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

increasinggly interconnected, the demand for effective multilingual communication tools has grown, making language identification more important than ever.In multilingual machine translation systems, the ability to correctly identify the source language of an input text is paramount. Incorrect language identification can lead to inappropriate translation models being applied, resulting in poor translation quality and potential miscommunication. This is particularly challenging in scenarios involving code-switching, where multiple languages are used within the same text, and in cases where low-resource languages are involved.The complexity of language identification arises from the vast diversity of languages and dialects, each with unique linguistic features. Traditional rule-based and statistical methods for language identification have shown limitations in handling such diversity, especially when dealing with short or noisy text data. Recent advancements in machine learning and deep learning offer promising solutions by leveraging large-scale multilingual datasets and sophisticated feature extraction techniques.This paper presents an advanced approach to language identification that integrates seamlessly with multilingual machine translation systems. Our method utilizes state-of-the-art machine learning algorithms to accurately detect the language of a given text, even in challenging conditions. By incorporating various linguistic features and training on extensive multilingual corpora, our model achieves high accuracy in language identification, which in turn enhances the overall performance of machine translation systems.

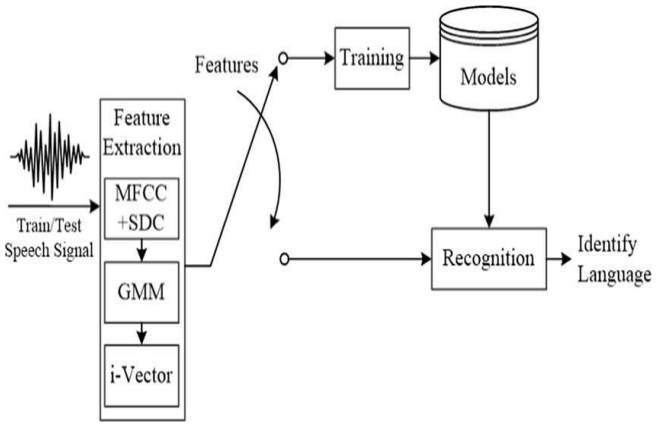


Fig1 : Block Diagram

II.EXISTING SYSTEM

Existing language identification systems for multilingual machine translation often rely on traditional rule-based or statistical methods, which have significant limitations. Rule-based approaches use predefined linguistic rules to identify languages, but they struggle with the diversity of language structures and can be cumbersome to maintain and update. Statistical

methods, which analyze character and word frequencies, offer some improvements but often fall short when dealing with short texts, noisy data, or mixed-language inputs.Moreover, many existing systems are not well-equipped to handle the complexities of code-switching, where multiple languages are interspersed within the same text, or to accurately identify low-resource languages due to insufficient training data. These limitations lead to frequent misidentifications, which can significantly degrade the quality of translations by applying incorrect translation models. Consequently, there is a growing need for more sophisticated and accurate language identification systems that can seamlessly integrate with modern multilingual machine translation frameworks to enhance translation quality and reliability.

DRAW BACKS

The existing systems for language identification in multilingual machine translation exhibit several drawbacks that hinder their effectiveness and accuracy:

1. **Limited Handling of Code-Switching:** Many existing systems struggle with code-switching scenarios where multiple languages are used within the same text. This leads to incorrect language identification and subsequently poor translation quality.
2. **Inaccuracy with Short Texts:** Traditional methods, especially statistical ones, often rely on the analysis of character and word frequencies. These methods are less effective for short texts, where insufficient data leads to inaccurate language identification.

III.PROPOSED SYSTEM

The proposed system for language identification in multilingual machine translation addresses the limitations of existing methods by employing advanced machine learning techniques and leveraging large-scale multilingual datasets. This system uses deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to automatically detect the language of a given text with high accuracy. These models are trained on diverse and extensive multilingual corpora, allowing them to capture complex linguistic patterns and nuances.One of the key features of the proposed system is its ability to handle

code-switching and mixed-language inputs effectively. By analyzing contextual information and linguistic features, the system can accurately identify languages even in challenging scenarios. Additionally, the proposed system includes robust preprocessing steps to handle noisy data, ensuring reliable performance across various text types and contexts.The integration of the proposed language identification system with multilingual machine translation frameworks is seamless, allowing for real-time language detection and appropriate model selection for translation tasks. This enhances the overall translation quality by ensuring that the correct translation models are applied based on the accurately identified source language.

