get camera access

Fig-5: Output for file input

The image input to the model can be given by capturing any images that can be handwritten or images that contain textual content. It showcases the ability to identify and classify by understanding the patterns or features embedded within the images.



Press Spacebar to Capture

may easily parse ... in
to be accurate and precise ... interpret, the
the algorithm should be able to
features efficiently

But most real-world datasets for machine learning
highly susceptible to being missing, inconsistent, and
noisy due to their heterogeneous origin. Applying data
analysis algorithms to this noisy data would not give
quality results as they would fail to identify patterns
effectively. Data processing is, therefore, essential to
of the overall statistics of data
values may give

Capute text

Access camera to capture image of text it can be hand written or typed text click on click here button to get camera access

Language detected is..English

Fig-6: Image input & output

CONCLUSION

Our language detection project has successfully addressed the fundamental challenge of identifying the language of text data, file content, and image content. Through a combination of robust data preprocessing, feature extraction, and machine learning techniques, we have developed a reliable and accurate language detection system. In conclusion, the proposed multimodal language detection model, integrating logistic regression within a comprehensive machine learning framework, represents a novel and effective solution in the realm of language processing. The model has been thoughtfully designed to address the limitations of existing systems by seamlessly handling three distinct input modalities such as text, files, and images providing a unified approach for accurate language identification.

REFERENCES

[1] Aaron Jaech, George Mulcaire, Shobhit Hathi, Mari Ostendor, Noah A. Smith: "Hierarchical Character-Word Models for Language Identification" (Aug 2016).

[2] Priyank Mathur, Arkajyoti Misra, Emrah Budur: "Language Identification from Text Documents" (2015).

[3] Shashank Simha B K, Rahul M, Jyoti R Munavalli, Prajwal Anand: "Dual-Language Detection using Machine Learning" (Dec 2022).

[4] Sowmya Vajjala, Sagnik Banerjee: "A study of N-gram and Embedding Representations for Native Language Identification" (September 8, 2017).

[5] Adarsh.D.Patil, Akshay Vishwas Joshi, Harsha. K.C, Pramod. N: "Spoken Language Identification Using Machine Learning" (May 2012).

[6] Marco Lui, Jey Han Lau, Timothy Baldwin: "Automatic Detection and Language Identification of Multilingual Documents" (Feb 2014).