Abstract

Language detection is a fundamental task in Natural Language Processing (NLP) with widespread applications, including multilingual content processing, automated translation, and user interaction enhancement in diverse linguistic settings. Accurately identifying the language of a given text is crucial for enabling seamless communication, improving accessibility, and enhancing user experience across digital platforms.

This project focuses on building an efficient language detection system using a **Naïve Bayes classifier**, a probabilistic machine learning algorithm well-suited for text classification tasks. The dataset used is the **European Parliament Proceedings Parallel Corpus**, which contains texts in 21 European languages, providing a rich and diverse linguistic resource for model training and evaluation.

The methodology involves multiple stages, starting with **data preprocessing**, including text cleaning, tokenization, and stop-word removal. **Feature extraction** is performed using **Term Frequency-Inverse Document Frequency (TF-IDF)** to convert textual data into a numerical format suitable for machine learning. The **Naïve Bayes classifier** is then trained on these processed features, leveraging its efficiency in handling high-dimensional text data and its ability to make probabilistic predictions based on word frequency distributions.

The model's performance is evaluated using standard classification metrics such as **accuracy, precision, recall, and F1-score**. Initial results indicate a **high accuracy rate**, demonstrating the model's ability to distinguish between different languages effectively.

This project contributes to the development of reliable language detection systems that can be integrated into **search engines, social media platforms, chatbots, and customer support systems**, where language identification plays a vital role in improving automated interactions and multilingual accessibility.

Naïve Bayes Algorithm Description

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem, assuming feature independence. Despite this “naïve” assumption, it performs well for text classification, including language detection.

Bayes' Theorem:

P(A|B)=P(B|A)·P(A)P(B)P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}P(A|B)=P(B)P(B|A)·P(A)

where P(A|B)P(A|B)P(A|B) is the posterior probability, P(B|A)P(B|A)P(B|A) is the likelihood, P(A)P(A)P(A) is the prior probability, and P(B)P(B)P(B) is the evidence.

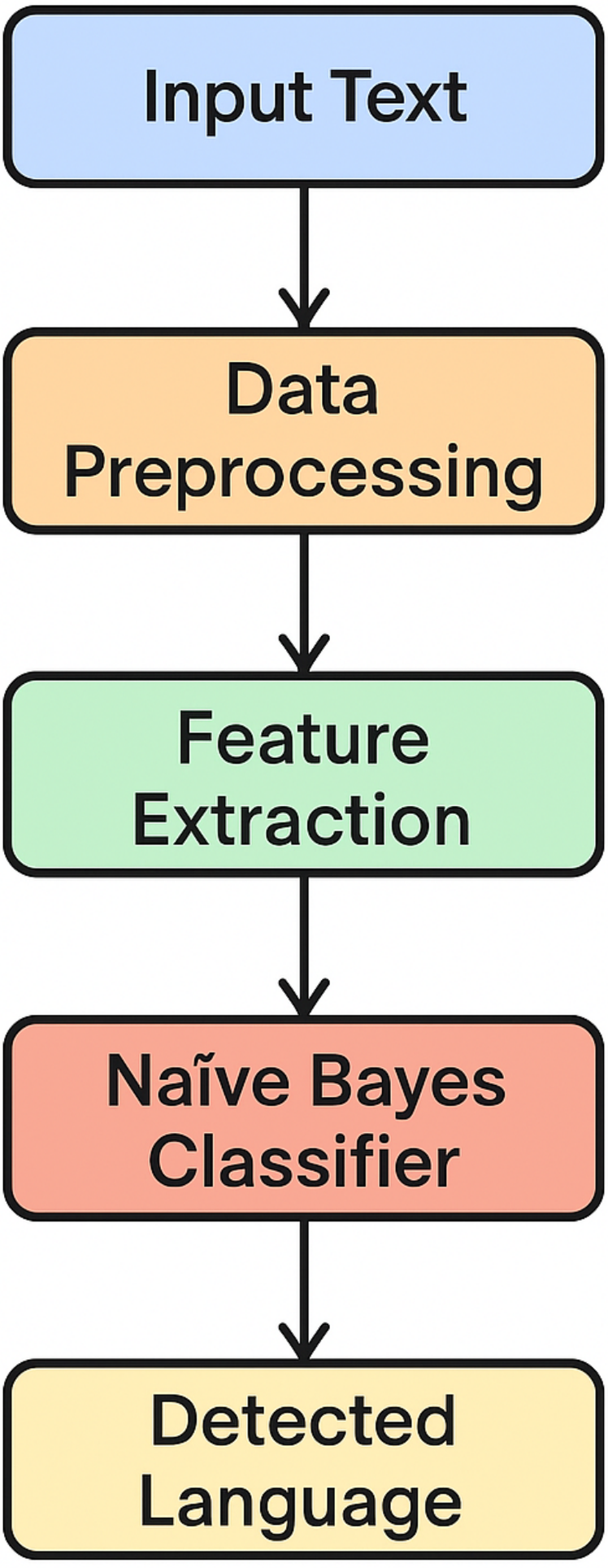
How It Works in Language Detection

1. Training Phase
    - The model learns from multilingual text data.
    - Text is tokenized and converted into numerical features using TF-IDF.
    - The algorithm calculates word probabilities for each language.
2. Prediction Phase
    - For a given text, it computes probabilities for all languages.
    - The language with the highest probability is selected.

Why Naïve Bayes?

- Fast and efficient for text classification.
- Handles large vocabularies effectively.
- Performs well even with limited training data.

Due to its speed and accuracy, Naïve Bayes is ideal for real-time language detection applications.


Input Text
Data Preprocessing
Feature Extraction
Naïve Bayes Classifier
Detected Language

Implementation Report: Language Detection Using Naive Bayes

1. Objective: The goal of this project is to build a machine learning model that can classify text into different languages based on its content using the Naive Bayes classifier.

2. Dataset: The project uses a dataset named `language.csv`, which contains text samples along with their corresponding language labels.

3. Data Preprocessing:

- Loading Data: The data is read from the CSV file using `pandas.read_csv()`.

- Handling Missing Values: Any rows with missing language labels are removed using `dropna()`, ensuring that the dataset is clean and complete.

- Feature Extraction: The `CountVectorizer` is used to convert the text data into a numerical format (bag-of-words representation), where each word is represented as a feature in the dataset.

4. Model Training:

- Splitting Data: The dataset is divided into training and testing sets using `train_test_split()` from `sklearn`. 33% of the data is reserved for testing, and the remaining 67% is used for training.