ADVANTAGES

The proposed system for language identification in multilingual machine translation offers several significant advantages over existing methods:

1. **High Accuracy:** By leveraging advanced machine learning techniques and deep learning models, the proposed system achieves high accuracy in language identification, even with short or noisy text inputs.
2. **Effective Code-Switching Handling:** The system is designed to accurately identify languages in texts that involve code-switching or mixed-language inputs, ensuring appropriate translation models are applied.

IV.MODULES

To implement this project we have designed following modules

1) Upload Language Dataset: using this module we will upload dataset and then remove all missing and special symbols from dataset
2) Pre-process Dataset: using this module we will convert above process dataset into numeric vector by employing 3 NGRAMS technique and then convert entire text data into numeric vector and then split training data into train and test where application using 80% dataset for training and 20% for testing
3) Train KNN Algorithm: 80% training data will be input to KNN algorithm to train a model and this model will be applied on

20% test data to calculate prediction accuracy

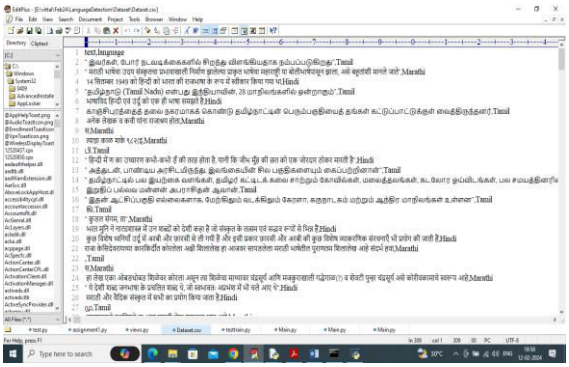
4) Train SVM Algorithm: 80% training data will be input to SVM algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
5) Train Random Forest Algorithm: 80% training data will be input to Random Forest algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
6) Comparison Graph: will plot comparison between all algorithms
7) Language Detection & Translation: here user can enter some text line and then application will predict language name and then translate that language into English using Google Translator.

In this project we have employed NGRAM and Machine learning algorithms to identify language names from given text. To evaluate performance we have utilized various machine learning algorithms such as SVM, KNN and Random Forest. Each algorithm performance is tested in terms of accuracy, precision, recall, Confusion matrix graph and FSCORE. Among all algorithms Random Forest is giving high accuracy.

To train above algorithms we have used dataset of languages such as Tamil, Hindi and Marathi and this dataset can be downloaded from below URL

https://www.kaggle.com/datasets/sandeepbelamagi/indian-local-languages

In below screen we are showing dataset details



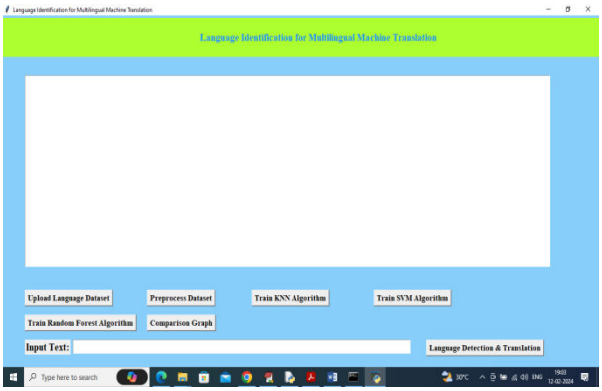
In above dataset first row contains dataset column names and remaining rows contains Text sentences and language names and by using above dataset we will train and test each algorithm performance.

