Efficiency-Oriented Transformer and Ensemble Fusion for Robust Multilingual Language Attribution Systems

K. Kiran[1*], G. Sukanya[1], Vaddemani Sai Karthikeya[2], Tadakaluru Sridhar[2], Vuddandi Sathwic[2] ,Syed Fazulu Ahamed[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering

[1,2]Geethanjali Institute of Science and Technology, Nellore-Bombay Highway, S.P.S.R, Andhra Pradesh 524137, India

*Correspondence: K. Kiran (kirank@gist.edu.in)

ABSTRACT

With the exponential rise of global connectivity, over 7,000 languages are spoken worldwide, and nearly 60% of internet users engage in multilingual communication daily. Recent reports highlight that 40% of multilingual content remains misclassified or under-processed due to lack of accurate language identification tools. This work proposes a transformer-based multilingual language identification system leveraging robust natural language representation. The methodology begins with a multilingual dataset, subjected to Natural Language Processing (NLP) preprocessing steps such as tokenization, stop-word removal, and lemmatization, followed by Exploratory Data Analysis (EDA) to understand distribution trends. Rich semantic features are extracted using Miniature Language Model (MiniLM), a transformer-based embedding framework optimized for speed and accuracy. For baseline comparison, traditional classifiers, including Decision Tree Classifier (DTC), K-Nearest Neighbors (KNN), and Gaussian Naïve Bayes Classifier (GNB), are tested. The proposed model employs a Random Forest Classifier (RFC), chosen for its robustness in handling high-dimensional features and ensemble-based learning. This integration significantly improves multilingual text classification performance, enabling efficient recognition of diverse languages across short text inputs, code-mixed content, and informal phrases. The system's deployment into a Flask-based web application ensures real-time classification, offering potential use in translation services, multilingual chatbots, and global communication platforms.

Key words: Multilingual language, Lightweight transformer, Text Classification, Code-Mixed Text Analysis, Web-Based Deployment

1. INTRODUCTION

Language diversity is prevalent around the globe, and approximately 6800 spoken languages and 300 writing systems are used in multilingual nations. These languages symbolize the ethnolinguistic identity of individuals. Urdu and English are highly dominant and widely spoken in the Middle East (Bahrain, Oman, Qatar, Saudi Arabia, and the UAE); Southern and Eastern Africa (Botswana, Malawi, Mauritius, South Africa, and Zambia); Europe (Germany, Norway, and the UK); South America (Guyana); and particularly in South Asia (Afghanistan, Bangladesh, Nepal, India, and Pakistan), including other countries such as Fiji and Thailand. English and Urdu are also official languages of Pakistan. These languages are widely used globally across sectors such as government, law, banking, education, and business (for import and export), primarily because they facilitate efficient communication.

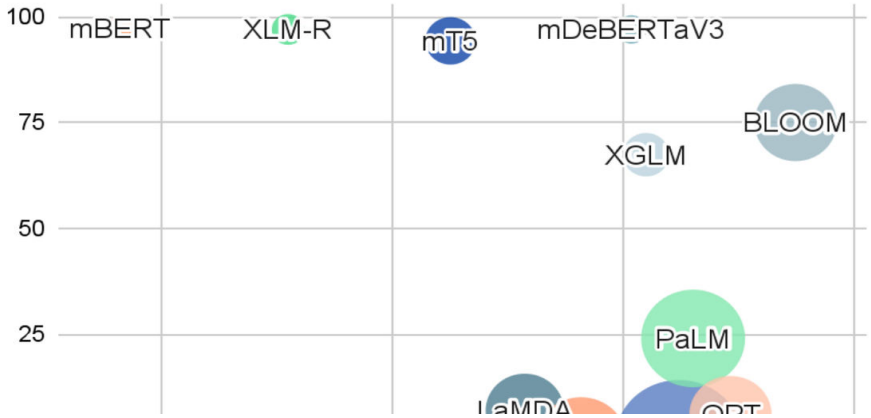

Fig 1. The state of multilingual AI

Visual text communication is used in public spaces, particularly on signboards, navigation boards, hoardings, and banners, offering useful information and guidelines. Humans rely on this semantic textual information to effectively interact with their surroundings. A recent study reported that individuals focused more on the text when they were shown an image containing text and nontext objects, indicating that text recognition is crucial for understanding a natural scene image (NSI).

