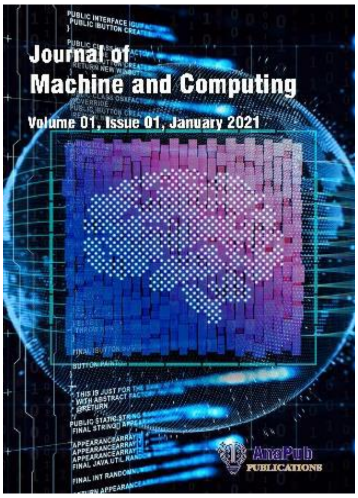
Journal Pre-proof

Expert Crawler: Amalgamation of Deep Learning Models for Multilingual Multiclass Classification of Product Reviews

Priyanka Sharma, Ganesh Gopal Devarajan and Manash Sarkar



DOI: 10.53759/7669/jmc202505058

Reference: JMC202505058

Journal: Journal of Machine and Computing.

Received 24 October 2024

Revised form 12 December 2024

Accepted 20 January 2025

Please cite this article as: Priyanka Sharma, Ganesh Gopal Devarajan and Manash Sarkar, "Expert Crawler: Amalgamation of Deep Learning Models for Multilingual Multiclass Classification of Product Reviews", Journal of Machine and Computing. (2025). Doi: https://doi.org/10.53759/7669/jmc202505058

Expert Crawler: Amalgamation of Deep Learning Models for Multilingual Multiclass Classification of Product Reviews

Priyanka Sharma[1], Ganesh Gopal Devarajan[1,*] and Manash Sarkar[2]

[1]Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Delhi-NCR Campus, Delhi – Meerut Road, Modinagar, Ghaziabad, Uttar Pradesh 201204, India.
[2] Department of Computer Science and Engineering, Atria Institute of Technology, Hebbal, Bengaluru, 560024, Karnataka, India.

*Corresponding author: Ganesh Gopal Devarajan(dganeshgopal@gmail.com)
Contributing authors E-mail(s): ps9627@srmist.edu.in, manash.sarkar26@gmail.com

Abstract

With the proliferation of social platforms for online shopping, accurately predicting item categories from multilingual reviews has become crucial for informed decision-making. This paper addresses the significant challenge of categorizing reviews across diverse languages by enhancing Transformer models for multilingual review classification, addressing key challenges such as efficiency, scalability, and interpretability. To improve model efficiency, we integrate sparse attention mechanisms using mBert, XLM-RoBERTa, and model distillation via DistilBERT, thus balancing performance with reduced computational cost. For data augmentation, we employ back-translation to enrich the training data, thereby enhancing model robustness and generalization across diverse languages. Additionally, to enhance model interpretability, we employ Local Interpretable Model-Agnostic Explanations to provide clear and actionable insights regarding model predictions. The proposed methods are applied to multilingual reviews sourced from products listed on Amazon covering the Spanish, English, German, Hindi, Chinese, Japanese, and French languages. The model achieves a classification accuracy of 88% across 32 product categories, demonstrating its effectiveness in solving the multilingual multiclass categorization problem in the retail sector. This work illustrates the potential of combining advanced natural language processing techniques with innovative approaches to improve the efficiency, accuracy, and interpretability of classification models, thereby facilitating better decision-making in online shopping platforms. With continued research, these models will offer increasingly robust solutions for processing and understanding multilingual data.

Keywords: Expert Crawler, Machine Learning XLM-RoBERTa, LIME, Natural Language Processing, optimizers

1 Introduction

In today's competitive business landscape, customers are pivotal for the success of any organization. The purchasing behavior of consumers, whether conducted online or offline, has become a fundamental aspect of daily life. This makes it imperative for businesses to align their strategies with customer preferences and demands so as to cultivate loyalty and drive sales in the retail sector. Analyzing data derived from customer reviews and feedback is among the most effective methods for gaining insights into customer needs and expectations. However, in a globalized and highly competitive market, customer reviews often span multiple languages and originate from diverse platforms, necessitating multilingual proficiency for accurate interpretation. Traditionally, the task of analyzing multilingual customer feedback has relied heavily on human expertise, which is both labor-intensive and cost-prohibitive. Customer data appears in various forms and must be monitored across multiple channels, including online retail platforms, such as Amazon, social media networks, such as Facebook and Twitter, and content-sharing platforms, such as YouTube. With the advent of advanced technologies, machine learning (ML), deep learning (DL), natural language processing (NLP), and artificial intelligence (AI) have emerged as powerful tools for automating the classification of customer feedback. Natural Language Processing (NLP) has witnessed significant advancements in multilingual applications, with models such as BERT and mBERT achieving considerable success. However, existing solutions face challenges in handling closely related languages and underrepresented dialects. Previous studies have identified gaps in feature engineering and domain adaptability. Addressing these gaps, this research aims to introduce a novel multilingual text processing system that incorporates advanced fine-tuning techniques, domain-adaptive training strategies, and improved feature engineering methods to enhance the accuracy and efficiency of language identification and classification. These technologies reduce the need for manual effort and enhance business decision-making processes [13].

In this context, addressing the challenges posed by multilingual data is particularly critical for large global organizations seeking to optimize their operations across different markets [1]. We propose a model that integrates a collocation-based approach with stochastic gradient descent optimization to tackle these challenges efficiently and cost-effectively. Fig. 1 illustrates the model generation process, which comprises three steps: data gathering, data wrangling, and classification of user reviews across multiple languages. By streamlining the analysis of multilingual [20] customer feedback, this approach offers a scalable solution that enhances the accuracy and efficiency of classification tasks in the retail sector.