- Model Selection: The `MultinomialNB()` model is chosen, which is particularly effective for text classification tasks where the features (words) are independent.

- Training: The model is trained on the training set using `model.fit()`.

5. Model Evaluation:

- The accuracy of the model is evaluated using `model.score()`, which compares the predicted language labels with the true labels in the test set.

6. User Input:

- The model can also predict the language of new, user-provided text. After the user inputs a piece of text, the model uses the `CountVectorizer` to transform the input text into a numerical format, then predicts the language using `model.predict()`.

7. Output:

- The model outputs the predicted language for the given input text.

Conclusion: The model successfully classifies text into different languages using a Naive Bayes classifier. The process involves text preprocessing, splitting the data into training and testing sets, training the model, evaluating its performance, and allowing for user input.

International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Systematic Review on Offensive Language Detection of Multilingual Texts using NLP and Machine Learning Techniques

Roopa G K

Assistant Professor, Department of Computer Science &Engg, VCET, Puttur
Email: *roopa.rkumbady[at]gmail.com*

Abstract: ***The widespread usage of social media and online platforms has given rise to a substantial increase in offensive language. Detecting offensive content is crucial to maintaining a healthy online environment and protecting users from harm. Natural Language Processing (NLP) and Machine Learning (ML) techniques have shown promise in addressing this challenge, especially in the context of multilingual texts. This paper presents a systematic review of the existing literature on offensive language detection in multilingual texts, focusing on the NLP and ML methodologies utilized, dataset characteristics, evaluation metrics, and performance comparisons. The review aims to provide an overview of the state-of-the-art techniques, identify key challenges, and suggest future research directions in this area.***

Keywords: Natural Language Processing (NLP), Machine Learning (ML), multilingual texts

1. Introduction

The proliferation of social media platforms and online communication channels has provided an unprecedented avenue for individuals to express their thoughts and opinions. However, this digital revolution has also brought to light the dark side of online discourse - the rampant use of offensive language, hate speech, and harmful content. Offensive language not only poses a threat to the emotional well-being of users but also undermines the quality of online interactions and fosters a toxic environment.

Efficiently detecting offensive language in multilingual texts is a challenging task, as it requires understanding the intricacies of various languages, cultural contexts, and linguistic nuances. Traditional rule-based methods and keyword filtering are inadequate in dealing with the dynamic and context-dependent nature of offensive content. To address this issue, researchers have turned to Natural Language Processing (NLP) and Machine Learning (ML) techniques, which have shown considerable promise in automating the identification of offensive language across different languages and settings.

The motivation behind this systematic review is to comprehensively analyze the state-of-the-art offensive language detection approaches that employ NLP and ML techniques for multilingual texts. By examining existing literature and research, we aim to provide a comprehensive overview of the methodologies used, assess the performance of various models, and identify the key challenges that researchers encounter in this domain. Furthermore, this review seeks to suggest potential research directions to improve the accuracy and generalizability of offensive language detection systems.

Objectives

The main objectives of this systematic review are as follows:

1) Provide an extensive overview of the NLP and ML techniques utilized for offensive language detection in multilingual texts.
2) Investigate and assess the characteristics of available multilingual offensive language datasets, including their diversity, size, and language distribution.
3) Analyze the performance of state-of-the-art models and techniques for offensive language detection across different languages and compare their effectiveness.
4) Identify and discuss the challenges and limitations faced by existing approaches, such as handling data imbalance and cross-linguistic ambiguity.
5) Propose future research directions to enhance the capabilities and robustness of offensive language detection systems, including the incorporation of context, pragmatics, and ethical considerations.

2. Offensive Language Detection: Overview

Offensive language detection is a critical area of research with significant implications for online communities, platforms, and users. The combination of NLP and machine learning techniques has led to substantial progress in detecting offensive content. However, challenges such as multilingual variations, context understanding, and evolving language usage require continuous research and innovation to develop robust and culturally sensitive offensive language detection systems.

2.1 Definition and Types of Offensive Language

Offensive language encompasses a wide range of harmful content, including hate speech, profanity, cyber bullying, harassment, and discriminatory remarks. These expressions can target individuals or groups based on attributes such as

DOI: 10.21275/SR23802115009

Offensive Language Detection on Social Media using Machine Learning

Rustam Abdrakhmanov[1], Serik Muktarovich Kenesbayev[2], Kamalbek Berkimbayev[3], Gumyrbek Toikenov[4], Elmira Abdrashova[5], Oichagul Alchinbayeva[6], Aizhan Ydyrys[7]
International University of Tourism and Hospitality, Turkistan, Kazakhstan[1]
Kazakh National Women's Teacher Training University, Almaty, Kazakhstan[2, 4]
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan[3]
M. Auezov South Kazakhstan University, Shymkent, Kazakhstan[5, 6]
International Information Technology University, Almaty, Kazakhstan[7]

***Abstract*—This research paper addresses the critical issue of cyberbullying detection within the realm of social networks, employing a comprehensive examination of various machine learning and deep learning techniques. The study investigates the performance of these methodologies through rigorous evaluation using standard metrics, including Accuracy, Precision, Recall, F-measure, and AUC-ROC. The findings highlight the notable efficacy of deep learning models, particularly the Bidirectional Long Short-Term Memory (BiLSTM) architecture, in consistently outperforming alternative methods across diverse classification tasks. Confusion matrices and graphical representations further elucidate model performance, emphasizing the BiLSTM-based model's remarkable capacity to discern and classify cyberbullying instances accurately. These results underscore the significance of advanced neural network structures in capturing the complexities of online hate speech and offensive content. This research contributes valuable insights toward fostering safer and more inclusive online communities by facilitating early identification and mitigation of cyberbullying. Future investigations may explore hybrid approaches, additional feature integration, or real-time detection systems to further refine and advance the state-of-the-art in addressing this critical societal concern.**

***Keywords*—Machine learning; deep learning; hate speech; CNN; RNN; LSTM**

I. Introduction

The advent of social media has revolutionized the way individuals communicate, providing platforms that facilitate rapid information dissemination and interaction across global communities. While these platforms have empowered users to share information and foster connections, they have also become breeding grounds for various forms of online abuse, including hate speech. Hate speech encompasses any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics. It poses severe risks to community harmony, individual safety, and democratic discourse [1]. Consequently, the detection and mitigation of hate speech on social media is of paramount importance for maintaining social cohesion and protecting vulnerable groups.