To implement this project we have designed following modules

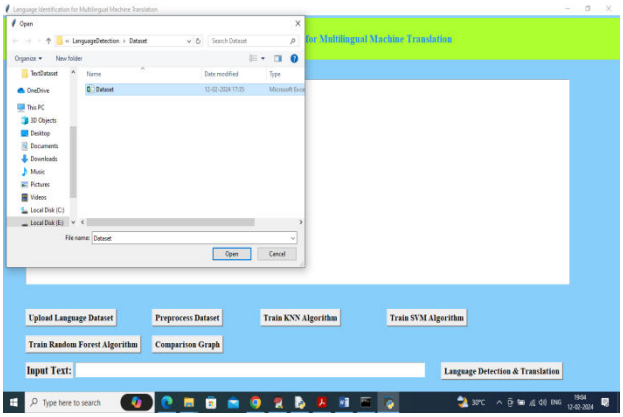
8) Upload Language Dataset: using this module we will upload dataset and then remove all missing and special symbols from dataset

9) Pre-process Dataset: using this module we will convert above process dataset into numeric vector by employing 3 NGRAMS technique and then convert entire text data into numeric vector and then split training data into train and test where application using 80% dataset for training and 20% for testing
10) Train KNN Algorithm: 80% training data will be input to KNN algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
11) Train SVM Algorithm: 80% training data will be input to SVM algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
12) Train Random Forest Algorithm: 80% training data will be input to Random Forest algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
13) Comparison Graph: will plot comparison between all algorithms
14) Language Detection & Translation: here user can enter some text line and then application will predict language name and then translate that language into English using Google Translator.

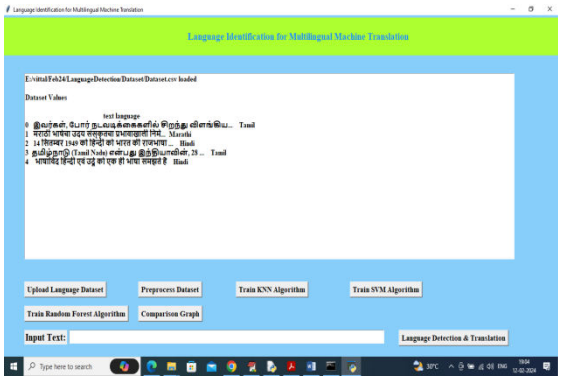
To run project double click on run.bat file to get below screen



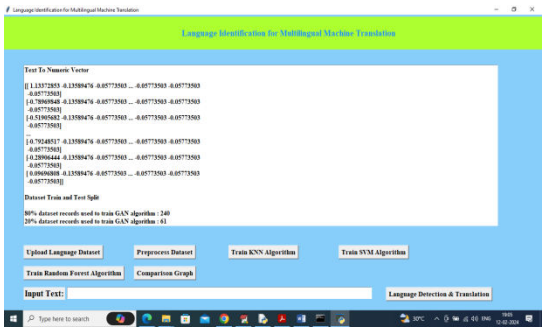
In above screen click on 'Upload Language Dataset' to load dataset and get below output



In above screen selecting and uploading dataset file and then click on 'Open' button to load dataset and get below page

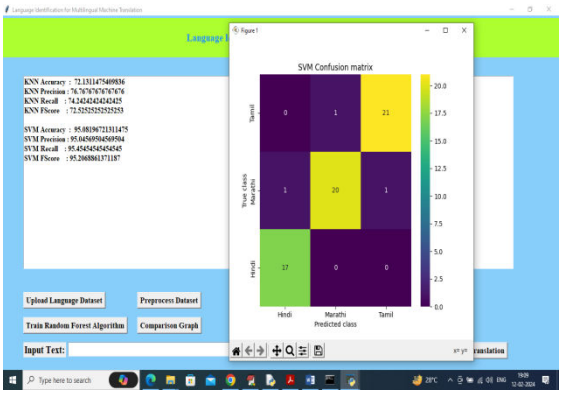


In above screen dataset loaded and now click on 'Pre-process Dataset' button to clean dataset and get below output