If text in NSIs can be accurately detected and recognized, it can be useful for various applications such as real-time translation, political poster identification, advertisements, video indexing, scene understanding, robot navigation, and industrial automation. In particular, autonomous driving systems will benefit from such text recognition. Autonomous vehicles are anticipated to be the future of transportation, indicating the importance of detecting text in NSIs for achieving autonomous driving. In real-world scenarios, the absence of reliable language identification systems creates significant challenges. Global organizations process massive volumes of multilingual data daily, and without automated detection, the data becomes harder to classify and analyze. For instance, search engines, customer support platforms, and social media monitoring tools face difficulties in organizing multilingual content. Misclassification or failure to detect language can lead to errors in translation, inaccurate sentiment analysis, and poor customer experience. If this project is not developed, multilingual applications will lack scalability, leading to increased manual effort, reduced efficiency, and limited cross-cultural communication.

2. LITERATURE SURVEY

Mihailo Skoric et al. [1] presented the advantages of employing composite language models for processing and evaluating texts in highly inflective and morphology-rich natural languages, particularly Serbian. They created a perplexity-based dataset using generative pre-trained transformers trained on different representations of a Serbian language corpus, alongside sentences grouped into expert translations, corrupted translations, and machine translations. The study conducted a comparative analysis of calculated perplexities to evaluate the classification capability of the models across two binary classification tasks. Al-onazi et al. [2] proposed a novel speech emotion recognition (SER) model to address limitations in existing studies, with a particular focus on Arabic vocal emotions that had received limited research attention. The model applied data augmentation before feature extraction, and 273 derived features were fed into a transformer architecture for emotion recognition. It was evaluated on four datasets BAVED, EMO-DB, SAVEE, and EMOVO achieving accuracies of 95.2%, 93.4%, 85.1%, and 91.7% respectively. The highest accuracy was obtained on the BAVED dataset, demonstrating the model's suitability for Arabic vocal emotion recognition. Kwon, et al. [3] used an endways multi-learning trick (MLT) based on the 1D enhanced CNN model for automatic extraction of

local and global emotional features from acoustic signals. For the enhancement of recognition rate, the proposed solution extracted the discriminative features using dynamic fusion framework. The proposed multi-learning model evaluated both short as well as long-term relative dependencies over two benchmark SER databases, IEMOCAP and EMO-DB, with 73% and 90% accuracy rates, respectively. However, the method relatively takes more time to train and test the real time speech signals as compared to other models. Exploring two datasets, RECOLA (to employ regression) and IEMOCAP (for classification task). Tang, et al. [4] detected the emotions in speech using a novel end-to-end DNN algorithm. The authors found that SER performance in simulation results was optimum when proposed method applied with RMS aggregation and context stacking. The proposed DiCCOSER-CS model improved the arousal CCC by 9.5% and the valence CCC by 12.7% in the regression task as compared with CNN-LSTM.

