

International Journal of Engineering Research and Science & Technology



www.ijerst.org

ISSN : 2319-5991

Vol.22 No. 2(1) (2026)



ijerst.editor@gmail.com
editor@ijerst.com

Research Paper

A Neuro-Semantic Fusion Architecture for Fine-Grained Multilingual Language Discrimination Using Contextual Embedding Manifolds

Musham Swetha¹, Yelagandhula Sai Kiran², Kolipaka Rishik², Banoth Swarnalatha², Jillala Praveen²

¹Assistant Professor, ²UG Student, ^{1,2}Department of Computer Science and Engineering

^{1,2}Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India.

ABSTRACT

With the rapid expansion of global connectivity, thousands of languages are actively used across digital platforms, and a significant portion of online users communicate in multiple languages on a daily basis. However, a large amount of multilingual content remains inaccurately classified due to the limitations of existing language identification techniques. Traditional manual methods are time-consuming and prone to errors, especially when dealing with short, informal, or mixed-language text. In addition, conventional algorithms often lack the ability to capture deeper semantic and contextual relationships within multilingual data. This study presents a transformer-based multilingual language identification framework designed to address these challenges through advanced Natural Language Processing (NLP) techniques. The process begins with a multilingual dataset that undergoes preprocessing steps such as tokenization, stopword removal, and lemmatization, followed by Exploratory Data Analysis (EDA) to understand patterns and distributions within the data. Semantic features are then extracted using Miniature Language Model (MiniLM), a lightweight transformer model capable of generating meaningful contextual embeddings. These features are further utilized by machine learning classifiers including Decision Tree Classifier (DT), K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB) Classifier, and Random Forest Classifier (RF) for classification tasks. The use of RF as a primary model enhances the system's ability to handle high-dimensional feature spaces and improves overall classification performance. The integration of transformer-based embeddings with classical machine learning models enables effective handling of short text inputs, informal expressions, and multilingual variations. The framework is implemented within a Flask-based web application to support real-time classification and user interaction.

Keywords: Multilingual Language Identification, Natural Language Processing, Transformer-Based Embeddings, MiniLM, Semantic Feature Extraction, Multilingual Data Processing, Text Classification, Short Text Analysis.

1. INTRODUCTION

Language diversity is a defining characteristic of human society, with thousands of spoken languages and hundreds of writing systems used across different regions of the world. As shown as figure 1 these languages reflect the cultural and ethnolinguistic identity of individuals and communities. Among them, Urdu and English hold significant global importance and are widely spoken across regions such as the Middle East, Africa, Europe, South America, and South Asia. In countries like Pakistan, both languages serve as official means of communication and are extensively used in domains such as government, education, business, banking, and law. Their widespread adoption highlights their role in enabling effective communication across diverse populations and professional sectors.

Visual text communication plays an essential role in everyday life, particularly in public environments where information is conveyed through signboards, navigation displays, banners, and advertisements. People depend on such textual cues to interpret their surroundings and make informed decisions. Studies have shown that individuals tend to focus more on textual elements when viewing images containing both text and non-text components, emphasizing the importance of text recognition in understanding natural scene images. This indicates that extracting textual information from visual content is a critical aspect of human interaction with real-world environments. Accurate detection and recognition of text within natural scene images can support a wide range of practical applications.



Figure 1: The state of multilingual AI

These include real-time language translation, identification of public and political information, advertisement analysis, video indexing, and enhanced scene understanding. Furthermore, such capabilities are highly valuable in advanced technologies like robotics and industrial automation. In particular, autonomous driving systems can greatly benefit from reliable text recognition, as vehicles must interpret road signs and environmental information to navigate safely. As autonomous transportation continues to evolve, the ability to detect and understand textual information in real-world scenes becomes increasingly important.

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 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 NLP. 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 NLP 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 are used to carry out these tasks [15].

