www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 02, February 2025, pp : 460-462

e-ISSN : 2583-1062
Impact Factor : 7.001

LANGUAGE IDENTIFICATION FROM TEXT

Poojashree Chandrashekar[1]

[1,2]Dept of Mathematical Science Stevens Institute of Technology Hoboken, New Jersey.

poojash@stevens.edu

ABSTRACT

This paper presents a comprehensive evaluation of several machine learning models designed to enhance language identification tasks. We compare traditional and advanced models including Naive Bayes, Bi-LSTM, CNN, and BERT, using met- rics such as accuracy, precision, recall, and F1-score to assess their performance. The study reveals that the BERT model, leveraging its transformer-based architecture and self-attention mechanisms, significantly outperforms others by achieving an exceptional accuracy of 99.92%. The robustness of BERT in handling complex linguistic features such as mixed and short text sequences, and its efficacy in processing code-mixed texts and phonetically similar languages, highlight its potential for advanced natural language processing tasks. These findings not only contribute to the theoretical advancements in machine learning but also offer practical insights for implementing effective language identification systems.

1. INTRODUCTION

With the rapid growth of digital content and global com- munication, multilingualism has become an integral aspect of many online platforms, requiring effective methods for automatic language detection. Accurate language identification is critical in numerous applications, such as content modera- tion, language-based routing, and machine translation. Without this capability, systems may misinterpret content, leading to inefficient processes and miscommunication.

This project aims to address the challenge of automatic language identification from text by developing a machine learning model capable of accurately predicting the language of a given text. The problem is particularly important in scenar- ios where short or multilingual text snippets are common, such as social media posts, chat applications, and user-generated content across different platforms. To accomplish this, we will utilize multilingual datasets such as Europarl, which contain a rich variety of text samples in different languages. These datasets will allow us to train and test models across various linguistic contexts.

The research will explore a range of machine learning and deep learning approaches, be- ginning with traditional models like Naive Bayes and progressing to more advanced models, including Bi-LSTM, CNN, and BERT, a state-of- the-art transformer-based model fine-tuned for multilingual tasks.Through this project, we seek to improve the accu- racy of language detec- tion systems and tackle complex scenarios such as distinguishing between phonetically similar languages or handling mixed-language text. The results will have practical implications for enhancing content management systems, translation services, and multilingual communication tools, contributing to the broader goal of enabling seamless interaction across languages in the digital age.

2. DATA COLLECTION AND PROCESSING

A. Data Source

The dataset used for this project is the Europarl-parallel- corpus-dataset available on Kaggle: https://www.kaggle.com/ datasets/dionafegnem/europarl-parallel-corpus-10962106.

These datasets provide a diverse collection of sentences in multiple languages, allowing for comprehensive model training and testing across different linguistic setups. The folder consists of 19 csv files with English sentences and its translations into Bulgarian, Italian, and Spanish text pairs for multilingual training.

B. Data Processing

- Python's Pandas library has been used to load the dataset already downloaded from the repository for this project. It is also used for various data manipulation tasks in our project.
- Regular expressions were applied to remove non- alphanumeric characters while retaining punctuation es- sential for contextual analysis. All text was converted to lowercase to ensure uniformity and reduce redundancy in text representation.
- The pandas.fillna() method was employed to handle miss- ing text entries. Empty text fields were replaced with empty strings, maintaining dataset completeness without introducing bias.
- A custom length-based filter was implemented to exclude short texts (less than three characters), ensuring only meaningful content was used for model training.
- TF-IDF vectorization was implemented using scikit- learn's TfidfVectorizer, converting textual data into nu-

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 02, February 2025, pp : 460-462
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

merical features. To improve efficiency during training, only the top 1000 features were retained, based on TF- IDF scores.

3. MODEL DEVELOPMENT

A. Machine Learning Model Considered

In this project, we aim to develop a robust system for language identification by exploring a variety of machine learning and deep learning models. The models selected for this task each offer distinct advantages based on their ability to handle various aspects of text, ranging from simple word fre- quencies to more complex sequential patterns and contextual embeddings.

- **Naive Bayes:** We begin with a basic model utilizing Naive Bayes for language classifi- cation. Naive Bayes is a probabilistic classifier based on Bayes' theorem, which assumes independence between features. This model is commonly used in text classification tasks because of its simplicity and effectiveness, particularly when dealing with word frequencies. By leveraging term frequency- inverse document frequency (TF-IDF) representations, Naive Bayes can provide a solid baseline for language identification. Despite its simplicity, it can achieve rea- sonable performance when text length is sufficient, and the differences in word usage between languages are distinct
- **Bi-LSTM (Bidirectional Long Short-Term Memory):** To improve upon the base- line Naive Bayes model, we implement a Bi-LSTM model. Bi-LSTM is a type of recurrent neural network (RNN) that captures both forward and backward dependencies in a sequence. This bi-directional approach allows the model to effectively learn from the context surrounding each word, making it especially suited for sequential data such as text. By capturing bi-directional word sequences, Bi-LSTM can learn more complex patterns in language structure, which are essential for identifying languages, especially when short text sequences or similar languages are involved.

3. Convolutional Neural Networks (CNNs): We further enhance our language detec- tion system by introducing a CNN model. While CNNs are traditionally used in image recognition tasks, they have shown great promise in text classification, particularly for de- tecting patterns in character sequences. CNNs excel in tasks involving local dependencies, such as identifying n-gram-like fea- tures, which can be crucial for distinguishing between languages that share similar alphabets or word structures. By applying convolutional filters to the character level, the model can learn patterns that differentiate languages at a more granular level.