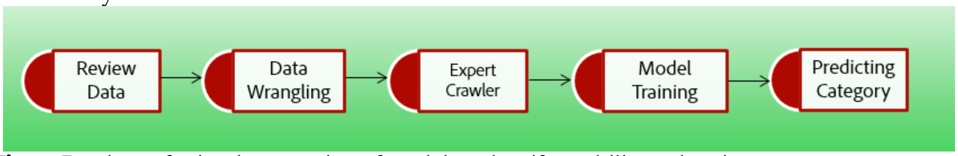


Fig. 1 Roadmap for implementation of model to classify multilingual reviews

Building on the need for the efficient analysis of multilingual customer feedback, this research focuses on automating the categorization of reviews across various plat- forms. A large number of people use social media and other online platforms to communicate in multiple languages. Understanding and classifying such multilingual data presents a formidable challenge, traditionally demanding extensive manual effort and linguistic expertise IIIT Hyderabad, [16]. Automating this process will not only reduce time and resource requirements but also offer valuable insights that have potentially been obscured by language barriers. To significantly enhance the accuracy and efficiency of categorizing multilingual reviews, this study leverages cutting-edge technologies, including advanced ML models, NLP techniques [10], small language models (SLMs), and optimizers. This research demonstrates the potential of automated systems to manage and interpret large volumes of multilingual data by concentrating on platforms such as Amazon and YouTube, ultimately facilitating more informed decision-making [21]. Understanding different languages and analyzing diverse reviews typically demands subject matter expertise and substantial manual labor. However, the implementation of sophisticated AI-driven systems promises to streamline these tasks, offering a scalable and effective solution to the challenges inherent in global customer feedback analysis [8]
This research paper is structured as follows: Section 2 presents the study's objectives and details the data collection process and methodology employed for the proposed model. This section also delves into the specifics of the implementation of the model, highlighting the use of SLMs and ML for classification tasks. Section 3 presents the study's results, Section 4 offers an in-depth discussion, and Section 5 concludes the study.

1.1 Literature Survey

Numerous studies have been considered Table 1, for this research, including a study by Zhu et al. where the researchers evaluated numerous factors affecting the performance of LLMs in translation tasks [27]. Another relevant study by Keung et al. examined an extensive collection of Amazon reviews for multilingual text classification [18]. The corpus included reviews in English, Japanese, German, French, Spanish, and Chinese collected between 2015 and 2019, forming a curated sub- set of reviews specifically designed for multilingual text classification research. With this contribution, the researchers aimed to provide a valuable resource to the research community.

Yet another significant study conducted by Yu et al. proposed a BERT-based text classification model named BERT4TC, which builds auxiliary sentences and convert a classification task into a binary sentence-pair format, with the aim of stating data problems related to limited training and task awareness [26]. The authors also presented the implementation and architecture details for BERT4TC, along with an approach for evaluating BERT's performance across different domains. Babhulgaonkar provided a summary of the challenges and significance of automated language identification using ML algorithms. This paper also emphasized the importance of "language identification" and "machine translation" in making cross-lingual information accessible [3]. The study highlights the challenge of distinguishing between closely related languages, such as Hindi, Marathi, and Sanskrit, that share many similarities but require unique attributes for accurate classification. The paper used Hindi and Sanskrit as examples to demonstrate the process of distinguishing between different languages.

Another research by Wu examines the challenges and applications of entity-linking, focusing on the prominent strategies to address these issues. The scholars also list the knowledge bases, datasets, estimation criteria, and associated challenges of entity-linking [11]. Notably, these existing methods are significant to the linking of analogous languages. While multilingual entity-linking has been a topic of interest for years, relatively little work has been done to refine and advance this specific area of research.

Table 1 Comparison of strength and gaps of existing strategies

Study	Strength	Gaps
Zhu et al. 2023	Evaluated numerous factors affecting NMT performance in translation tasks.	Limited exploration of fine-tuning methods for translation tasks.
Keung et al. 2020	Extensive collection of multilingual Amazon reviews for text classification.	Potential dataset biases affecting classification accuracy.
Yu et al. 2019	BERT4TC model for text classification with auxiliary sentence conversion.	Need for generalization of BERT-based models to varied NLP tasks.
Babhulgaonkar and Sonavane 2020	Emphasized challenges in automated language identification.	Insufficient feature differentiation leading to classification errors.
De Cao et al. 2022	Examined challenges and applications of entity-linking with knowledge bases and datasets.	Lack of refined approaches for advancing multilingual entity-linking.

2. Methods

The data gathering process started with the collection of reviews from Amazon Web Services and YouTube in Spanish, English, German, Hindi, Chinese, Japanese, and French, spanning 32 product categories. Subsequently, data preprocessing and wrangling techniques were applied to filter, clean, and merge the reviews. The proposed method, termed "Expert Crawler," aims to ease multilingual language understanding, feature identification, and extraction, culminating in model construction based on the training dataset. The model's performance was assessed using accuracy, Matthew's correlation coefficient (MCC) on test data set along with Precision, Recall, and F1-score across all 32 product categories. Ultimately, the trained model was employed to predict new, unseen product categories written in the seven languages considered in this study.

2.1 Proposed Methodology: Expert Crawler

Transformers, introduced by, have significantly advanced the field of NLP. However, despite their success, these models face challenges in computational efficiency, scalability, and interpretability. In this paper, we propose "Expert Crawler" to evaluate several techniques and address these issues, specifically focusing on classification tasks. We explore "Efficiency Improvements" by utilizing sparse attention mechanisms and model distillation to reduce computational costs. We also implement "Data Augmentation" by applying back-translation to enhance training data diversity. Additionally, we enhance "Model Interpretability" by employing LIME to explain the model predictions.

2.1.1 Reviews were Splitting and Tokenization

We employed language-specific tokenizers [23]to handle the unique characteristics of each language:

a) For English, Spanish, French, and German, we used the SpaCy library, which provides robust tokenization for these languages.

b) For Hindi, we utilized the iNLTK library, which is specifically designed for Indian languages.

c) For Chinese, we used Jieba, a popular Chinese text segmentation library.

d) For Japanese, we employed MeCab, a part-of-speech and morphological analyzer, for tokenization.

2.1.2 Text Normalization

This includes processes of contraction, spelling correction, and lowercasing.

a) Contractions: We implemented language-specific contraction expansion for languages that use them (primarily English, French, and Spanish).

b) Spelling Correction: We used language-specific dictionaries and the SymSpell algorithm, adapting it to the orthography of each language.

c) Lowercasing: We applied to languages with case distinction (not applied to Chinese and Japanese).

2.1.3 Text Simplification

a) Lemmatization: Language-specific lemmatizers were used, as follows:

1. For French, German, and Spanish: SpaCy lemmatizers
2. For Hindi: iNLTK lemmatizer
3. For Chinese and Japanese: Custom rule-based approaches, as these languages do not utilize traditional lemmatization

b) Stop Words Removal: Custom stop word lists were implemented for each language, accounting for linguistic and domain-specific factors.