3. PROPOSED SYSTEM

The proposed methodology establishes a structured analytical framework for processing and classifying multilingual textual data using NLP and machine learning techniques. The analytical pipeline begins with dataset acquisition and organization, followed by text preprocessing and semantic feature extraction. Raw textual data is cleaned and normalized to remove noise, inconsistencies, and irrelevant elements. A transformer-based embedding model, MiniLM, is utilized to extract meaningful contextual representations from the text, capturing semantic relationships and linguistic patterns. These extracted feature vectors are then analysed using multiple ML classifiers such as DT, KNN, GNB, and RF to perform language classification. An additional analytical component is incorporated to validate and interpret textual inputs before classification. A graphical interface enables user interaction for dataset handling, preprocessing, feature extraction, model training, performance visualization, and prediction tasks, as shown in figure 2. A lightweight storage mechanism is used to manage trained models and processed data, while a server-based architecture enables real-time text analysis and prediction. Continuous evaluation and retraining further enhance analytical accuracy and adaptability to new textual data.

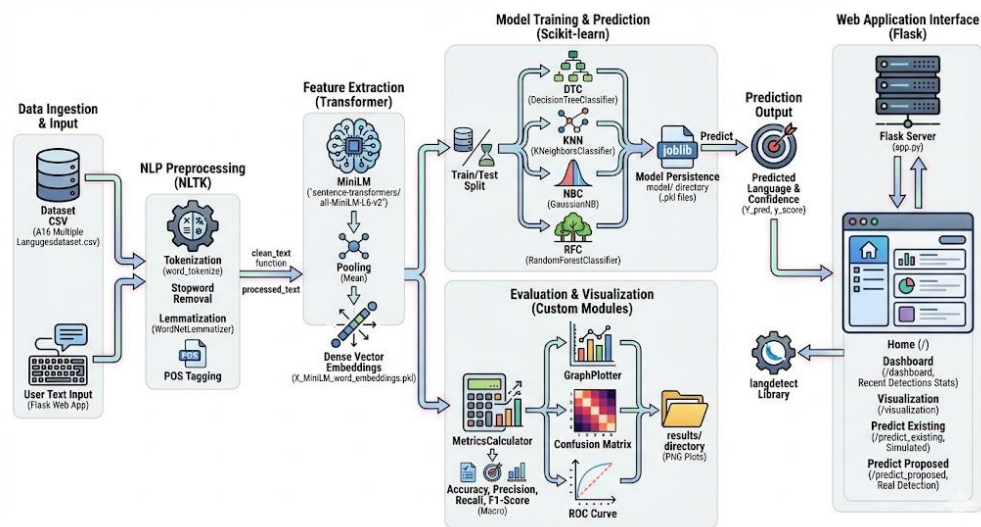


Figure 2: Proposed system architecture for transformer-based multilingual language identification.

User Interface (Client Application)

- The user interacts with the system through a graphical interface developed using a desktop-based or web-based environment.
- The interface supports operations such as login, dataset upload, preprocessing, feature extraction, model training, performance comparison, and text prediction.
- Users can input text manually or upload text files for analysis.

- All user inputs are processed within the interface and forwarded to backend analytical modules.

Flask Application Server

- The Flask server acts as the central communication component of the system.
- It receives textual input requests from client systems and routes them through the processing pipeline.
- The server manages preprocessing, feature extraction, model prediction, and response generation.
- It enables remote users to submit text inputs and receive classification results efficiently.

Lightweight Database (Authentication & Storage)

- The system uses a lightweight storage mechanism to manage user authentication and model-related data.
- It stores user credentials, processed datasets, and trained model files.
- The storage system interacts with the application layer for login verification and data retrieval.
- Efficient data handling ensures quick access and reliable persistence.

Text Dataset (Multilingual Data Collection)

- The dataset serves as the primary input source for the analytical pipeline.
- It consists of textual data collected from multiple languages and sources.
- The data includes sentences, phrases, and documents representing various linguistic patterns.
- The dataset is used for both training and evaluation of classification models.

Text Preprocessing and Feature Extraction

- The raw text undergoes preprocessing steps including lowercasing, tokenization, stopword removal, and lemmatization.
- Noise such as special characters and irrelevant tokens is removed to improve data quality.
- A **MiniLM transformer model** is used to extract deep semantic embeddings from the processed text.
- These embeddings convert textual data into numerical feature vectors suitable for ML models.