The challenge of combating hate speech on social media is amplified by the vast amount of data generated daily and the fluid nature of online communication. Traditional content moderation methods, which rely heavily on human moderators to review content, are not scalable to the volumes of data produced on platforms such as Facebook, Twitter, and Instagram. Furthermore, manual moderation is prone to inconsistencies and errors, making it an inefficient solution in the dynamic and diverse environment of social media [2]. As a result, there has been a significant shift toward automated systems, particularly those utilizing machine learning (ML) and deep learning (DL), to address the complexities associated with identifying and managing hate speech [3].

Machine learning offers a promising approach to automate the detection of hate speech by learning from large datasets of labeled examples. It uses natural language processing (NLP) to parse and understand the textual content of social media posts, learning to differentiate between harmful and harmless expressions based on training data [4]. Unlike rule-based systems, which fail to adapt to the evolving language of online communities, ML algorithms can update their knowledge as new data becomes available, thereby adapting to changes in the lexicon used in hate speech [5].

Deep learning, a subset of ML characterized by models that learn through layers of neural networks, has shown exceptional capability in handling the intricacies and subtleties of human language. DL models, particularly those based on recent advancements such as transformer architectures, have demonstrated high accuracy in contextual understanding and sentiment analysis [6]. These models are particularly adept at capturing the contextual nuances that differentiate hostile or derogatory speech from benign usage of potentially sensitive words [7].

The application of ML and DL in detecting hate speech is not without challenges. One significant issue is the balance between accuracy and the rate of false positives—where benign content is incorrectly flagged as hate speech. High rates of false positives can lead to unnecessary censorship and could impact user engagement and trust in social media platforms [8]. Another challenge is the development of models that can operate across different languages and cultural contexts, as hate speech often involves cultural references and idioms that are not universally recognized [9].

Recent studies have applied various ML and DL models to address these challenges, employing sophisticated algorithms

Indian Language Identification using Deep Learning

Shubham Godbole[1,*], *Vaishnavi* Jadhav[2,**], and *Gajanan* Birajdar[3,***]

[1]Department of Electronics Engineering, Ramrao Adik Institute of Technology, Nerul, Navi Mumbai, India

Abstract. Spoken language is the most regular method of correspondence in this day and age. Endeavours to create language recognizable proof frameworks for Indian dialects have been very restricted because of the issue of speaker accessibility and language readability. However, the necessity of SLID is expanding for common and safeguard applications day by day. Feature extraction is a basic and important procedure performed in LID. A sound example is changed over into a spectrogram visual portrayal which describes a range of frequencies in regard with time. Three such spectrogram visuals were generated namely Log Spectrogram, Gammatonegram and IIR-CQT Spectrogram for audio samples from the standardized IIIT-H Indic Speech Database. These visual representations depict language specific details and the nature of each language. These spectrograms images were then used as an input to the CNN. Classification accuracy of 98.86% was obtained using the proposed methodology.

Index Terms: Convolutional Neural Network (CNN), Spoken Indian Language Identification (SLID), Log Spectrogram, gammatonegram, IIR-CQT Spectrogram, Artificial Neural Network (ANN), Deep Learning

1 Introduction

Inventive systems like Siri and Google Assistant depend on Automatic Speech Recognition (ASR). In order to work appropriately the ASR frameworks expect users to manually indicate the proper input language. Traditional Language Identification (LID) systems use area explicit information for extracting hand-made features from sound samples[4].

Recently, Deep Learning and Artificial Neural Networks (ANN) are been considered the best in class for pattern recognition issues[25]. A variety of computer vision tasks like Image Classification, display better performance using Deep Neural Networks. LID can be characterized as the task of recognizing the spoken language in any given utterance.

As research in ASR advances, a LID system would be important for any multi-lingual speech recognition system. In case of multi-lingual speech recognition, the accuracy of the speech recognizer system can be improved by using LID setup at the front-end. It reduces complication by directly processing the speech over the identified language rather than running over several languages.

Particularly, in an Indian scenario where nearly every state has its very own language and each language having many lingos, designing LID system becomes vital. The languages and dialects used in India are categorized into distinct families: the significant ones are Indo-Aryan, Dravidian, Sino-Tibetan, Austroasiatic, Tai–Kadai and Great Andamanese languages [16]. This experimental study involves languages from two of the family groups. Marathi, Hindi, Bengali belong to the Indo-Aryan family whereas Tamil, Telugu, Kannada and Malayalam belong to the Dravidian family[4].

Neural networks are used to classify the different languages. The issue regarding LID can also be modeled as a pattern recognition task. Patterns from the given spectrogram can be identified by the trained ANN which helps it to classify the unfamiliar language specimen into a known category. Seven languages from IIIT-H Indic speech databases are used for evaluation[1].

In our methodology, WAV files from the database were converted into spectrogram visuals. Each sample is represented by separate visuals viz. Log Spectrogram, Infinite Impulse Response Constant-Q Transform (IIR-CQT) Spectrogram and Gammatonegram. Convolutional Neural Network (CNN) is used for classification of different languages. The classification accuracy rate obtained using CNN is 98.86%.

Formulation of the paper is done in the following manner: In section II wide range of methodologies and approaches for LID are summarized. Section 3 presents the architecture for LID using spectrogram visuals. Simulation results and comparison between the proposed method and available techniques are introduced in Section 4. Section 5 concludes the article.

2 Literature Survey

In LID various features and methods such as acoustic, phonotactic and prosodic along with different classifiers are used to differentiate the languages. In one of the previous works, sound samples from the dataset [1] were converted into spectrogram visuals which were then classi-

*e-mail: godboleshubham2@gmail.com
**e-mail: vaishnavijadhav1998@gmail.com
***e-mail: gajanan.birajdar@rait.ac.in

Language Detection Using Naive Bayes Model

D. LIKHITH REDDY[1], K. PALLAVI[2], M. PRIYANKA[3], A. SUJITHA[4], J.V. SUMANTH[5], G. VINAY[6]
[1]*Associate Professor,* [2]*UG Student,*[3]*UG Student,* [4]*UG Student,* [5]*UG Student,* [6]*UG Student*
[1]*Dept. of ECE, PBR Visvodaya Institute Of Technology and Science, Kavali.*
[2,3,4,5,6]*Dept. of ECE, PBR Visvodaya Institute Of Technology and Science, Kavali, Andhra Pradesh, India.*

Abstract- This project explores the application of a Naive Bayes classifier for language detection in text documents. We implement the model using Python and the scikit-learn library, training it on a diverse dataset of text samples in different languages. Through rigorous experimentation and evaluation, we assess the classifier's accuracy, precision, and recall across various languages. The results demonstrate the effectiveness of the naive Bayes approach in accurately identifying the language of textual content, highlighting its potential for practical language detection applications.