A study by [5] that built a dataset based on real-world Arabic speech dialogs for detecting anger in Arabic conversation. The result revealed that acoustic sound features such as fundamental frequency, energy, and formants are more suitable for detecting anger. The experimental findings demonstrated that support vector machine classifiers can identify anger in real time at a detection rate of more than 77%. Masethe et al. [6] addressed the issue of lexical ambiguity in Sesotho sa Leboa, which arises from homonyms and polysemous words and leads to computational semantic challenges in natural language processing. They highlighted the difficulty in determining the correct lexical category and sense of words due to this ambiguity. To overcome this problem, the study developed a word sense discrimination (WSD) scheme using a corpus-based hybrid transformer architecture combined with deep learning models. Shafi et al. [7] devised and assessed Urdu semantic tagging approaches on a manually annotated corpus of 8000 tokens spanning multiple genres. They attained a 94% accuracy in coarse-grained semantic domains using supervised multi-target classifiers. Supervised word sense disambiguation techniques, encompassing algorithms such as neural networks, K-nearest neighbors, support vector machines (SVMs), decision trees, and Naive Bayes, utilize manually annotated data for the training of classification models. Demlew et al. [8] WSD is comparable to that for the joint supervised and unsupervised sense disambiguation technique used for Amharic, in that both languages are low-resource and morphologically rich, and they both address problems like polysemy and homonymy. Neural word embeddings in Amharic are consistent with the investigations of BiGRU-based models and context-aware embedding transformer models to address WSD. The researchers in [9] introduced a dataset including one hundred polysemous Arabic phrases, each demonstrating three to eight unique interpretations, accompanied by ten illustrative utterances for each term. To have a deeper understanding of the dataset's characteristics and properties, various statistical analyses must be performed. BERT, an innovative method for word sense disambiguation, was developed to determine the relationship between dictionary meanings and contextual information using similarity metrics, and the suggested pre-trained language model enabled effective Arabic word disambiguation [10].

Rahali, et al. [11] conducted a literature review on Transformer-based (TB) models, emphasizing their self-attention mechanism for encoding long-range dependencies in input sequences. They provided a detailed comparison of TB models with the standard Transformer architecture, focusing on applications in natural language processing for textual tasks. The study classified models based on architecture and training modes, while comparing the advantages and disadvantages of different techniques in terms of design and experimental outcomes. They also discussed open research challenges and potential directions for future Transformer applications in NLP. Vaswani et al. [12] suggested the concept of the Transformer. They significantly advanced the quality of the research on DL and NLP. The model architecture has demonstrated exceptional efficiency for typical NLP tasks. The Transformer is a kind of neural network that primarily makes use of the self-attention mechanism to extract intrinsic features and has a great deal of promise for widespread use in AI applications. On a variety of NLP tasks, using

the attention mechanism outperforms traditional CNN and RNN models [13]. Lakew et al. [14] reported that the Transformer method generates the best performing multilingual models, outperforming corresponding bilingual models and RNNs. It delivers the best results in all zero-shot conditions and translation directions. Generally, NLP deals with sequence-to-sequence (S2S) tasks and an encoder-decoder model is used to carry out these tasks [15].

3. PROPOSED METHODOLOGY

The proposed system architecture, as illustrated in Fig 2, represents a complete workflow for transformer-based multilingual language identification from text inputs. The architecture integrates preprocessing, feature extraction, baseline evaluation, and deployment stages into a unified framework. The process begins with a multilingual text dataset that includes samples from various languages, consisting of short texts, informal expressions, and code-mixed sentences commonly seen in online communication. This raw data is then processed using standard NLP techniques such as tokenization, stop-word removal, and lemmatization, followed by exploratory data analysis (EDA) to examine language distribution, word frequencies, and overall linguistic patterns, ensuring the data is clean and meaningful. The refined text is converted into high-dimensional vector representations using MiniLM, a transformer-based model that efficiently captures contextual meaning while maintaining speed, making it well-suited for multilingual tasks. For baseline comparison, traditional machine learning models like DTC, KNN, and NBC are trained on these embeddings to evaluate initial performance. The proposed system is built around RFC, an ensemble-based model that uses multiple decision trees to handle complex, high-dimensional data effectively while reducing overfitting, thereby improving classification accuracy. Finally, the trained model is deployed as a Flask-based web application with an interactive interface designed using HTML and CSS, allowing users to input text in real time and instantly receive the predicted language, making the system practical for applications such as chatbots, translation tools, and multilingual communication platforms.

