Enhancing Cross-Lingual Embedding Alignment with Additive Keywords for International Trade Product Classification

Angga Wahyu Anggoro
School of Computer Science and Inform.
Cardiff University
Cardiff, United Kingdom
AnggoroAW@cardiff.ac.uk

Padraig Corcoran
School of Computer Science and Inform.
Cardiff University
Cardiff, United Kingdom
CorcoranP@cardiff.ac.uk

Yuhua Li
School of Computer Science and Inform.
Cardiff University
Cardiff, United Kingdom
liy180@cardiff.ac.uk

***Abstract*—Cross-lingual embedding alignment plays an important role in enabling effective multilingual classification tasks. Although multilingual pretrained language models and fine-tuning techniques are increasingly adopted, current approaches inadequately address specialised domains, where domain-specific terminology and mixed-language content present unique challenges that hinder classification accuracy. This work considers the problem of automatically classifying text-based descriptions of international trade transactions with respect to an international standard Harmonized System (HS) code taxonomy. We propose a novel method that incorporates mixed-language keyword embeddings to improve cross-lingual alignment, focusing on bilingual models, and subsequently leverages this alignment for downstream classification tasks, with particular applicability to low-resource domains. Using a supervised learning framework implemented through neural network architectures, the model is trained on pairs of product descriptions and their corresponding extracted keywords. Experimental results on benchmark bilingual datasets demonstrate significant and consistent improvements in classification performance over baseline models, including in low-resource target language scenarios. The findings demonstrate the effectiveness of incorporating additive keywords as a strategy for cross-lingual embedding alignment, thereby enhancing representation quality and improving classification accuracy.**

***Index Terms*—Bilingual product classification, Keyword-based features, Cross-lingual embedding alignment, Low-resource languages, Harmonized System codes**

I. Introduction

Language diversity plays a critical role in strategic decisions, particularly in the interpretation of regulations to comply with obligations [1]. In the domain of international trade, it is evident that language differences can hinder bilateral trade flows [2]. Trade documents in international trade are governed by the Harmonized System (HS) [3], which presents inherent complexity related to classifying products involving a large number of codes and the use of multiple languages in trade declarations [4], [5]. Product descriptions typically follow local language, but traders widely use international languages like English for clarity and industry relevance [4], [5]. This treatment is permitted under Indonesian law to allow English as a supplementary language to improve the completeness of information [6]. The presence of this bilingual language in the product declarations complicates the trade facilitation process and highlights the necessity for robust classification systems [7].

Most existing studies evaluating product classification models have utilised monolingual datasets and mainly in English [8]–[11]. Monolingual language models limit their applicability in multilingual trade environments, where linguistic diversity is prevalent. On the other hand, non-English datasets have been utilised individually to evaluate model performance in specific regional contexts, such as those in Brazilian Portuguese [12], Korean [13], Chilean [14] and Chinese [15]. Notably, the availability of non-English datasets in the public domain is limited due to non-disclosure agreements with data providers.

Several studies utilised a multilingual model by exploiting cross-lingual features for improved performance [16], [17]. Cross-lingual word embeddings facilitate the transfer of lexical knowledge to align word representations from equivalent text data in embeddings space, which offers benefits for cross-lingual document classification [18]. A method for cross-lingual word embeddings uses two monolingual embeddings and bilingual dictionaries to align their embedding spaces with the MUSE model, which presented this approach effectively [19]. The model has been utilised in a study to classify product titles using German and French languages [20]. Another method uses parallel data consisting of text paired with translations in other languages, demonstrated in the LASER model [21]. LASER employs an encoder-decoder architecture to generate ready-to-use cross-lingual embeddings that work across the languages it was trained on. The XLM model utilised a Transformer-based architecture for cross-language representation learning through contextual representations rather than word-level translation dictionaries and facilitates fine-tuning for domain-specific applications [16]. Additionally, a fine-tuning approach to multilingual models, such as multilingual BERT [22], was utilised in dealing with Chilean and Spanish in trade product classification [12].