2.1.4 Feature Engineering

a) Contextual embedding and attention (Pre-trained Transformers): We employed the XLM-RoBERTa [9] model, a multilingual variant of RoBERTa pre-trained on 100 languages. This model generates contextual embeddings that functioned across all our target languages, facilitating unified representation and potentially enabling zero-shot cross-lingual transfer.

b) Factorization (Product categories): We implemented a multilingual product category embedding system. Category names were machine-translated into all tar- get languages and then embedded within a shared multilingual space. This approach enabled consistent category representation across languages.

2.1.5 Advanced Learning Techniques

a) Sparse attention: We implemented a language-aware sparse attention mechanism that dynamically adjusted the sparsity based on the language, accounting for variations in average sentence length and information density across languages (e.g., sparser for Japanese, which typically requires fewer characters to convey the same information as English). We implemented this sparse attention mechanism in a language-agnostic manner, operating on the token-level representations derived from XLM-RoBERTa. This approach ensured consistent application across all languages, regardless of their script or grammatical structure.

b) Few-shot learning: We extended our few-shot learning approach to accommodate multilingual scenarios:

1. The meta-learning model was trained on a diverse set of tasks spanning all target languages.
2. Language-specific features were incorporated into the task representations, enabling the model to adapt to language-specific nuances [5].
3. A cross-lingual few-shot learning framework was developed. This framework leverages examples from resource-rich languages (such as English) to enhance clas- sification for low-resource languages. We extended the model-agnostic meta-learning (MAML) algorithm to incorporate language-agnostic features, enabling effective knowledge transfer across languages.

2.1.5.1 Language-Specific Considerations

a) Script handling: For languages with non-Latin scripts (Hindi, Chinese, Japanese), Unicode normalization was implemented to ensure consistent text representation.

b) Translation augmentation: Applying back-translation to enhance training data diversity. c) Model interpretability: Employing LIME to explain model predictions.

2.1.6 Model Distillation

In this ML technique, knowledge from a large, complex model (the "teacher") is transferred to a smaller, more efficient model (the "student"), thus allowing the latter to achieve performance comparable to the former, despite its fewer parameters. To create a more efficient model, we utilized DistilBERT, a smaller and faster variant of BERT. DistilBERT retains 97% of BERT's language understanding while being 60% faster and 40% smaller.

2.1.7 Multilingual Model Architecture

We designed a hierarchical attention network that first processed each language separately and then combined the language-specific features. This allowed the model to capture both language-specific nuances and cross-lingual patterns.

2.2 Data Collection

Reviews were collected from Amazon's marketplace (https://registry.opendata.aws/amazon-reviews/) in the US, Spain, Germany, China, Japan and France for English, Spanish, German, Chinese, Japanese and French languages, respectively. The data, initially in Java Script Object Notation (JSON) format, was transformed into the Comma-Separated Value (CSV) format. Hindi language reviews were gathered separately from a GitHub repository (https://github.com/MrRaghav/Complaints-mining-from-Hindi-product-reviews) in excel format [17]. The Hindi reviews were also converted into the CSV for- mat and then combined with the Spanish, French, Chinese, German, Japanese, and English language reviews to create a comprehensive dataset of seven languages. We consolidated the shared categories across all seven datasets. The dataset comprised three sections: *Train, Dev, and Test*. The training dataset encompassed roughly 61,963 reviews across all 32 product categories. The data was split into three subsets: 70% allocated to training, 15% to validation, and the remaining 15% to testing. This distribution ensured that the training set, being the largest portion, allowed the model to learn from as much data as possible. The validation set was used to tune the hyperparameters and evaluate the model during training. This subset had to be large enough to provide reliable performance estimates. The test set was used to assess the performance of the final model. This subset was designed to provide a statistically significant measure of the model's capabilities.

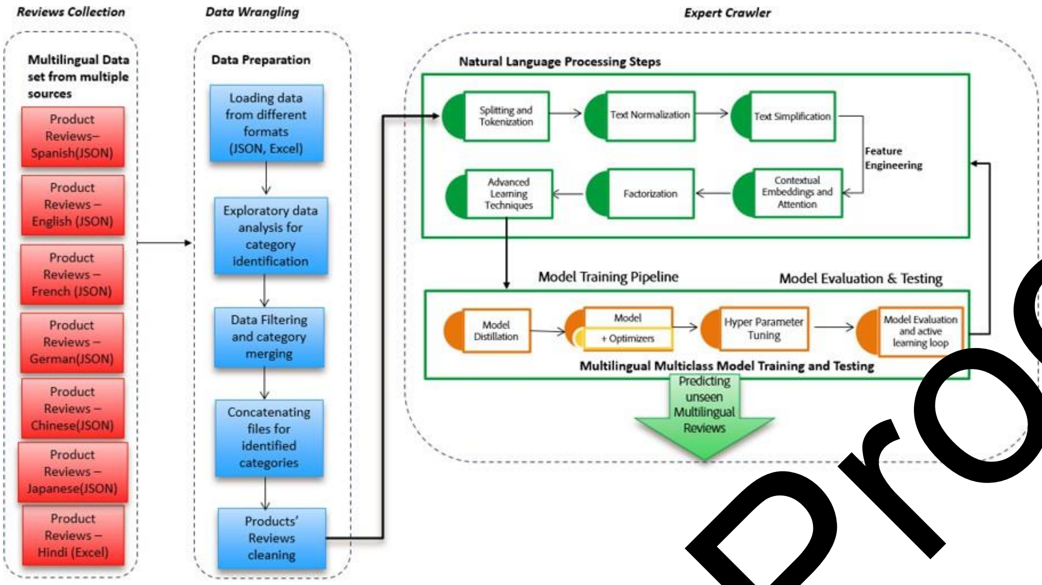


Fig. 2 Expert Crawler Approach for Multilingual Multiclass classification of online reviews

As shown in Fig. 2, the data collection phase began with sourcing reviews in Hindi, Spanish, French, Chinese, German, Japanese, and English from diverse platforms. These reviews, acquired in various formats such as JSON and Excel, were subjected to subsequent processing in the *Data Wrangling* stage. In this stage, we conducted a comprehensive exploration of the reviews, aiming to align them with appropri- ate categories, rectify spelling errors in product category names, standardize product category names to lowercase, and merge categories with similar designations. Subsequently, the refined dataset proceeded to the Expert Crawler phase. In this phase, the reviews underwent exploratory data analysis, followed by a series of NLP techniques, including *Splitting*, *Tokenization*, *Contractions*, *lowercasing*, *Spelling Correction*, *and Lemmatization*, which refined the dataset for further analysis. Next, we implemented a factorization process to convert the English category labels into numerical values, assigning a numerical identifier ranging from 1 to 32 to each unique category. After factorization, we eliminated stop words from the reviews in Hindi, Spanish, French, Chinese, German, Japanese, and English. Following this, we employed collocation analysis, which examined the proximity of words or phrases within a text corpus. This technique, often used in NLP, aims to identify meaningful word combinations that convey specific contexts or meanings.