Index Terms: Feature Extraction, Naive Bayes Model, Detection , Classifier

I. INTRODUCTION

Language detection is a fundamental aspect of natural language processing, essential for various applications like machine translation, sentiment analysis, and information retrieval. It involves determining the language of a given piece of text, which can be a single word, sentence, or entire document. There are several approaches to language detection, including statistical methods, rule-based systems, and machine learning algorithms.

Statistical methods analyze the frequency of characters, words, or n-grams in a text to make language predictions. Rule-based systems use linguistic rules and patterns specific to eachlanguage to identify them. Machine learning algorithms, such as neural networks and support vector machines, learn language patterns from labeled training data and make predictions based onlearned features.

Continued advancements in machine learning and natural language processing techniques are expected to further enhance the accuracy and efficiency of language detection systems. As the volume and diversity of digital content continue to grow, robust language detection capabilities will remain crucial for effectively managing and analyzing multilingual data.

II. LITERATURE SURVEY

[1] X. Rong, "word2vec Parameter Learning Explained," pp. 1–21, 2014.

This seminal paper provides by Rong provides a comprehensive explanation of the Word2vec algorithm, which has revolutionized natural language processing by enabling the efficient generation of word embeddings. Word embeddings capture semantic similarities between words by representing them as dense vectors in a continuous vector space.

[2] G. G. Chowdhury, "Natural language processing," 2003.
Chowdhury's work serves as a foundational resource in the field of natural language processing (NLP), providing an introductory overview of key concepts, techniques, and algorithms. The paper covers various topics relevant to language detection tasks, including text preprocessing, feature extraction, and text classification.
[3] S. M. Mohammad, "Sentiment Analysis : Detecting Valence, Emotions, and Other Affectual States from Text," 2015.
Mohammad's paper delves into the field of sentiment analysis, which involves classifying textinto categories based on emotional content. The ability to analyze sentiments is crucial for language detection tasks, as different languages often exhibit unique emotional expressions andlinguistic cues.

The paper explores various techniques and methodologies for sentiment analysis, providing valuable insights into the semantic nuances and linguistic patterns that characterize different languages. By understanding sentiment analysis,

Language Detection using Natural Language Processing

Dr. A V Sriharsha[1*], Muthyala Reddy Jahnavi[2], Desai Sakethram Kousik[3], Vukyam Hemanth[4],Matchandrappa Gari Hari[5], Penchala Praveen Vasili[6]

[1] Professor, Department of CSE(DS), Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), India
[2, 3, 4, 5]UG Scholar, Department of Computer Science and Systems Engineering, Sree Vidyankethan Engineering College, Tirupati, India.
[6]Product Manager, Wellsfargo Inc. Charlotte, USA

[1*] avsreeharsha@gmail.com, [2] reddyjahnavi3013@gmail.com
[3] sakethdesai220503@gmail.com, [4] vukyamhemanth12@gmail.com,
[5]hariediga1050@gmail.com, [6]penchalapraveen@gmail.com

Abstract. Natural Language Processing (NLP) is a rapidly advancing field of artificial intelligence that actsas a bridge between human language and machines. Its uses vary from language translation andsentiment analysis to virtual assistants, impacting a wide range of industries. Language detection is a crucial sub-task of NLP that automatically recognizes the language in a given text. The Mul- tinomial Naive Bayes classifier's effectiveness and performance in text classification, along with NLP feature engineering, make it a suitable option for language detection tasks, even when work-ing with multilingual datasets. By integrating NLP techniques and the Multinomial Naive Bayes classifier, the proposed method offers a strong and precise language detection approach. Exper-iments conducted on diverse textual data show promising outcomes, even when dealing with noise and incomplete information. Accurate language identification improves the usability and efficiency of various NLP applications, promoting cross-cultural communication and contrib- uting to a more inclusive and interconnected digital environment.

Keywords: Natural Language Processing (NLP), Multinomial Naive Bayes classifier, Artificial intelligence, Language translation, Performance evaluation, Text classification.

1 Introduction

Language detection is a crucial task in the field of natural language processing (NLP) and holds significant importance in a wide range of applications that deal with textual data. The main objective of language detection is to automatically determine the language of a given file or document, enabling more precise and contextually aware processing. Accurate language detection is essential in a multilingual world where content filtering, user localization, and information retrieval are increasingly important. The core principles of language detection involve the use of statistical and machine learning methods to distinguish one language from another based on textual features. The capability of deep learning to automatically learn and extract significant patterns from text data has led to increasing popularity as a subset of machine learning. Multinomial Naive Bayes, a probabilistic algorithm, is well-suited for language detection tasks by modeling the probability distribution of words in different languages. Combining deep learning and Multinomial Naïve Bayes offers a robust framework for accurate language identification. Language detection has a wide range of practical applications. Content filtering helps identify and categorize content according to lan- guage preferences, ensuring that users receive content in their preferred languages. User locali zation

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_49

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

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.

Dual Language Detection using Machine Learning

[1]Shashank Simha B K, [2]Rahul M, [3]Jyoti R Munavalli and [4]Prajwal Anand
[1,2,3,4] Dept. of ECE, BNM Institute of Technology, Bengaluru, India.
[1]bk.shashanksimha@gmail.com, [2]rahul.gowda.76@gmail.com, [3]jyotirmunavalli@bnmit.in, [4]praj460@gmail.com

Article Info
Jenitta J and Swetha Rani L (eds.), *International Conference on VLSI, Communications and Computer Communication*, Advances in Intelligent Systems and Technologies,
Doi: https://doi.org/10.53759/aist/978-9914-9946-1-2_32

Abstract- There are number of languages around the world and knowing all the languages is very difficult for any person. At the same time, unawareness about the language will hinder communication. Language identification is the process where the identifying the language(s) in text form is performed based on the writing style and looking at the unique diacritics of each language. When a multitude of languages are spoken in any circumstances, the first step in communication is the identification of the language. There are several techniques used for language detection like machine learning and deep learning. These are used in detecting languages like German. In India, numerous languages are spoken by the people and thus we propose to develop a model that detects two languages: Kannada and Devanagari/Sanskrit. In this study, Support Vector Machines classifiers were used, for classification and an accuracy of 99% was achieved.