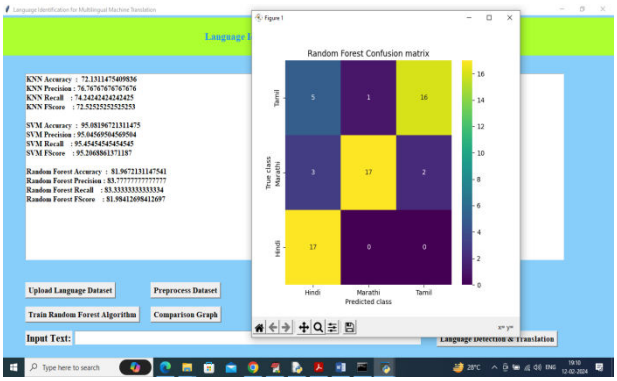


In above screen entire text data converted to numeric vector by using 3 NGRAM techniques and then can see train and test split details and now click on 'Run KNN Algorithm' to train KNN and get below output

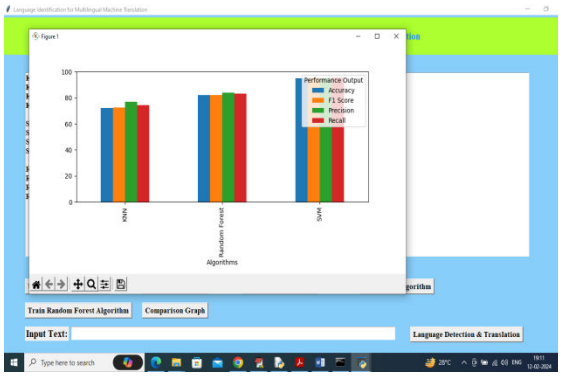
In above screen KNN training completed and it got accuracy as 72% and can see other metrics also and in confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels and all yellow and green colour boxes in diagnol represents correct prediction count and remaining blue boxes represents incorrect prediction count and now close above graph and then click on 'Train SVM' button to get below output



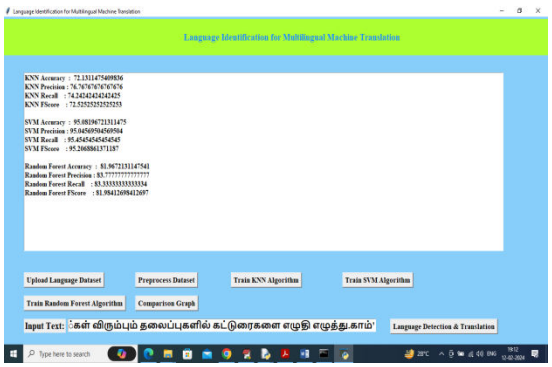
In above screen SVM got 95% accuracy and can see other metrics also and now click on 'Train Random Forest' to get below output



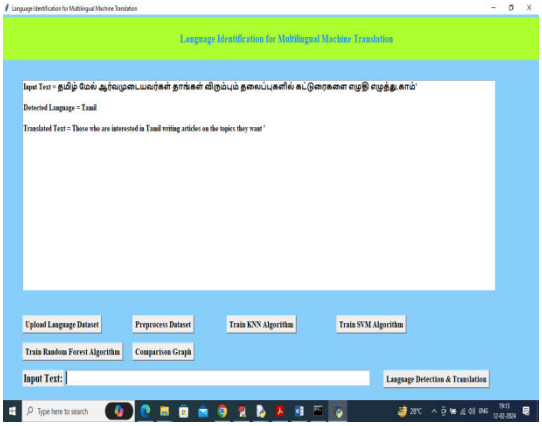
In above screen Random Forest got 81% accuracy and now click on Comparison Graph button to get below output



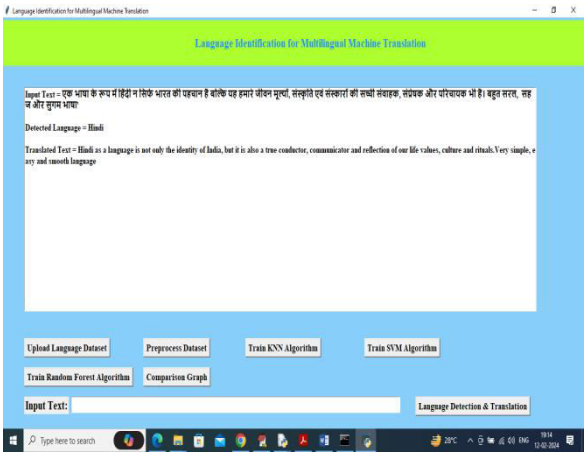
In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars and in all algorithms SVM got high accuracy and now enter some sentence in text field and then press 'Language Detection and Translation' button