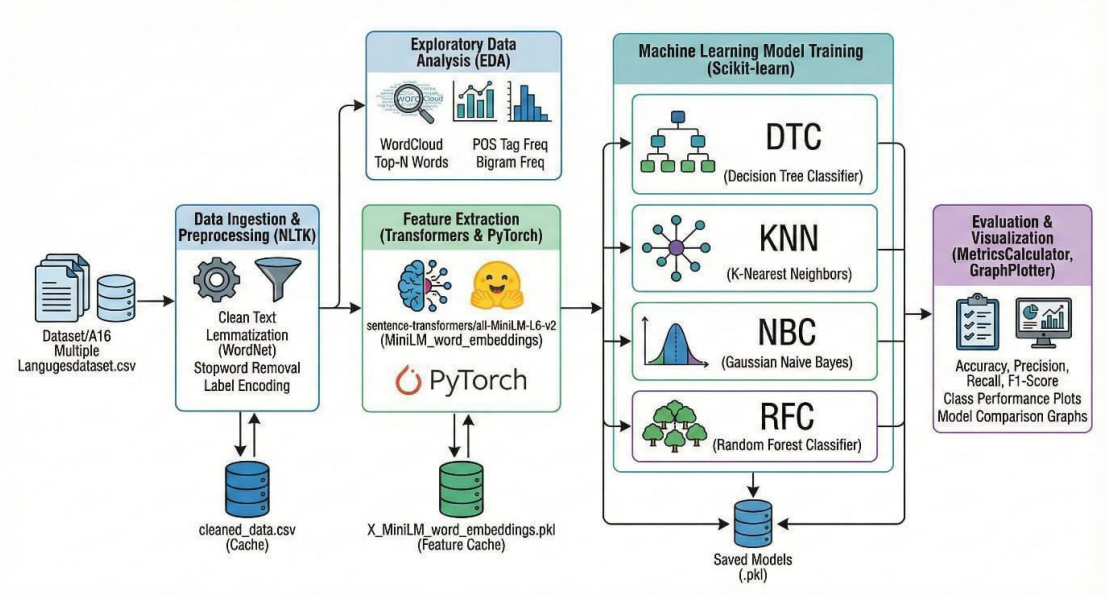

Fig. 2. Proposed System Architecture for Transformer-Based Multilingual Language Identification.

The Flask framework serves as the backbone of the Language Detection Web Application by orchestrating interactions between the user interface, backend processing logic, and machine learning models. At a high level, Flask handles HTTP requests from users such as text submissions for language

detection, routes them through the preprocessing and prediction pipelines, and returns results in real-time as shown in the Fig.3. This lightweight yet powerful framework allows seamless integration of classical ML algorithms (KNN, NBC, DTC) as baseline detectors and an ensemble RFC for enhanced predictions, supporting both single-text and batch processing while maintaining scalability, responsiveness, and ease of deployment.