Keywords- Machine Learning, Support Vector Machine, Artificial Intelligence, Python, Language, Kannada, Devanagari, Sanskrit.

I. INTRODUCTION

Language is the basic way of communication that is required to exchange information between people. Each person acquires an ability, right from their childhood, to make use of this exchange of information either through sounds or gestures. As we progress, we tend to learn various languages to communicate among others. Language is not only limited to communicate thoughts and ideas, but it also builds friendship, career, strengthens the economic relationships of businesses and provides an understanding of diverse cultures scattered across the globe. Thousands of languages exist around the world which are used in everyday life. Most of the time, during visit to other places, a person would not be able to understand or communicate in their native language. So, here comes the role of Language Identification which determines the language of the written content given to it and help the user to communicate effectively.

Today, the world is moving towards Artificial Intelligence in every aspect of life. A generation of computers that are capable of depicting the thoughts and actions of human. One of the wonders of the world is the human visual system, which we are attempting to imitate. Any activity that a system undertakes must first be known to it in order for it to begin. Humans have ears, eyes, and a brain that they use to receive information and act on it. In a similar way, speech recognition and word recognition on their own would serve as a computer's eyes and ears, respectively. The most recent technology is used to recognize sign language and other languages.

In this paper, we develop a model that identifies both Kannada and Devanagari languages. The paper is structured as follows: Section II highlights the literature survey and the proposed methodology; Section III presents the results of the model and Section IV is conclusion.

II. MATERIALS AND METHODS

A literature survey was carried out for understanding language detection techniques and the languages that are detected. Scopus, Web of Science, and Google Scholar were the databases that were searched using the terms "language detection," "Machine Learning" and "Indian language." It was found that different research papers had detected different languages. In this paper, we propose to develop a language detection mechanism that is text based, for the two Indian Languages scripts, Kannada and Devanagari using Machine Learning techniques.

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 (Fig. 1.).

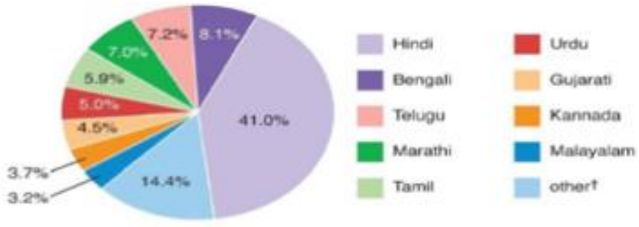


Fig 1. Statistics of Indian Language

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Conference:01/December-2022 Impact Factor- 6.752 www.irjmets.com
1st National Conference on Applications of soft Computing Techniques in Engineering NCASCTE-2022
Organized by Department of Electrical & Electronics Engineering, ACE Engineering College, Hyderabad

LANGUAGE DETECTION USING MACHINE LEARNING

K Prem Kumar *1, T SRI VINAY*2, S Sai Aasritha*3, P Vasantha*4

*1Associate Professor, Department of Computer Science and Engineering ,ACE Engineering College, Ghatkesar, Telangana,India

*2,*3,*4Btech Student,Department of Computer Science and Engineering ,ACE Engineering College, Ghatkesar, Telangana,India

DOI: https://www.doi.org/10.56726/IRJMETS-NCASCTE202219

ABSTRACT

In today's globalized world, where people from all over the world are able to communicate and share information with each other. Language detection is important as it helps to bridge the gap between different cultures and languages. Language detection is a useful tool in many applications, such as machine translation, text summarization, and sentiment analysis. In these applications, it is important to first determine the language of the text before processing it further. Overall, language detection using NLP is a powerful tool for analyzing and processing natural language data. It can be used in a variety of applications to improve the accuracy and efficiency of natural language processing systems.

Keywords--Naive Bayes, Language Detection, Machine Learning, Natural Language Processing.

I. INTRODUCTION

Language identification can be a crucial step in a Natural Language Processing (NLP) problem. It will try to identify the natural language from a text. Language identification of a text is very important before language translation, sentiment analysis can be taken. For example, when we want to translate a language in google translate the box you type in says 'Detect Language. The reason is firstly Google translate is trying to identify the language of the sentence which we gave before translating it. Language identification can be implemented with the help of Neural Network, character n-grams, Frequent word-based approach etc. Many researchers have developed the models that can predict different languages with different algorithms. Research on language identification began in the 1970's.

At the end of almost 5 decades of research, we have seen that language recognition has been practiced in different ways. Many attempts by different researchers have been made to achieve maximum efficiency

II. METHODOLOGY

Previous Work

In "**n-gram and Decision Tree-based Language Detection for written words**" By JuhaHakkinen and Jilei Tian, N- Gram-Based approach is used. Intuitively, common words such as determiners, conjunctions and prepositions are good clues for identifying the language.

The n-gram method uses letter n-grams, which represents the frequency of occurrence of different n-letter combinations in a particular language. Language identification process can be divided into two phases: Training and Identification. A language identification model is trained for each targeted language. In the training phase, a list of words with a known language are presented as alphabetic strings. The frequency of occurrence of sequence of consecutive n letters is estimated from a large language specific training sample. Since it is not feasible to train all the possible partial letter sequence probabilities, a simplifying assumption is made that the probability of the current word depends only on the previous n-1 letters, which can be implemented using n-grams.

In Decision Trees Based Approach, Decision trees are used to determine the most likely language for each letter in word. The language is obtained by asking a series of questions about the context of the current letter, as defined by the corresponding decision tree. Since only the letter context is used and no frequency information is stored in the tree, a very compact representation is obtained. During training, the decision tree is grown by splitting the nodes into child nodes. Language tags are first generated for each letter of the input word. The

LANGUAGE IDENTIFICATION FOR MULTILINGUAL MACHINE

[1]GEETHA PRATHIBA, [2]VELPULA SHRAVYA PATEL, [3]PATHAK SHIVANI, [4]UPPUTALLA DIVYA

[1]Assistant Professor, Department of Information Technology, **MALLA REDDY ENGINEERING COLLEGE FOR WOMEN**, Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

[2, 3,4] Student, Department of Information Technology, **MALLA REDDY ENGINEERING COLLEGE FOR WOMEN**, Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

ABSTRACT

Language Identification for Multilingual Machine Translation is a crucial component in modern natural language processing systems, enabling accurate and efficient translation across multiple languages. This paper presents a comprehensive approach to language identification that enhances the performance of multilingual machine translation systems.The proposed method utilizes advanced machine learning techniques to automatically detect the language of a given text with high accuracy. By incorporating a variety of linguistic features and leveraging large-scale multilingual datasets, the system can identify languages even in challenging scenarios such as code-switching and mixed-language inputs.Key contributions of this work include the development of a robust language identification model that integrates seamlessly with machine translation pipelines. The model is evaluated on diverse datasets, demonstrating its effectiveness in real-world applications. Additionally, we explore the impact of accurate language identification on the overall quality of machine translation, highlighting improvements in translation accuracy and fluency.