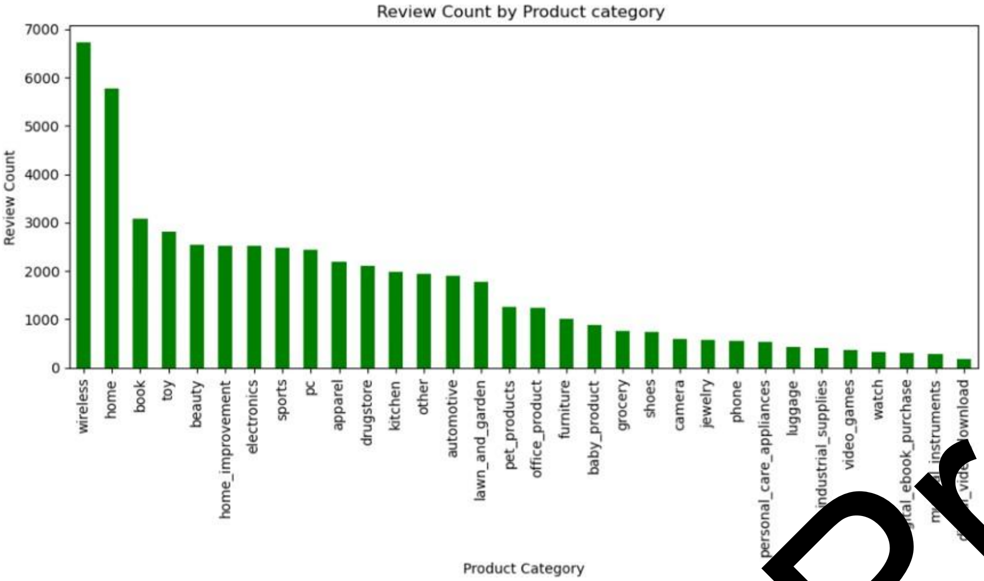


Fig. 3 Chart to show all 32 Product categories with its review count.

Fig. 3 presents an overview of the review counts across 32 product categories, highlighting consumer interest and engagement levels in each segment. The wireless category has the highest number of reviews, indicating strong demand for products such as mobile phones, routers, and accessories. Following closely, the home category also enjoys a significant number of reviews, reflecting the popularity of household items, furniture, and decor. Books, a timeless category, continue to garner attention, suggesting a consistent readership across various genres. Toys and beauty products are also highly reviewed, showing strong interest from parents and individuals focused on self-care. The home improvement sector, encompassing tools and building materials, sees considerable engagement, indicating a growing trend in DIY projects. Similarly, electronics such as gadgets and home appliances remain a favorite among consumers.

Other notable categories include sports, which highlights interest in fitness and outdoor activities, and PCs, with users actively engaging in reviewing computing devices and accessories. Apparel and drugstore products also see a high volume of reviews, reflecting ongoing interest in fashion and personal wellness. Kitchen appliances and accessories, categorized under kitchen, are frequently reviewed by cooking enthusiasts. Meanwhile, specialized categories such as automotive, lawn and garden, and pet products have a steady review presence, showing targeted consumer interest. Office products and furniture receive attention from both home and professional users.

Categories with moderate review counts include baby products, indicating the cautious nature of parents seeking quality, and grocery, which reflects the increasing shift toward online shopping for daily essentials. Shoes, cameras, and jewelry maintain steady consumer feedback based on style, quality, and personal preferences. Some categories, such as personal care appliances, luggage, and industrial supplies, receive fewer reviews but still represent niche markets with dedicated buyers. Appliances, including large household items like refrigerators and washing machines, have a moderate review count, whereas video games and watches attract feedback primarily from enthusiasts.

Towards the lower end of the review spectrum, categories such as digital ebooks purchases, music instruments, and digital video downloads have relatively fewer reviews, likely due to the digital nature of the products and their specialized audience. Overall, the distribution of reviews suggests that consumer engagement is highest in essential and widely used product categories, while niche or digital products receive comparatively less feedback. This analysis provides valuable insights into consumer behavior and market trends across diverse product segments across English, French, Spanish, German and Hindi languages, with the product category names considered in English.

2.3 Implementation Details

The sequential execution of the process outlined for Expert Crawler was conducted using both CPUs and 16 GB of GPU RAM with Nvidia A100 machine. This approach ensured a step-by-step progression through Expert Crawler's various tasks, enabling efficient processing and analysis of the multilingual data collected from diverse sources. Notably, we implemented the pro- posed approach using Python to obtain the desired results. Python's versatility and rich ecosystem of libraries make it well suited for multilingual review classification tasks, such as data loading, data pre-processing, NLP, ML and DL model development, and result analysis. The reviews were then loaded from all the seven files in various formats and, as mentioned in the proposed plan, the data pre-processing step was performed. In the Expert Crawler process, we sequentially applied the NLP steps to attain the desired features.

2.3.1 Advanced Learning Techniques

a) Sparse attention: We implemented a language-aware sparse attention mechanism that dynamically adjusted the sparsity based on the language, accounting for variations in average sentence length and information density across languages (e.g., sparser for Japanese, which typically requires fewer characters to convey the same information as English). We implemented this sparse attention mechanism in a language-agnostic manner, operating on the token-level representations derived from XLM-RoBERTa. This approach ensured consistent application across all languages, regardless of their script or grammatical structure.