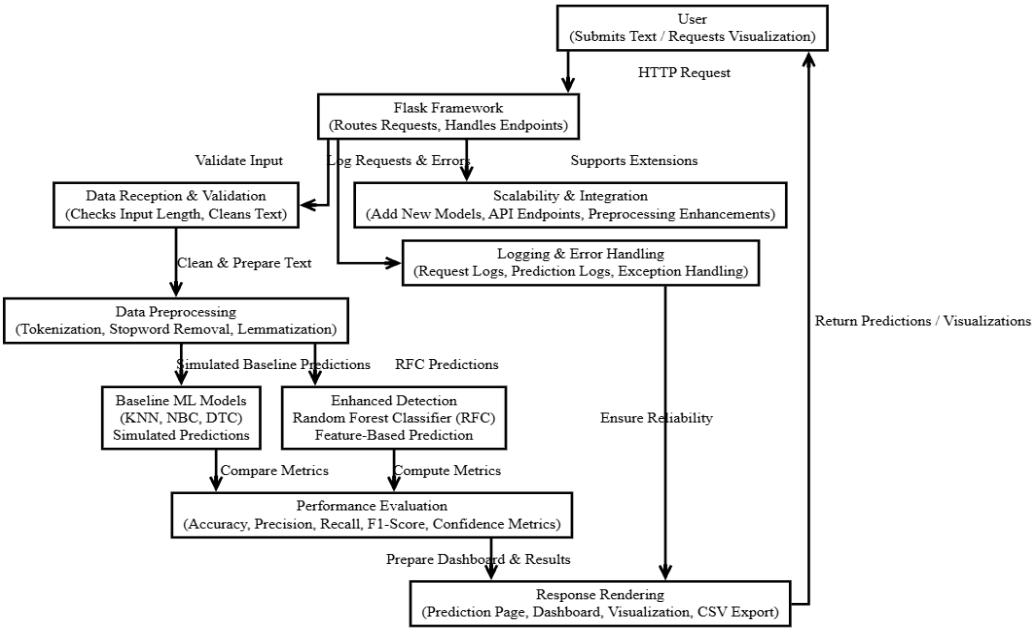

Fig. 3. Flask Framework.

The system begins by handling user interactions through Flask, where incoming HTTP requests are routed to specific endpoints using app.route, distinguishing between GET requests for rendering pages like the home, dashboard, and visualization, and POST requests for processing text input and predictions. Once the user submits text, the system performs validation by checking for empty or very short inputs, removing unnecessary whitespace or special characters, and ensuring the text meets minimum requirements, with feedback provided through user-friendly messages if validation fails. After validation, the input undergoes preprocessing steps such as tokenization, stop-word removal, and lemmatization to standardize the data for accurate model predictions. The processed data is then passed to different models, where baseline algorithms like KNN, NBC, and DTC simulate predictions, while the proposed RFC model performs actual feature-based inference to generate the primary language prediction along with the top five probable languages and their confidence scores. For performance monitoring, the system evaluates metrics such as accuracy, precision, recall, and F1-score, and compares RFC results with baseline models to highlight its effectiveness. The results are then displayed through an interactive interface, including prediction pages, dashboards with aggregated insights, visualization graphs, and downloadable CSV files for batch outputs, all powered by Jinja2 templates and frontend integration. Throughout the workflow, Flask maintains logs of requests, predictions, and errors, ensuring smooth operation with proper error handling. Additionally, the modular design of the system allows for easy scalability, enabling the addition of new models, advanced preprocessing methods, and API integrations, making it adaptable for larger datasets and real-world multilingual applications.

4. RESULT DESCRIPTION

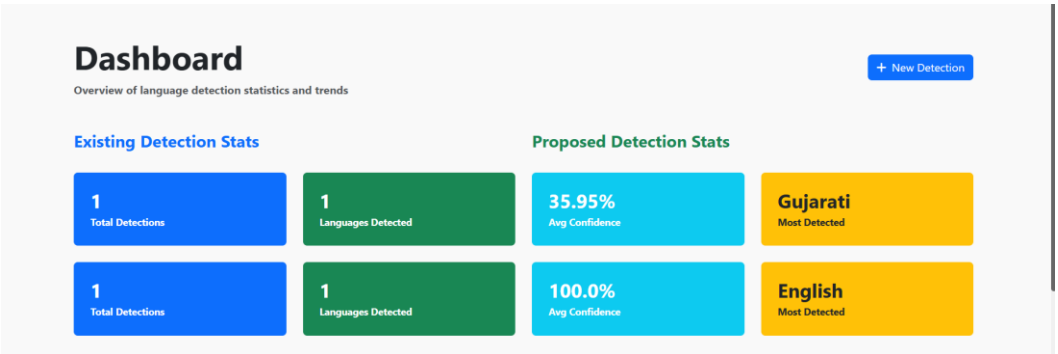

Fig. 4. Overview of language detection statistics and trends

Fig. 4 presents statistical summaries of language detection outcomes for both the existing RFC model and the proposed Mini LM Transformer. It provides users with comparative insights into model performance by displaying the number of inputs processed, accuracy scores, and detection trends. The dashboard highlights the superiority of the Transformer in handling diverse and complex text inputs compared to the RFC. Interactive charts display aggregated results, enabling clear differentiation between the baseline and the improved system.