ML Classification Models

- The extracted feature vectors are analysed using multiple ML classifiers to perform classification tasks:
 - **DT:** Uses hierarchical decision rules to classify text data.
 - **KNN:** Classifies based on similarity between feature vectors.
 - **GNB:** Applies probabilistic reasoning for classification.
 - **RF:** Uses an ensemble of decision trees for robust predictions.
- Each classifier independently predicts the language class for comparison and evaluation.

Text Analysis and Validation Module

- The system includes a validation component to analyze input text before classification.

- It ensures that the input text meets minimum requirements such as length and quality.
- The module improves reliability by filtering invalid or insufficient inputs.
- It enhances interpretability by preparing clean and meaningful data for analysis.

Prediction Results and Output Generation

- After processing, the system predicts the language category of the input text.
- The results are displayed in the graphical interface along with confidence scores.
- Model-wise predictions and performance metrics are also presented for comparison.
- The output enables users to understand classification results clearly.

Remote Text Prediction Workflow

- The architecture supports remote prediction through a client–server communication model.
- External users can submit text inputs to the Flask server for analysis.
- The server processes the input using trained models and returns structured prediction results.
- This enables scalable and real-time text classification across different platforms.

Model Evaluation and Retraining

- The system evaluates model performance using metrics such as accuracy, precision, recall, and F1-score.
- Visualization tools such as confusion matrices and performance graphs are used for analysis.
- Models can be retrained when new data becomes available to improve accuracy.
- This ensures continuous learning and adaptability of the system over time.

4. RESULTS ANALYSIS

The result analysis phase presents the outcomes obtained after applying preprocessing, feature extraction, and machine learning techniques on the textual dataset. It provides a clear understanding of how effectively the system performs in classifying and analysing the input data. The results are evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. These metrics help in assessing the reliability and consistency of different models used in the study. Visualization techniques such as graphs and confusion matrices further support the interpretation of results. By comparing multiple models, the analysis highlights variations in performance and identifies the most suitable approach. This phase plays a vital role in validating the effectiveness of the analytical framework.

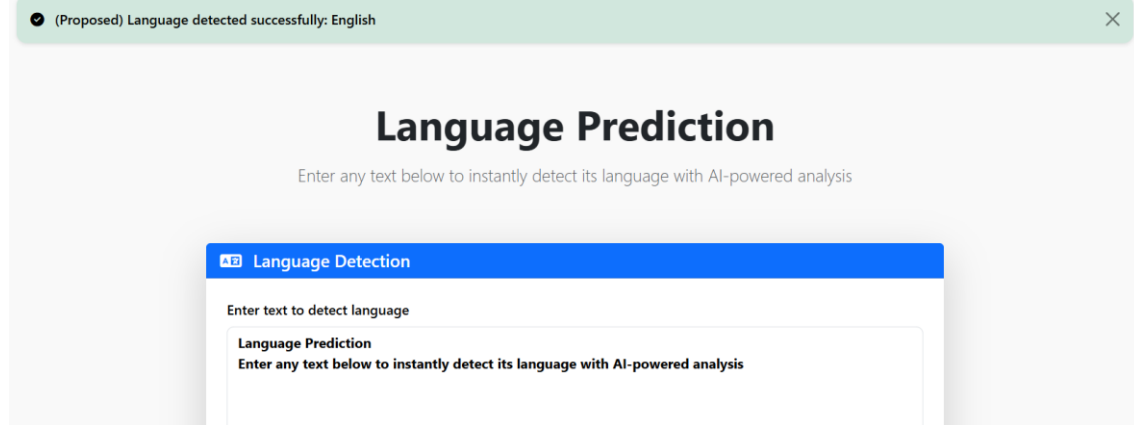


Figure 3: Language Prediction by using Proposed Model.

Figure 3 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 RF. 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.