- **BERT(Bidirectional Encoder Representations from Transformers):** Lastly, we fine-tune BERT, a pre-trained transformer model, specifically for multilingual text clas- sification tasks. BERT is a state-of-the-art model designed to capture deep contextual relationships in text by using self-attention mechanisms. It has been pre-trained on a large corpus in multiple languages, making it particularly suitable for multilingual tasks. By fine-tuning BERT on our language identification dataset, we leverage its pow- erful con- textual embedding capabilities, enabling it to understand the nuances of each language beyond simple word frequencies or character patterns. This approach is expected to yield the best performance, especially in handling complex or ambiguous text scenarios.

B. Chosen Model Architecture

After evaluating multiple machine learning and deep learn- ing models, the architecture that showed the best performance for language identification is BERT (Bidirectional Encoder Representations from Transformers). BERT was chosen due to its ability to handle complex text sequences and its proven efficiency in multilingual tasks.

BERT's transformer-based architecture utilizes self- attention mechanisms, allowing it to capture deep contextual relationships in text. This architecture is particularly advantageous for tasks involving language detection, as it can understand the contextual meaning of words in a sequence, rather than relying on just their individual occurrences or local patterns.

1) Key Components of the BERT Model Architecture:

- **Transformer Encoder:** The core of BERT consists of multiple transformer encoder layers, which allow the model to focus on different parts of a sentence simultane- ously. This is crucial for tasks like language identification, where the context and structure of the sentence play a significant role in determining the language.
- **Multilingual Pre-Training:** BERT has been pre-trained on large corpora across multiple languages, making it inherently multilingual. This enables the model to handle a wide range of languages without needing to train separate models for each language.
- **Fine-Tuning for Classification:** For our task, the pre- trained BERT model is fine- tuned specifically for multi- lingual text classification. By adjusting the model weights
- on our language identification dataset, BERT can learn the specific characteristics of different languages in the dataset.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
www.ijprems.com
editor@ijprems.com
Vol. 05, Issue 02, February 2025, pp : 460-462
e-ISSN : 2583-1062
Impact Factor : 7.001

- **Self-Attention Mechanism:** BERT's self-attention mech- anism enables it to capture long-range dependencies in text, making it well-suited for detecting subtle language patterns that other models might miss.

2) Final Model Selection

After comparing the performance of various models like Naive Bayes, Bi-LSTM, CNN, and BERT, the fine-tuned BERT model demonstrated superior results in terms of ac- curacy, precision, recall, and F1-score. Its ability to handle both word-level and sentence-level nuances, along with its multilingual capabilities, makes BERT the most suitable model architecture for this language identification task. The fine-

tuned BERT model not only provided the highest accuracy but also handled challenging cases, such as distinguishing between phonetically similar languages and han- dling code-mixed text (i.e., sentences containing multiple languages). Thus, BERT was selected as the final model architecture for the language identification system.

4. RESULT AND ANALYSIS

The performance of the models was evaluated using ac- curacy, precision, recall, and F1-score. The results for each model are as follows:

TABLE 1- PERFORMANCE OF VARIOUS MODELS

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	56%	0.76	0.55	0.45
Bi-LSTM	99.81%	0.99	0.99	0.99
CNN	96.50%	0.97	0.96	0.96
BERT	99.92%	0.99	0.99	0.99

BERT significantly outperformed all other models, demon- strating exceptional performance with an accuracy of 99.92%. Its ability to handle mixed and short text sequences proved highly advantageous, especially in challenging scenarios such as code-mixed text or phonetically similar languages.

5. CONCLUSION

- BERT demonstrated superior performance compared to other models, including Naive Bayes, Bi-LSTM, and CNN, particularly in handling multilingual text and mixed-language data. Its ability to effectively capture contextual relationships using self-attention makes it an optimal choice for this task
- BERT achieved an impressive accuracy of 99.92%, under- scoring its ability to understand complex linguistic pat- terns and provide reliable language identification across diverse scenarios.
- The models were trained on the Europarl parallel corpus, which facilitated effective language identification across a wide range of languages, including English, Spanish, Bulgarian, and Italian. This dataset provided a robust foundation for developing and testing the models.
- The fine-tuned BERT model is highly applicable to real- world use cases such as content moderation, machine translation, and multilingual communication tools. Its versatility and accuracy make it a valuable asset for enhancing user experiences in global, multilingual envi- ronments.

6. REFERENCES

[1] S. Bo, Y. Zhang, J. Huang, S. Liu, Z. Chen and Z. Li, "At- tention Mechanism and Context Modeling System for Text Min- ing Machine Translation," 2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS), Hangzhou, China, 2024, pp. 857-863, doi: 10.1109/DOCS63458.2024.10704434.

[2] keywords: Training; Text mining; Attention mechanisms; Computational modeling; Semantics;Clustering algorithms; Transformers; Machine trans- lation; Standards; Context modeling; Transformer; Text mining; Machine translation; K-Means,

[3] E. Boonchieng, W. Nadda, W. Liawrungrueang and W. Boonchieng, "Enhancing Dis- ease Symptom Analysis in Thai Text: Methods for Text Oversampling in Imbalanced Data for Disease Detection," 2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI), San Jose, CA, USA, 2024, pp. 302-307, doi: 10.1109/IRI62200.2024.00068.

[4] keywords: Accuracy; Migraine; Influenza; Machine learning; Natural language processing; Data models; Complexity theory; Rough surfaces; Hemorrhaging; Medical diagnostic imaging; Machine Learning; Text Classification; Oversampling; Digital Disease Detection; Imbalanced Data Problem,