Few-shot learning: We extended our few-shot learning approach to accommodate multilingual scenarios:

1. The meta-learning model was trained on a diverse set of tasks spanning all target languages.
2. Language-specific features were incorporated into the task representations, enabling the model to adapt to language-specific nuances [5].
3. A cross-lingual few-shot learning framework was developed. This framework leverages examples from resource-rich languages (such as English) to enhance classification for low-resource languages. We extended the model-agnostic meta-learning (MAML) algorithm to

incorporate language-agnostic features, enabling effective knowledge transfer across languages.

2.3.2 Language-Specific Considerations

a) Script handling: For languages with non-Latin scripts (Hindi, Chinese, Japanese), Unicode normalization was implemented to ensure consistent text representation.
b) Translation augmentation: Applying back-translation to enhance training data diversity.
c) Model interpretability: Employing LIME to explain model predictions.

2.3.3 Model Distillation

In this ML technique, knowledge from a large, complex model (the "teacher") is transferred to a smaller, more efficient model (the "student"), thus allowing the latter to achieve performance comparable to the former, despite its fewer parameters. To create a more efficient model, we utilized DistilBERT [24], a smaller and faster variant of BERT. DistilBERT retains 97% of BERT's language understanding while being 60% faster and 40% smaller.

2.3.4 Multilingual Model Architecture

We designed a hierarchical attention network that first processed each language separately and then combined the language-specific features. This allowed the model to capture both language-specific nuances and cross-lingual patterns.

2.3.5 Hyperparameter Tuning

The following steps must be followed for hyperparameter tuning:

2.3.5.1.1 Define a search space for hyperparameters, including the learning rate, training batch size, evaluation batch size, number of epochs, epsilon, learning rate scheduler, warm up steps and optimizer

2.3.5.1.2 Employ a hyperparameter optimization technique to explore the search space efficiently.

2.3.5.1.3 Evaluate model performance using a validation set to select the optimal hyperparameter configuration.

ExpertCrawler combined XLM-RoBERTa and DistilBERT to effectively process and understand multilingual text, leveraging amalgamated features of XLM-RoBERTa and DistilBERT for model training, thus offering a promising approach.As mentioned in Algorithm 1, XLM-RoBERTa's expertise in handling multiple languages ensures accurate tokenization, while DistilBERT's efficiency and performance enable building smaller yet powerful models. This combination offers advantages in speed, accuracy, and adaptability in carrying out various multilingual tasks. Furthermore, it ensures that the system can effectively process and analyze product reviews in English, Spanish, French, German, Hindi, Chinese, and Japanese languages [12], leveraging both language-specific tools and cross-lingual models to achieve robust performance across diverse linguistic contexts.

```
# 'toy':
  . Most correlated bigrams:
. faltan piezas
. fallos pintura
. funko pop
  . Most correlated trigrams:
. dicen dicen viene
. popped first day
. schon beim anziehen
# 'video_games':
  . Most correlated bigrams:
. game play
. viene juegos
. wrong game
  . Most correlated trigrams:
. nul aucun intérêt
. caja mas grande
. sent wrong game
# 'watch':
  . Most correlated bigrams:
. montre fonctionne
. calidad reloj
. uhr nie
  . Most correlated trigrams:
. marca bien hora
. बह अच बह
. schöne uhr leider
# 'jewelry':
  . Most correlated bigrams:
. cuello negro
. creo plata
. pulsera pandora
  . Most correlated trigrams:
. gaudy cheap looking
. beim ersten tragen
. one star turned
# 'kitchen':
  . Most correlated bigrams:
. mucha potencia
. coffee maker
. non stick
  . Most correlated trigrams:
. solo sale chorro
. months use longer
. lids stay closed
# 'lawn_and_garden':
  . Most correlated bigrams:
. mayoría semillas
. [illegible]
. ningún mosquito
  . Most correlated trigrams:
. voy devolver vale
. worst hose ever
. return even last
```

Fig. 4 bi-grams and tri-grams using Collocation.

For feature extraction, the most correlated N-grams had to be identified, as depicted in Fig. 4. Collocations and chi-square methods were employed for this purpose, along with State-of-the-Art (SOTA) models. This process aimed to capture the most relevant linguistic patterns and associations within the text data, enabling effective classification across multiple languages and product categories. The formula for the chi-square is provided in equation 1:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{1}$$

where i denotes the feature, j refers to the specific class, Oij is the frequency of feature i and class j occurring together, and Eij is the frequency of feature i occurring without class j. The chi-square between each feature and class was computed, and the features with the highest chi scores were chosen.

Combining the n-grams derived from chi-square analysis using a vectorizer yielded effective results in multi-language classification. The SOTA model, initially introduced outlines the Transformer model [25]. Notably, this model relies solely on self-attention to compute the representation of a sequence or sentence, allowing for the connection of different words within the same sequence. Following feature extrac- tion, we employed various ML algorithms, including multinomial naïve Bayes, support vector machine, stochastic gradient descent, logistic regression, decision tree, random forest, mBert, XLM-RoBERTa, and DistilBert, Expert Crawler to train and evaluate the validation and test datasets. Each algorithm contributed uniquely to developing a robust final model, with methods ranging from probabilistic and optimization- based approaches to ensemble learning applied to ensure comprehensive and precise classification performance. Expert Crawler uses the following algorithm:

Algorithm 1 Training Algorithm for Expert Crawler

Input: Input data X, teacher model M_T (XLM-RoBERTa), student model M_S (DistilBERT), learning rate η, temperature T, balancing factor α

Output: Optimized student model M_S

Start

for each encoder layer do

Compute attention weights using query Q, key K, and value V

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Integrate multiple attention heads for enhanced repre- sentation

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

$$\text{head} = \text{Attention}(QW^Q, KW^K, VW^V)$$

Apply non-linearity and transformation position wise feed-forward network

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Normalize the output for stability

$$\text{LayerNorm}(x) = \frac{x - \mu}{\sigma + \epsilon}\gamma + \beta$$

Embed positional information

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right)$$

Use MLM loss for pre-training on multilingual data

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in \text{masked tokens}} \log P(x_i | x_{\text{context}})$$

Combine teacher model outputs with MLM loss

/* **Knowledge Distillation Loss** */

$$\mathcal{L}_{\text{distill}} = \alpha\mathcal{L}_{\text{CE}}(y, \hat{y}) + (1 - \alpha)T^2\mathcal{L}_{\text{KL}}\left(\frac{S}{T}, \frac{T}{T}\right)$$