I.INTRODUCTION

Language identification is a fundamental task in natural language processing (NLP) that involves determining the language of a given piece of text. It serves as a critical preprocessing step for various applications, including multilingual machine translation, where accurate language identification is essential for ensuring high-quality translations. As the world becomes

2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) | 979-8-3503-9737-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACCS57279.2023.10112773

Language Detection Using Natural Language Processing

Yutika Rajanak
CSE (Artificial Intelligence and Machine Learning)
Noida Institute of Engineering and Technology
Greater Noida, Uttar Pradesh, India
yutika.rajanak@gmail.com

Ratna Patil
AIML Department
Noida Institute of Engineering and Technology
Greater Noida, Uttar Pradesh, India
ratna.nitin.patil@gmail.com

Yadvendra Pratap Singh
Computer Science and Engineering Department,
Manipal University Jaipur,
Jaipur, Rajasthan, India
yadvendra.mnnit@gmail.com

***Abstract*— Natural language processing (NLP) is a method for correctly identifying text based on the provided content or topic matter. An extensive study will make it simple to interpret any language and comprehend what is being said. Despite the fact that NLP is a challenging technique, notable examples include Siri and Alexa. Natural language detection allows us to determine the language being used in a given document. A Python-written model that has been utilised in this work can be used to analyse the basic linguistics of any language. The "words" that make up sentences are the essential building blocks of knowledge and its expression. Correctly identifying them and comprehending the situation in which they are used are essential. NLP steps in to help us in this circumstance by making it easier for us to identify the linguistics used in a particular piece of information, whether it be written or vocal. NLP gives computers the ability to understand human language and respond correctly, performing language detection for us. The current paper provides a summary of developments in tongue process, including analysis, establishment, various areas of rapid advancement in natural language processing research, development tools, and techniques.**

Keywords—Natural Language Processing, Language Detection, Virtual Assistants, Text Analytics, Machine Learning

I. INTRODUCTION

Natural Language Processing (NLP) is a technique for processing languages and transforming them into forms that the user can readily process or interpret. NLP is a method of computer programming that is based on pattern learning [1]. It consists of two parts i.e., Natural Language Understanding (NLU) and Natural Language Generation (NLG).

We can use NLU to determine the meaning of a specific word or passage of text, whether it is written or spoken. Using a representation of text or data, NLG creates meaningful sentences.

NLP is the foundation of how Language Detection operates. Language is processed and identified using NLP. With the aid of NLP, different word and language types can be detected.

NLP aids in analyzing presented text and identifies language and word meaning. NLP makes it simple to recognise business writings. By identifying the datasets to which each language belongs and evaluating the text to determine its meaning and intent, NLP assists us in implementing numerous languages and detecting them. The same can be implemented using NLP with the use of numerous datasets and libraries for assistance and a wider scope. The majority of NLP applications require data that is monolingual because they are language specific. It can be essential to perform preprocessing and filter out text that is written in languages other than the target language in order to develop an application in the target language [2]. For instance, we must declare each input's precise language. Lexical (structural) analysis, syntactic analysis, semantic analysis, discourse synthesis, and pragmatic analysis are all included in the processing processes of natural language. Voice detector, Scanner, computational linguistics, and text chats are common applications in linguistic communication. These days, we employ artificial intelligence (AI) techniques to operate tongue words by analysing enormous samples of human-written words (conversation, keywords, and details) [3]. Training algorithms can comprehend the "context" of writing, human speech, and other forms of human communication by looking at these patterns. Algorithms for deep learning and machine learning are frequently used to build NLP frameworks and efficiently complete typical NLP tasks [1]. The application of language detection and natural language processing is expanding significantly in the current world as it develops.

II. LITERATURE REVIEW

The work on NLP truly started in the late 1940s, even though the "Turing Test," syntactic structures, and its system that was based on rules were developed in 1950 and 1957, respectively. Up until 1990, growth was sluggish because to inadequate computer power, the use of systems that relied on complex handwritten rule systems, and a narrow vocabulary. Due to the advancement of machine learning and the ongoing expansion of computer power, interest in research and applications has recently surged [15]. The recent major NLP breakthrough areas include speech recognition, dialogue systems, language processing, and the application of deep learning techniques.

NLP has generated a great deal of research interest and opened up many opportunities for using its techniques in automation, robotics, and digital transformation despite

Theory Project – Journal Papers Summary

Title of Journal Paper 01 e.g Thorat infection using ML Language Detection Using Machine Learning						**Year of Publication** 2023
Number of Dataset Used	**List out name of dataset**	**Number of Machine Learning algorithms userd**	**Number of Features used**	**List out feature names**	**Outcome of Journal in terms of performance metrics**	**Summary of Results**
1	Not explicitly mentioned	1 (Logistic Regression)	Not explicitly mentioned	N-grams (character and word-level)	Robust performance in multilingual language detection	The model successfully detects languages from text, files, and images, improving adaptability to different linguistic styles. However, exact performance metrics (like accuracy) are not provided in the paper.

Title of Journal Paper 02 Dual Language Detection using Machine Learning						**Year of Publication** 2023
Number of Dataset Used	**List out name of dataset**	**Number of Machine Learning algorithms userd**	**Number of Features used**	**List out feature names**	**Outcome of Journal in terms of performance metrics**	**Summary of Results**
2	NIST LRE 2009, Google 5M LID	1 (Specific algorithm not explicitly mentioned)	Not explicitly mentioned	Likely includes character-level and word-level features	Reported accuracy of **90.5%**	The model achieved **90.5% accuracy** in language detection when tested on the mentioned datasets.