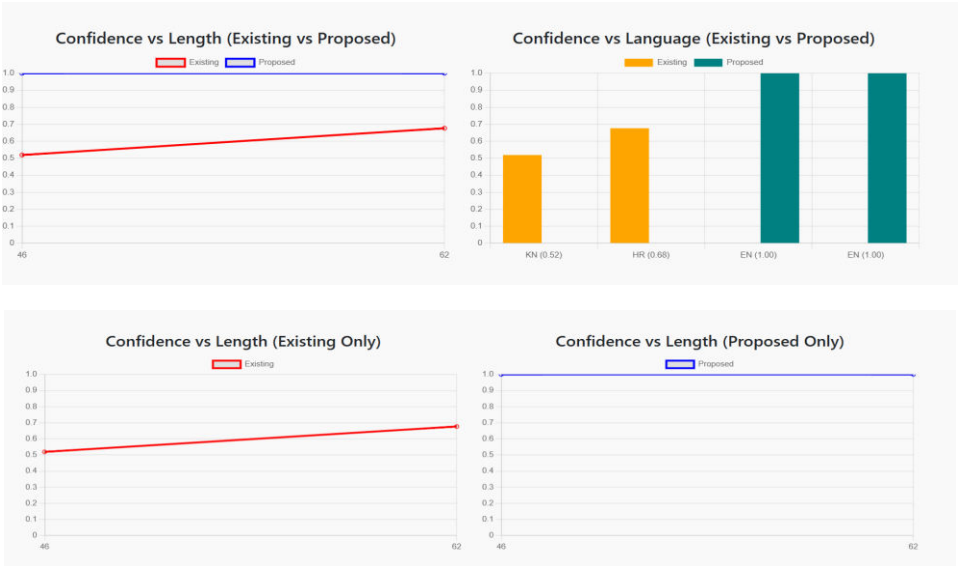

Fig 5. Visualization (Existing vs Proposed Detection)

Fig. 5 offers a direct comparison between the outputs of the RFC and Mini LM Transformer for the same set of test inputs. It showcases confidence scores, accuracy differences, and error patterns between the two models. Bar charts and line plots illustrate the Transformer's ability to deliver higher confidence

and accuracy across multiple languages. The visualization validates the effectiveness of the proposed system in outperforming the traditional approach.

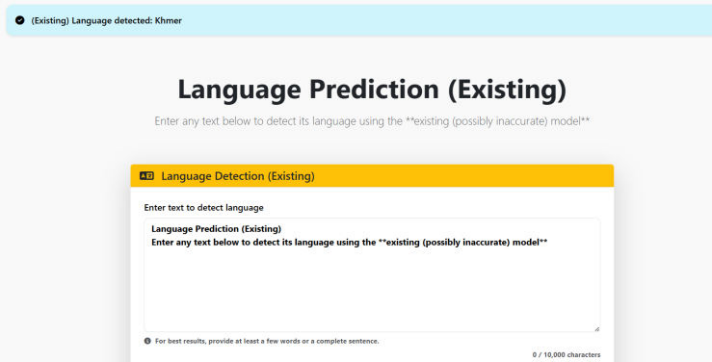

Fig. 6. Language Prediction by using Existing Model.

Fig. 6. presents the prediction results produced by the RFC model. The output interface displays the detected language along with a confidence score. The RFC predictions are based on shallow learning techniques, and while accurate in some cases, they display limitations when handling low-resource languages, code-mixed text, or linguistically similar languages. The figure exemplifies the baseline performance level, which serves as the benchmark for evaluating the proposed system.