$L_{\text{CE}}(y, \hat{y})$: Cross-entropy loss between true labels y and predictions $\hat{y}$

$\mathcal{L}_{\text{KL}}$: Kullback-Leibler divergence between student and teacher model outputs

Compute softened outputs for distillation

$$S = \text{softmax}\left(\frac{\text{logits}}{T}\right)$$

12

Update M_S using gradient descent:

$$M_S \leftarrow M_S - \eta\nabla\mathcal{L}_{\text{distill}}$$

end for loop

return

end

3 Results

The performance assessment of the proposed "Expert Crawler" technique involved the analysis of various accuracy metrics 2, such as precision 3, recall 4, F1 Score 5, confusion matrix, and Matthews correlation coefficient (MCC) 6. To mitigate loss in multilingual data, SGD optimizer [19] was employed along with the Modified Huber loss as parameters. Before applying the model, the reviews were partitioned into training, development, and testing datasets, and then assessed using different ML algorithms. For quantitative comparison, multiclass accuracy was utilized as the performance metric, calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Here
TP (True Positive): Classes that are correctly predicted as positive.
FP (False Positive): Classes that are incorrectly predicted as positive.
TN (True Negative): Classes that are correctly predicted as negative.
FN (False Negative): Classes that are incorrectly predicted as negative.

A high accuracy value indicates that the model is making correct predictions most of the time. However, in cases of imbalanced datasets, accuracy might not be the best metric to rely on.
Precision quantifies the accuracy of positive predictions. It is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Precision measures how many of the predicted positive instances are actually correct. A high precision score indicates that the model produces fewer false positives, which is particularly useful in applications like spam detection or medical diagnosis.
Recall, also known as sensitivity or true positive rate, measures the model's ability to detect actual positives and is given by:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

Recall focuses on identifying all positive instances in the dataset. A high recall value ensures that most of the actual positive cases are detected, which is crucial in applications like fraud detection or disease diagnosis where missing a positive case is costly.
The F1 Score provides a balance between precision and recall, and is computed using the harmonic mean of the two:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

The F1 score is particularly useful when there is an uneven class distribution, as it considers both false positives and false negatives. A higher F1 score signifies a better balance between precision and recall.
Furthermore, MCC is a more robust evaluation metric that considers all four confusion matrix components and provides a balanced measure of the model's quality:

$$\text{MCC} = \frac{(TN \times TP) - (FN \times FP)}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FP)}} \tag{6}$$

MCC ranges between -1 and +1, here
+1 means the best arrangement between the predicted values and actual values.
0 means no arrangement, i.e., the prediction is random with respect to the actuals.

MCC is particularly useful in evaluating model performance on imbalanced datasets, as it considers all classes equally.

Also, LIME was applied to generate interpretable explanations for individual predictions. It involved the following steps:

a) Perturbing the input features and observing its impact on the model's output.
b) Identifying the most important features contributing to the prediction.

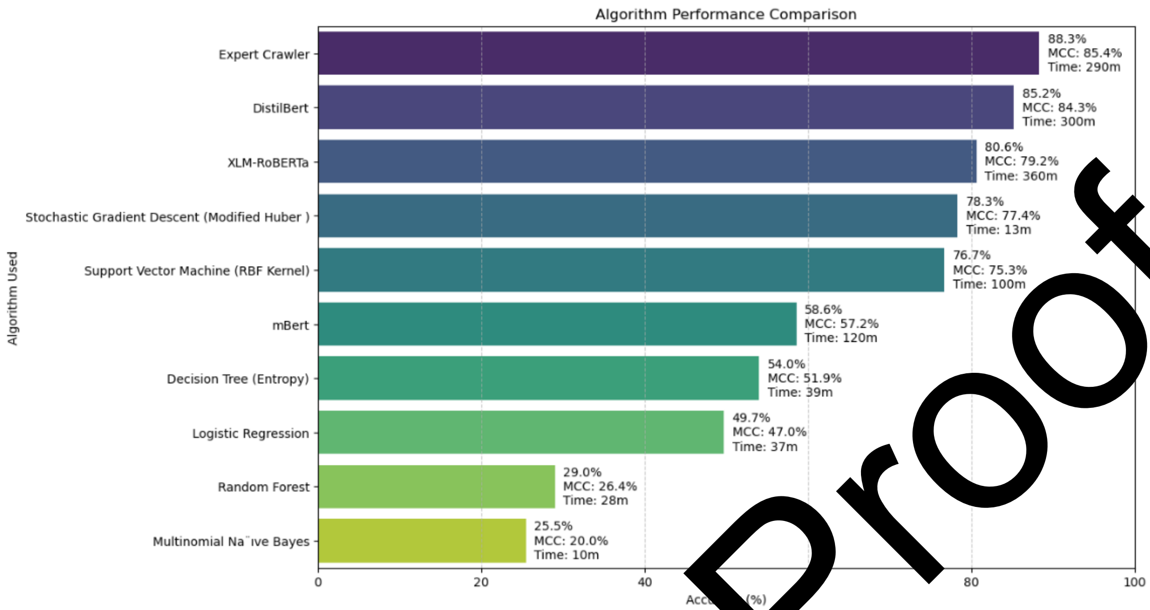


Fig. 5 Accuracy, MCC and Time comparison for the algorithms implemented

Fig. 5 presents a comparison of various ML algorithms in terms of multiclass accuracy, MCC, and training time. The algorithms evaluated include decision tree with both Gini index and entropy criteria, multinomial naïve Bayes, logistic regression, random forest, support vector machine with "rbf" kernel, and SGD utilizing loss functions, such as modified Huber.

Table 2 Comparison of accuracy among proposed and existing strategies

Algorithm Used	Accuracy	MCC
Decision Tree (Gini)	25.3	28.5
Multinomial Naïve Bayes	25.5	20.0
Random Forest	29.0	26.4
Logistic Regression	49.7	47.0
Decision Tree (Entropy)	54.0	51.9
mBert	58.6	57.2
Support Vector Machine (RBF Kernel)	76.7	75.3
Stochastic Gradient Descent (Modified Huber)	78.3	77.4
XLM-RoBERTa	80.6	79.2
DistilBert	85.2	84.3
Expert Crawler	88.3	85.4

As evident from Table 2, the Expert Crawler outperformed the other algorithms by 88.3%. This study compared traditional machine learning algorithms with transformer-based models [2] for text classification. Classical models like Decision Trees and Logistic Regression showed limited accuracy, while small language models like XLM-

RoBERTa (80.6%) and DistilBERT (85.2%) performed significantly better highlighting the clear advantage of transformer models for efficient and accurate text classification.