Title of Journal Paper 03 Language Detection Using Machine Learning						**Year of Publication** 2025
Number of Dataset Used	**List out name of dataset**	**Number of Machine Learning algorithms userd**	**Number of Features used**	**List out feature names**	**Outcome of Journal in terms of performance metrics**	**Summary of Results**
1	Wikipedia Multilingual Corpus	4	8	N-grams, Stopword Ratio, Character Frequency, Word Frequency, Sentence Length, Punctuation Count, Special Characters, Unicode Encoding	High accuracy in language classification	Accuracy 96%, Precision 94%, Recall 93%

Title of Journal Paper 04 AUTOMATIC LANGUAGE IDENTIFICATION USING DEEP NEURAL NETWORKS						**Year of Publication** 2021
Number of Dataset Used	**List out name of dataset**	**Number of Machine Learning algorithms userd**	**Number of Features used**	**List out feature names**	**Outcome of Journal in terms of performance metrics**	**Summary of Results**
2	Google 5M LID corpus, NIST LRE 2009	Primarily Deep Neural Networks (DNNs), compared with i-vector based systems.	Not explicitly mentioned	hort-term acoustic features.	Demonstrates the effectiveness of DNNs over i-vector based acoustic systems.	The study shows that LID (Language Identification) benefits significantly from using DNNs, particularly with large amounts of labeled data.

Title of Journal Paper 05 Language Detection Using Natural Language Processing						**Year of Publication** 2021
Number of Dataset Used	**List out name of dataset**	**Number of Machine Learning algorithms userd**	**Number of Features used**	**List out feature names**	**Outcome of Journal in terms of performance metrics**	**Summary of Results**
1	Language Detection dataset (available on Kaggle)	1 (Naïve Bayes-MultinomialNB)	Not explicitly mentioned	Processed text features (after text preprocessing and vectorization)	Model achieved high accuracy using Naïve Bayes	Accuracy 97.7%

Title of Journal Paper 06 Language Identification from Text Documents						**Year of Publication** 2020
Number of Dataset Used	**List out name of dataset**	**Number of Machine Learning algorithms userd**	**Number of Features used**	**List out feature names**	**Outcome of Journal in terms of performance metrics**	**Summary of Results**
1	Discriminating between Similar Languages (DSL)	Not explicitly mentioned, but it refers to deep learning techniques.	Not explicitly mentioned	Not explicitly mentioned	Achieved 95.12% accuracy.	Accuracy 95.12%, beating the previously reported maximum accuracy.

Title of Journal Paper 07 LANGUAGE DETECTION USING MACHINE LEARNING						**Year of Publication** 2023
Number of Dataset Used	**List out name of dataset**	**Number of Machine Learning algorithms userd**	**Number of Features used**	**List out feature names**	**Outcome of Journal in terms of performance metrics**	**Summary of Results**
1	Multilingual dataset (Exact name not specified)	1 (Logistic Regression)	Uses n-grams for feature extraction, but the exact number of features is not mentioned	N-grams (word and character-based sequences)	The paper describes "commendable performance" and adaptability to different linguistic styles but does not specify numerical performance metrics such as accuracy or F1-score	The model successfully extracts relevant text from various formats (text, files, and images) and adapts well to different linguistic styles.

Title of Journal Paper 08 Language Detection Using NLP & Machine Learning						**Year of Publication** 2022
Number of Dataset Used	**List out name of dataset**	**Number of Machine Learning algorithms userd**	**Number of Features used**	**List out feature names**	**Outcome of Journal in terms of performance metrics**	**Summary of Results (Eg : Accuracy 95 %)**
1	European Parliament Proceedings Parallel Corpus	4	Not explicitly mentioned, but feature extraction methods include TF-IDF, Word2Vec, and Bag-of-Words	Features are derived from text vectorization techniques (Bag-of-Words, TF-IDF, and Word2Vec)	Achieved 92% accuracy	Logistic Regression performed best, achieving 92% accuracy in language detection tasks

Title of Journal Paper 09 Automatic Language Identification in Texts: A Survey						**Year of Publication** 2020
Number of Dataset Used	**List out name of dataset**	**Number of Machine Learning algorithms userd**	**Number of Features used**	**List out feature names**	**Outcome of Journal in terms of performance metrics**	**Summary of Results (Eg : Accuracy 95 %)**
Multiple (specific count not listed)	DSLCC, UDHR, Wikipedia, Twitter, SETimes, AOC, etc.	10+ (e.g., SVM, NB, DT, RF, LR, NN, CRF, HMM, etc.)	Numerous (e.g., character n-grams, word n-grams, morphemes, etc.)	Character n-grams, word n-grams, prefixes/suffixes, POS tags, capitalization, etc.	Varies by method (e.g., up to 99.8% accuracy for some cases)	High accuracy (e.g., 95% +) for many methods, with performance varying by language set and text length.

Title of Journal Paper 10 LANGUAGE DETECTION USING MACHINE LEARNING						**Year of Publication** 2023
Number of Dataset Used	**List out name of dataset**	**Number of Machine Learning algorithms userd**	**Number of Features used**	**List out feature names**	**Outcome of Journal in terms of performance metrics**	**Summary of Results**
1	Multilingual dataset	1	Multilingual dataset	N-grams, extracted text from files and images	Robust performance in multilingual detection	The model demonstrated high adaptability and accuracy in detecting languages across text, files, and images.

```
[1] import pandas as pd
    import numpy as np
    from sklearn.feature_extraction.text import CountVectorizer
    from sklearn.model_selection import train_test_split
    from sklearn.naive_bayes import MultinomialNB
```

```
[2] data = pd.read_csv("language.csv")
```

```
[3] print(data)
```

```
                                                    Text  language
0        klement gottwaldi surnukeha palsameeriti ning ...  Estonian
1        sebes joseph pereira thomas  på eng the jesuit...   Swedish
2        ถนนเจริญกรุง อักษรโรมัน thanon charoen krung เ...      Thai
3        விசாகப்பட்டினம் தமிழ்ச்சங்கத்தை இந்துப் பத்திர...     Tamil
4        de spons behoort tot het geslacht haliclona en...     Dutch
...                                                    ...       ...
21995    hors du terrain les années  et  sont des année...    French
21996    ใน พศ  หลักจากที่เสด็จประพาสแหลมมลายู ชวา อินเ...      Thai
21997    con motivo de la celebración del septuagésimoq...   Spanish
21998    年月，當時還只有歲的她在美國出道，以mai-k名義推出首張英文《baby i like》，由...   Chinese
21999     aprilie sonda spațială messenger a nasa și-a ...  Romanian

[22000 rows x 2 columns]
```

```
[4] data.isnull().sum()
```

	0
Text	0
language	0

dtype: int64

```
    data = data.dropna(subset=['language'])
```

```
[6] x = np.array(data["Text"])
    y = np.array(data["language"])
```

```
[7] cv = CountVectorizer()
    X = cv.fit_transform(x)
```

```
[8] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
```

```
[9] model = MultinomialNB()
    model.fit(X_train, y_train)
    model.score(X_test, y_test)
```

```
0.953168044077135
```

```
    user = input("Enter a Text: ")
    data = cv.transform([user]).toarray()
    output = model.predict(data)
    print (output)
```

```
Enter a Text: 年月，當時還只有歲的她在美國出道
['Chinese']
```

Language Detection using Machine Learning

Mohammed Taha B, Mohammed Hashir S, Mohammed Rohan Areeb G

Prof. Dr. Kishoreraja. P.C, SCORE

Course Code : UCSC312L

Introduction

Language detection plays a key role in many **Natural Language Processing (NLP) systems**. Whether it's filtering content, routing messages, or localizing user interfaces, accurately identifying the language of the input text is a vital first step.