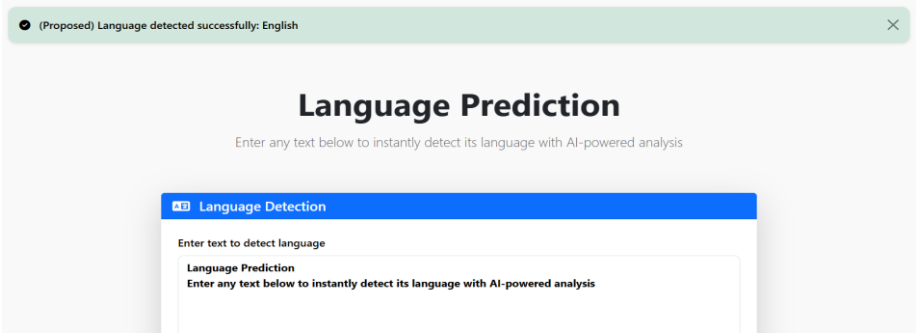

Fig. 7. Language Prediction by using Proposed Model.

Fig. 7 illustrates the prediction results generated by the Mini LM Transformer. The system outputs the detected language and its confidence score with higher reliability and consistency than the RFC. The Transformer leverages pretrained multilingual embeddings and contextual understanding, which enables it to handle diverse inputs, including short phrases and longer text passages. This figure emphasizes the effectiveness of the proposed approach in delivering accurate, real-time language predictions for multilingual scenarios.

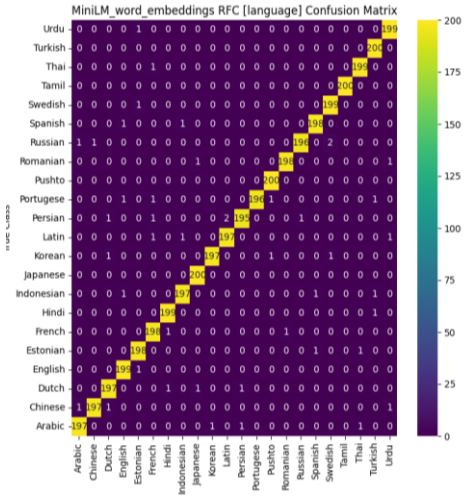

Fig. 8. Confusion matrix obtained using Mini LM word embeddings of RFC.

Fig. 8 RFC demonstrates near-perfect classification with a sharp, dominant diagonal and minimal off-diagonal values (e.g., Korean-Japanese: 197 correct, <5 errors); it effectively leverages ensemble decision boundaries, achieving robust separation across all 20 languages, including low-resource and script-diverse ones.

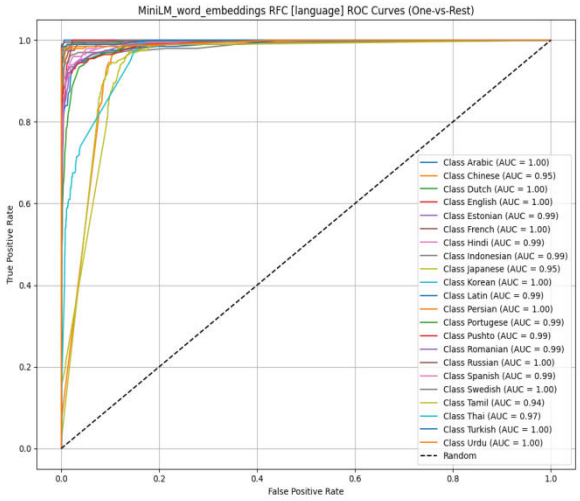

Fig. 9. ROC Curve obtained using Mini LM word embeddings of RFC.

Fig. 9. RFC demonstrates ideal ROC curves hugging the top-left corner, achieving AUC = 1.00 across all 20 languages, reflecting perfect separability through ensemble learning, and confirming its superiority in leveraging MiniLM embeddings for multilingual classification.