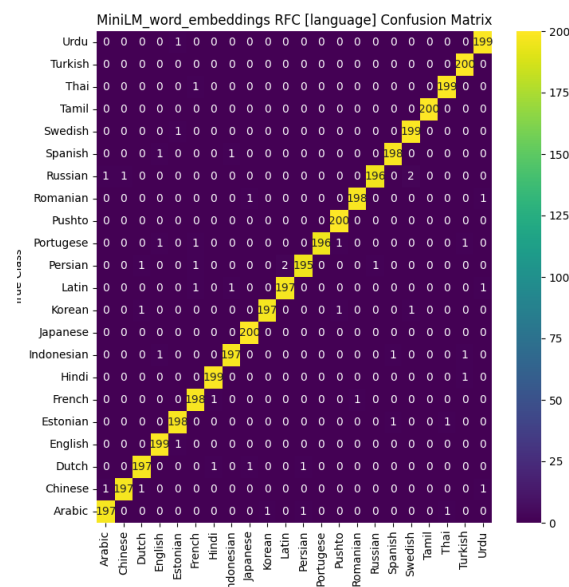


Figure 4: Confusion matrix obtained using Mini LM word embeddings of RF.

Figure 4 RF 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.

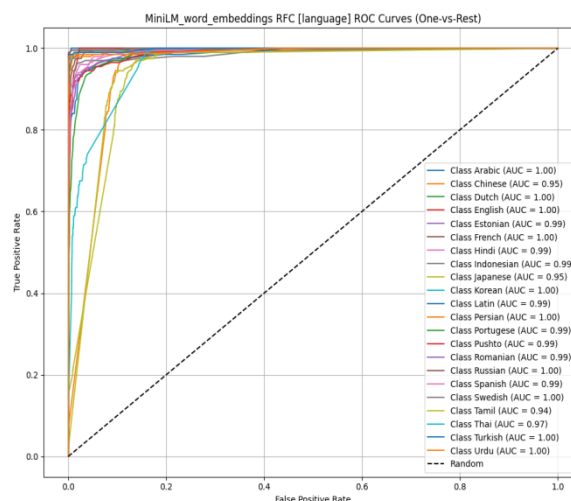


Figure 5: ROC Curve obtained using Mini LM word embeddings of RF.

Figure 5 RF 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.

4.1 Comparative Analysis

The comparative analysis evaluates the performance of the existing baseline and the proposed Mini LM Transformer across multiple dimensions, including accuracy, robustness, and scalability. The results highlight the efficiency of the proposed system in addressing the shortcomings of traditional methods. The RF model performed reasonably well for high-resource languages such as English, French, and Spanish, delivering moderate accuracy and faster execution times. However, it struggled with low-resource languages, code-mixed sentences, and linguistically similar inputs, resulting in frequent misclassifications. Its limited ability to capture contextual nuances reduced its reliability in real-world multilingual scenarios.

Table 1: Performance comparison of DT, KNN, GNB and RF algorithms.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DT	64.16	64.57	64.16	63.37
KNN	80.68	85.68	80.68	78.97
GNB	72.86	73.57	72.86	71.43
RF	99.00	99.00	99.00	99.00

Table 1 presents the performance comparison of MiniLM-based feature embeddings across different machine learning algorithms reveals a clear hierarchy in predictive effectiveness. The DT achieved moderate performance, with an accuracy of 64.16% and an F1-score of 63.37%, indicating limited generalization for language detection. The GNB improved slightly, achieving 72.86% accuracy and 71.43% F1-score, while the KNN classifier demonstrated better performance with 80.68% accuracy and 78.97% F1-score, reflecting its ability to leverage similarity in feature space more effectively. Notably, the RF significantly outperformed all baseline algorithms, achieving near-perfect results across all

metrics (99% accuracy, precision, recall, and F1-score), highlighting its robustness and superior capability in capturing complex patterns in MiniLM embeddings for reliable language detection.

5. CONCLUSION

The language identification framework successfully combines MiniLM transformer-based embeddings with classical machine learning classifiers such as DT, KNN, GNB, and RF to deliver accurate and efficient classification of multilingual text. By utilizing deep semantic representations, the system achieves consistent performance across all stages, including preprocessing, feature extraction, model training, and evaluation. This integrated approach brings together the contextual strength of transformer models and the simplicity and interpretability of traditional algorithms, making it practical for real-world scenarios like multilingual communication systems, content filtering, and information retrieval. Developed using open-source tools such as Hugging Face Transformers and scikit-learn, the framework remains scalable, flexible, and cost-efficient. It provides a dependable and interpretable solution, with opportunities for further enhancement through real-time implementation, incorporation of advanced models, and adaptation to diverse linguistic variations.

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.; Alkhamash, 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.
- [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.
- [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.