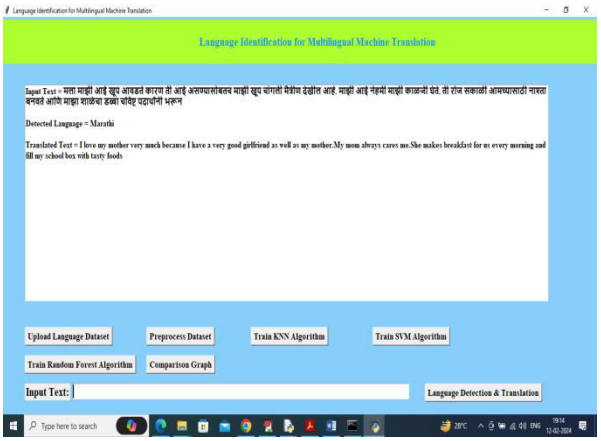
In above screen I entered some text in text field and then press Language Detect button to get below output



In above screen in text area can see Detected Language is Tamil and can see Translated text in English and below is another example



In above screen detected language is Hindi with translation



In above screen detected language is Marathi with English translation.

Similarly enter sentence in text field and get detected language and translation

V.CONCLUSION

In conclusion, language identification plays a critical role in the effectiveness of multilingual machine translation systems. Accurate language identification ensures that text is correctly classified into the appropriate language before translation, which is essential for the downstream translation tasks to function correctly. This study highlights the importance of leveraging advanced techniques such as deep learning and neural networks for improving the accuracy and robustness of language identification systems.The integration of sophisticated models, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers, has proven to enhance language identification performance, especially in handling short texts and code-switching scenarios. These models benefit from their ability to capture contextual information and manage a large variety of linguistic patterns, leading to more reliable language detection.However, challenges remain, including the handling of low-resource languages, dialects, and mixed-language inputs. Addressing these issues requires continuous advancements in model architectures, the inclusion of diverse and extensive training datasets, and the development of more effective pre-processing techniques.

VI.REFERENCES

- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). "Bag of Tricks for Efficient Text Classification." *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427-431.
- Discusses efficient text classification techniques, which are foundational for language identification tasks.
- Lui, M., & Baldwin, T. (2012). "langid.py: An Off-the-shelf Language Identification Tool." *Proceedings of the ACL 2012 System Demonstrations*, 25-30.
- Introduces langid.py, a widely used language identification tool, and its applications.
- Kocmi, T., & Bojar, O. (2017). "LanideNN: Multilingual Language Identification on Character

Window." *arXiv preprint arXiv:1701.03338*.

- Proposes a neural network approach for multilingual language identification using character windows.

- King, B., & Abney, S. (2013). "Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods." *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1110-1119.

- Discusses methods for language identification in mixed-language documents, addressing the challenge of code-switching.

- Jurgens, D., & Klapaftis, I. (2013). "Semeval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses." *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, 290-299.

- Provides insights into semantic evaluation tasks, relevant for understanding context in language identification.

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.

- Introduces BERT, a state-of-the-art model for language understanding, which can be adapted for language identification.

- Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2018). "Unsupervised Machine Translation Using Monolingual Corpora Only." *International Conference on Learning Representations*.

- Explores unsupervised methods for machine translation, highlighting techniques that can be applied to multilingual systems.

- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). "Improving Language Understanding by Generative Pre-Training." *OpenAI Technical Report*.

- Discusses generative pre-training models, such as GPT, which are

relevant for improving language identification.

- Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
- Presents XGBoost, a machine learning algorithm that can be used for feature-based language identification.
- Platanios, E. A., Chris Dyer, C., & Neubig, G. (2018). "Contextual Parameter Generation for Universal Neural Machine Translation." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 425-435.
- Describes methods for parameter generation in multilingual neural machine translation systems.
- Ling, W., Dyer, C., Black, A. W., & Trancoso, I. (2015). "Two/Too Simple Adaptations of Word2Vec for Syntax Problems." *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*