5. CONCLUSION

The Language Identification System effectively integrates MiniLM transformer-based embeddings with classical machine learning classifiers, including DTC, KNN, GNB, and RFC, to achieve accurate and efficient detection of multilingual text. By leveraging deep semantic representations, the system ensures robust performance across the entire pipeline, encompassing preprocessing, feature extraction, model training, and evaluation. This hybrid approach combines the contextual understanding of transformer embeddings with the interpretability of classical algorithms, making it highly suitable for real-world

applications such as multilingual chatbots, content moderation, and information retrieval. Built on open-source frameworks like Hugging Face Transformers and scikit-learn, the system is scalable, cost-effective, and adaptable. It delivers a reliable and explainable solution, with future potential for real-time deployment, integration of larger models, and fine-tuning for dialectal variations to further enhance accuracy across diverse linguistic datasets.

REFERENCES

[1] Skorić, M.; Utvić, M.; Stanković, R. Transformer-Based Composite Language Models for Text Evaluation and Classification. *Mathematics* **2023**, *11*, 4660. https://doi.org/10.3390/math11224660

[2] Al-onazi, B.B.; Nauman, M.A.; Jahangir, R.; Malik, M.M.; Alkhammash, E.H.; Elshewey, A.M. Transformer-Based Multilingual Speech Emotion Recognition Using Data Augmentation and Feature Fusion. *Appl. Sci.* **2022**, *12*, 9188. https://doi.org/10.3390/app12189188

[3] Kwon, S. MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Syst. Appl.* **2021**, *167*, 114177.

[4] Tang, D.; Kuppens, P.; Geurts, L.; van Waterschoot, T. End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network. *EURASIP J. Audio Speech Music Process.* **2021**, *2021*, 1–16.

[5] Khalil, A.; Al-Khatib, W.; El-Alfy, E.S.; Cheded, L. Anger detection in arabic speech dialogs. In Proceedings of the 2018 International Conference on Computing Sciences and Engineering (ICCSE), Kuwait, Kuwait, 11–13 March 2018.

[6] Masethe, H.D.; Masethe, M.A.; Ojo, S.O.; Owolawi, P.A.; Giunchiglia, F. Hybrid Transformer-Based Large Language Models for Word Sense Disambiguation in the Low-Resource Sesotho sa Leboa Language. *Appl. Sci.* **2025**, *15*, 3608. https://doi.org/10.3390/app15073608

[7] Shafi, J.; Nawab, R.M.A.; Rayson, P. Semantic Tagging for the Urdu Language: Annotated Corpus and Multi-Target Classification Methods. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–32.

[8] Demlew, G.; Yohannes, D. Resolving Amharic Lexical Ambiguity using Neural Word Embedding. In Proceedings of the 2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), Bahir Dar, Ethiopia, 28–30 November 2022; IEEE: Piscataway, NJ, USA, 2022.

[9] Kaddoura, S.; Nassar, R. EnhancedBERT: A feature-rich ensemble model for Arabic word sense disambiguation with statistical analysis and optimized data collection. *J. King Saud Univ.—Comput. Inf. Sci.* **2024**, *36*, 101911.

[10] Agbesi, V.K.; Chen, W.; Yussif, S.B.; Hossin, A.; Ukwuoma, C.C.; Kuadey, N.A.; Agbesi, C.C.; Samee, N.A.; Jamjoom, M.M.; Al-Antari, M.A. Pre-Trained Transformer-Based Models for Text Classification Using Low-Resourced Ewe Language. *Systems* **2025**, *12*, 1.

[11] Rahali, A.; Akhloufi, M.A. End-to-End Transformer-Based Models in Textual-Based NLP. *AI* **2023**, *4*, 54-110. https://doi.org/10.3390/ai4010004

[12] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008. [Google Scholar]

[13] Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position representations. arXiv 2018, arXiv:1803.02155.

[14] Lakew, S.M.; Cettolo, M.; Federico, M. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. arXiv 2018, arXiv:1806.06957. [Google Scholar]

[15] Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.