Table 3 Comparison of accuracy among proposed and existing strategies

Method	Languages	Category Average Score(%)
Fine grained Classification	En, Fr, De, Es, Za, Jh	59.2
Zero-Shot Cross-lingual	En, Fr, De, Es, Za, Jh	44.0
Few-Shot Cross-lingual	En, Fr, De, Es, Za, Jh	78.0
Expert Crawler	En, Fr, De, Es, Za, Jh, Hi	88.3

Table 3 presents a comparison of the accuracy metrics of the proposed model and existing technologies. Fine-grained classification achieved a 59.2% category average, while zero-shot cross-lingual attained 44% accuracy across six languages: English, French, Spanish, Japanese, German, and Chinese. Few-shot cross-lingual achieved 78% accuracy, while the proposed model achieved an average accuracy of 88.3% across seven languages: Hindi, Spanish, French, Chinese, German, Japanese, and English.

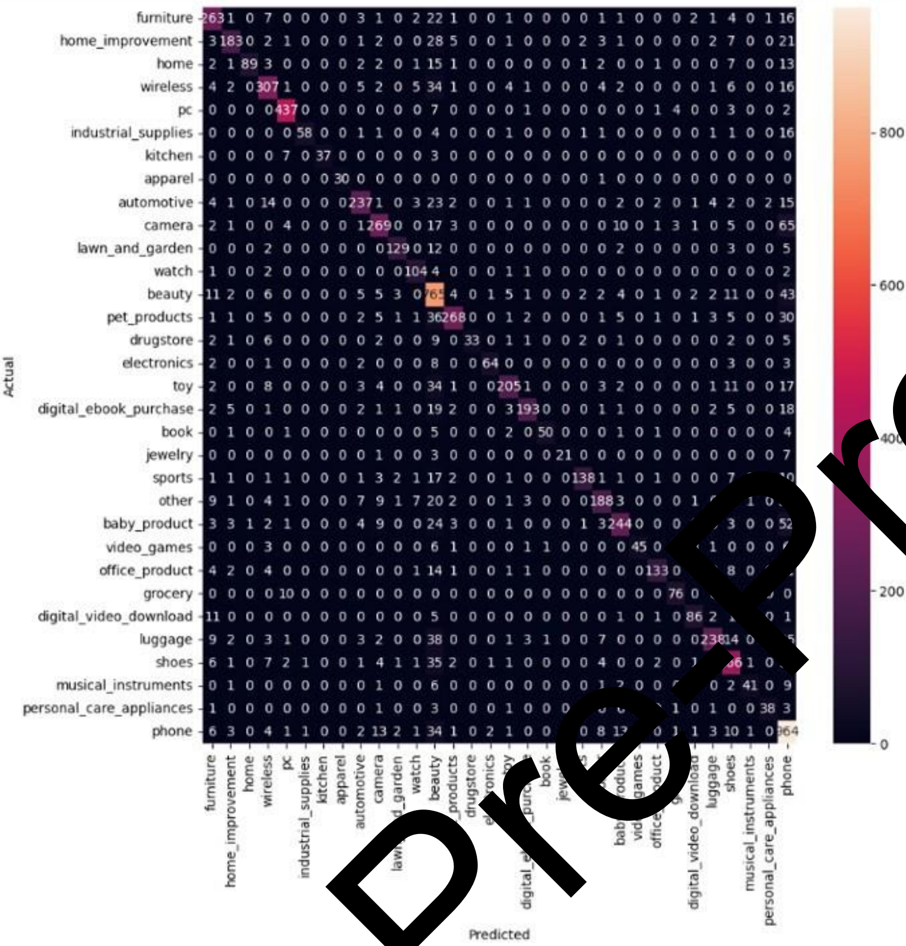


Fig. 6 Heatmap showing 32 product categories.

Fig. 6 presents a heatmap depicting all 32 categories. The contingency matrix illustrates the classification of reviews in seven different languages across these 32 categories, highlighting the values for each category. Correct predictions are shown along the diagonal of the matrix in their corresponding colors, while misclassified values are indicated in the off-diagonal cells. The rows denote the actual values of the 32 categories, while the columns represent the predicted values for these categories. By employing our proposed approach and fine-tuning the parameters, we achieved a respectable accuracy in classifying multilingual customer reviews [6]. The multicolored bar on the right side of the matrix represents the number of reviews per category.

4 Conclusion & Future Work

In the present implementation, the proposed model achieved an average accuracy of 88.3% across seven languages: Hindi, Spanish, French, Chinese, German, Japanese, and English with hyperparameters learning rate of 2e-5, training and evaluation batch sizes of 8, Adam optimizer with betas (0.9, 0.999) and epsilon of 1e-8, a linear learning rate scheduler with 500 warmup steps, and 25 epochs.As a prospective avenue for further exploration, expanding the scope to include additional languages could enhance the validation of our results. Furthermore, increasing the sample size for each language across various categories is another potential direction for future research.

The Expert Crawler process demonstrates superior time and cost efficiency compared to other multilingual classification approaches [15]. The process operates efficiently, scaling seamlessly from smaller to larger datasets by leveraging open-source libraries and state-of-the-art models on both CPU and GPU architectures. This method demonstrates efficacy in handling imbalanced data, which is a common occurrence in many significant business scenarios [7].

Furthermore, the Expert Crawler approach offers versatility by easily adapting to diverse languages and applications, including sentiment analysis [22], biomedical literature [4], spam detection, fake news detection [11], and hate speech identification [14]. Another notable advantage of the Expert Crawler approach lies in its integration of both traditional ML and DL techniques. This amalgamation enables the model to effectively capture intricate relationships between features and code categories in multilingual scenarios.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] M. Artetxe and H. Schwenk, “Massively multilingual sentence embeddings for zero shot cross-lingual transfer and beyond,” arXiv preprint arXiv:1812.10464, 2018. doi: 10.48550/ARXIV.1812.10464.