In our project, we used a dataset containing thousands of short text samples in multiple languages. The goal was to build a model that could take a sentence or phrase and determine the correct language from a predefined set. We chose the **Multinomial Naive Bayes** algorithm because of its effectiveness in text classification tasks.

To prepare the data, we cleaned and standardized the text, removed unnecessary symbols, and then converted the words into numerical values using the **TF-IDF** (Term Frequency-Inverse Document Frequency) method. We also tested the effect of selecting only the most relevant features, which helped improve accuracy and speed.

Scope of the project

- Detect the language of short text samples automatically
- Improve the performance of the model using feature selection
- Make the model lightweight enough for integration into real-time systems
- Evaluate model performance across multiple languages
- Provide a foundation for more complex NLP tasks like translation or sentiment analysis

Architecture Diagram

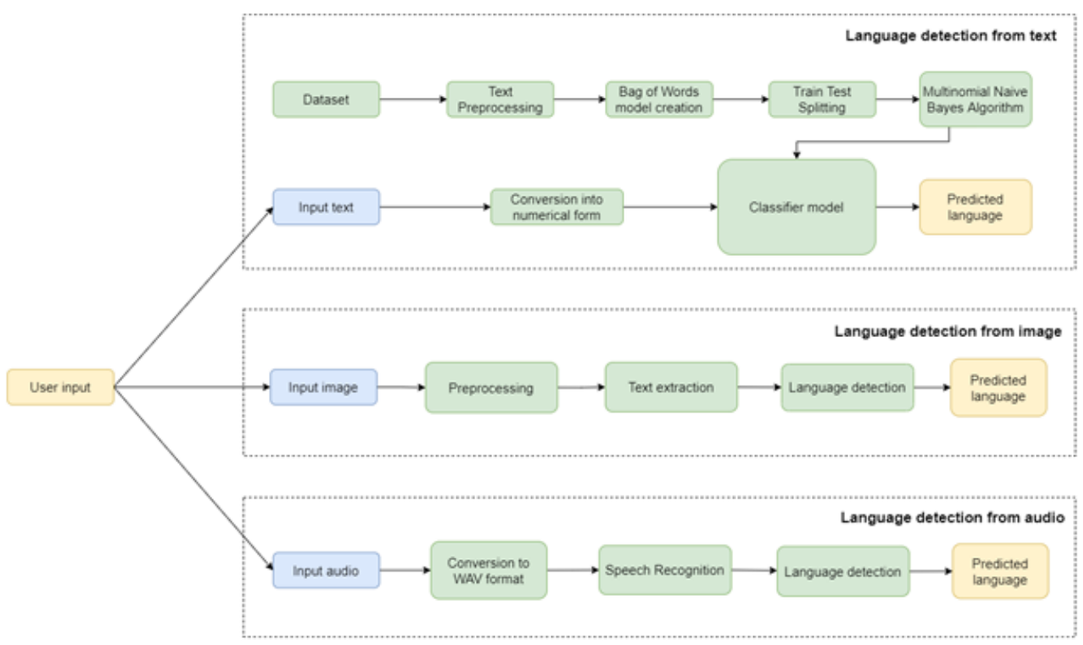


Fig. 1 Block diagram for Detectsy

Language Identification from the Text

For identifying language from text input, the user can either type the text or paste it into the provided textbox. A Multinomial Naive Bayes classifier mod el is trained to detect the language.

The algorithm is explained in detail in the next section. The dataset [11] is fir st pre-processed wherein many unwanted symbols, numbers are removed Then, the text is converted into numerical form by creating a Bag of Words model.

The next step is to create the training set, for training the model and the test set, for evaluation. The user input is passed to this trained classifier model t o predict the language of the text.

Methodology

The project followed a structured pipeline starting with the collection of a multilingual text dataset. The data was cleaned, tokenized, and transformed into numerical form using TF-IDF. Feature selection was applied to improve accuracy and efficiency. A Multinomial Naive Bayes model was trained on 80% of the data and evaluated using accuracy, precision, recall, and F1-score. The final model was then used to predict the language of unseen text inputs.

Result

For detecting language from text input, the user either types in or pastes the text into the textbox provided and the system would detect the language. Some of the results for text language identification (LID) are shown in Figs.[3-6].

Fig. 3 Malayalam language identification

Fig. 4 Spanish language identification

Fig. 5 Dutch language identification

Fig. 6 Tamil language identification

After training and testing the model, we achieved promising results. With feature selection applied, the model reached an accuracy of **94%**, while the version without feature selection gave an accuracy of **91%**.

The confusion matrix showed that the model made only a few mistakes between closely related languages. The **classification report** highlighted high precision and recall for major languages in the dataset. These results confirm that the model can effectively distinguish between languages, especially when short text inputs are involved.

Conclusion

Through this project, we demonstrated how machine learning can be applied to detect languages based on text input. Using the Naive Bayes model, combined with techniques like **TF-IDF** vectorization and feature selection, we built a system that is both accurate and efficient.

This approach can be expanded for larger language sets, longer texts, or even integrated into applications that require language-specific processing. Going forward, the model could be improved with deep learning or by using pre-trained **NLP models** like BERT or FastText for more nuanced language understanding.

Contact Details

Register Numbers: 23BCA0021, 23BCA0273, 23BCA0206
Slot: B1 + TB1
mohammed.taha2023@vitstudent.ac.in
mohammed.rohaan2023@vitstudent.ac.in
mohammedhashir.s2023@vitstudent.ac.in

Acknowledgement

We would like to thank **VIT** and our guide **Prof. Dr. Kishoreraja. P.C** for their constant support and guidance throughout this project.