[2] S. H. Asefa and Y. Assabie, “Transformer-based Amharic-to-English machine translation with character embedding and combined regularization techniques,” IEEE Access, vol. 12, pp. 1–12, 2024. doi: 10.1109/ACCESS.2024.10813169.

[3] A. Babhulgaonkar and S. Sonavane, “Language identification for multilingual machine translation,” in Proc. 2020 Int. Conf. Commun. Signal Process. (ICCSP), Chennai, India, Jul. 2020, pp. 1–5. doi: 10.1109/ICCSP48568.2020.9182184.

[4] A. Basile and C. Rubagotti, “CrotoneMilano for AMI at Evalita2018: A performant cross-lingual misogyny detection system,” in EVALITA Evaluation of NLP and Speech Tools for Italian, Accademia University Press, 2018, pp. 206–210. doi: 10.4000/books.aaccademia.4734.

[5] E. Y. A. Charles and R. Weerasinghe, “Machine reading comprehension for the Tamil language with translated SQuAD,” IEEE Access, vol. 13, pp. 1–[illegible], 2025. doi: 10.1109/ACCESS.2025.10844269.

[6] X. Chen *et al.*, “Adversarial deep averaging networks for cross-lingual sentiment classification,” Trans. Assoc. Comput. Linguist., vol. [illegible], pp. 557–[illegible]70, Dec. 2018. doi: 10.1162/tacl_a_00039.

[7] S. K. W. Chu, R. Xie, and Y. Wang, “Cross-language fake news detection,” Data Inf. Manage., vol. 5, no. 1, pp. 100–10[illegible], Jan. 202[illegible]. doi: 10.2[illegible]78/dim-2020-0025.

[8] A. Conneau, K. Khandelwal, *et al.*, “Unsupervised cross-lingual representation learning at scale,” arXiv preprint arXiv:1911.02116, 2019. doi: 10.48550/ARXIV.1911.02116.

[9] A. Conneau, G. Lample *et al.*, “XNLI: Evaluating cross-lingual sentence representations,” arXiv preprint arXiv:1809.05053, 2018. doi: 10.48550/ARXIV.1809.05053.

[10] A. De *et al.*, “A transformer-based approach to multilingual fake news detection in low-resource language,” ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 21, no. 1, pp. 1–[illegible]0, Jan. 2022. doi: 10.1145/3472619.

[11] N. De Cao, *et al.*, “Multilingual autoregressive entity linking,” Trans. Assoc. Comput. Linguist., vol. 10, pp. 274–290, Mar. 2022. doi: 10.1162/tacl_a_00460.

[12] V. Dogra, *et al.*, “A complete process of text classification system using state-of-the-art NLP model,” Comput. Intell. Neurosci., vol. 2022, pp. 1–26, Jun. 2022. doi: 10.1155/2022/1883698.

[13] J. M. Eisenschlos, *et al.*, "MultiFiT: Efficient multi-lingual language model fine-tuning," arXiv preprint arXiv:1909.04761, 2019. doi: 10.48550/ARXIV.1909.04761.

[14] H. Fei and P. Li, "Cross-lingual unsupervised sentiment classification with multi-view transfer learning," in Proc. 58th Annu. Meeting Assoc. Comput. Linguist., Association for Computational Linguistics, 2020, pp. 5759–5771. doi: 10.18653/v1/2020.acl-main.510.

[15] N. Goyal, *et al.*, "Larger-scale transformers for multilingual masked language modeling," in Proc. 6th Workshop Represent. Learn. NLP (RepL4NLP-2021), Association for Computational Linguistics, 2021, pp. 29–33. doi: 10.18653/v1/2021.repl4nlp-1.4.

[16] S. M. Aggarwal, R. Kumar, and A. K., "Efficient multilingual text classification for Indian languages," in Proc. Recent Adv. Natural Lang. Process. - Deep Learn. Natural Lang. Process. Methods Appl., INCOMA Ltd., 2021, pp. 19–25. doi: 10.26615/978-954-452-072-4_003.

[17] K. Karthikeyan, *et al.*, "Cross-lingual ability of multilingual BERT: An empirical study," arXiv preprint arXiv:1912.07840, 2019. doi: 10.48550/ARXIV.1912.07840.

[18] P. Keung, *et al.*, "The multilingual Amazon reviews corpus," arXiv preprint arXiv:2010.02573, 2020. doi: 10.48550/ARXIV.2010.02573.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014. doi: 10.48550/ARXIV.1412.6980.

[20] A. Kumar, "Multilingual natural language processing," IEEE Trans. Neural Netw. Learn. Syst., vol. 1, 2025. doi: 10.1109/TNNLS.2025.10830644.

[21] Z. Li, *et al.*, "Learn to cross-lingual transfer with meta graph learning across heterogeneous language," in Proc. 2020 Conf. Empirical Methods Natural Lang. Process. (EMNLP), Association for Computational Linguistics, 2020, pp. 2290–2301. doi: 10.18653/v1/2020.emnlp-main.179.

[22] G. Manias, *et al.*, "Multilingual text categorization and sentiment analysis: A comparative analysis of the utilization of multilingual approaches for classifying Twitter data," Neural Comput. Appl., vol. 35, no. 29, pp. 21415–21431, Oct. 2023. doi: 10.1007/s00521-023-08629-3.

[23] M. [illegible] Ms[illegible] and Y. S. Jeong, "Tokenizers for African languages," IEEE Access, vol. 1, 20[illegible]. doi: 10.1109/ACCESS.2024.10815724.

[24] V. Sanh, *et al.*, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019. doi: 10.48550/ARXIV.1910.01108.

[25] A. Vaswani, *et al.*, "Attention is all you need," arXiv preprint arXiv:1706.03762, Aug. 2023.

[26] S. Yu, J. Su, and D. Luo, "Improving BERT-based text classification with auxiliary sentence and domain knowledge," *IEEE Access*, vol. 7, pp. 176600–176612, 2019. doi: 10.1109/ACCESS.2019.2953990.

[27] W. Zhu, *et al.*, "Multilingual machine translation with large language models: Empirical results and analysis," arXiv preprint arXiv:2304.04675, 2023. doi: 10.48550/ARXIV.2304.04675.