The characteristics of trade datasets show they are clearly

dominated by domain-specific product terminology, such as the brevity of the text length, which often yields insufficient features resulting in poor classification performance [23], [24]. For text classification tasks using short texts, a few words often define the model outcomes [25]. Given these aspects, involving keywords of the main product description is potentially supportive in representing product information in a better way for classification tasks. A keyword is a word or phrase that encapsulates the essential information of a text, which potentially serves as a guiding model to emphasise essential features [23], [26].

This study hypothesises that integrating original product descriptions with extracted keywords improves bilingual classification for trade product classification employing HS codes. Utilising a transformer-based language model in this integration and prediction methodology is expected to enhance the representation quality of both text and keywords, which in turn improves the classification performance.

Our main contributions to this paper are summarised as follows.

- We propose a novel configuration of paired training samples consisting of product descriptions and their corresponding cross-lingual keyword features, referred to as mixed keywords. The mixed keywords are generated using a simple yet effective keyword model, which is tailored and subsequently used as a representation of the product category.
- We propose an architecture for a bilingual model consisting of several components of the neural network: a dual encoder, a fusion layer, and deep classification layers, designed for product category prediction (e.g., HS code).
- We propose a bilingual model trained through incremental training to generate a bilingual model that effectively achieves high performance in low-resource settings. The empirical evaluation compared to several established multilingual baselines shows the effectiveness of our proposed method as a potential application for a real-world system.

The paper is structured as follows: first, it outlines the model design, then presents the experimental setup, and subsequently, it presents key results with corresponding analyses, concluding with the main findings.

II. Methodology

Keywords refer to words or phrases that encapsulate the main topic, theme, or essential information within a text and they are valuable in domain-specific contexts where terminology is often highly specialised [27]–[29]. Terminology in domain-specific areas, such as trade, legal, or medical domains, is often complex and noisy, making keywords particularly valuable for narrowing the focus to the core problem [30]. In the classification of particularly short texts with a limited word count, the results are often determined by just one or two key terms [25]. Therefore, keywords can act as discriminative indicators by directing models toward the most informative features of the text [23], [26].

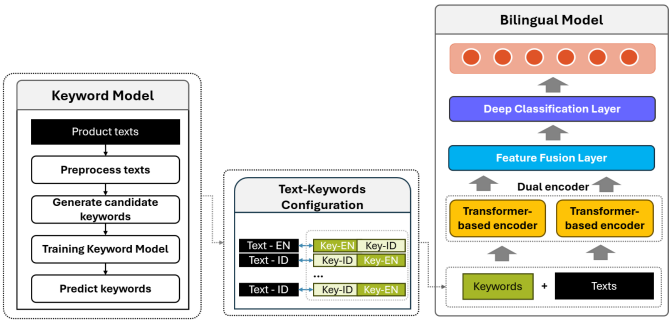

Fig. 1. Overview of the methodology for text and mixed keyword integration in bilingual product classification, consisting of: a keyword prediction model, sample configuration, and a bilingual classification model for predicting product categories

As such, integrating keywords with the original text can intensify contextual signals, enrich textual representation, and improve model performance, especially when applied to domain-specific datasets. In trade data, product declarations are typically brief, contain mixtures of product attributes, lack grammatical rules, and are intricate to understand the context. The use of keywords in classification tasks is expected to enrich text representation by providing more informative features for effective models' understanding. Keywords integration has the prospect to help in preserving important semantic information from minority classes to improve model generalisation, although this approach does not serve as a primary solution for data imbalance. In bilingual trade datasets, product descriptions are grouped with their respective categories, but they may lack shared semantic alignment due to language-specific differences. Additional keywords become essential to provide higher-level information, acting as semantic anchors that bridge and generalise across category labels.

We extract keywords from the product description and subsequently use them to guide the model and facilitate product representations across different languages, expecting the products belonging to the same category to be similarly represented. Our approach differs from previous studies, such as Onan [31], where keywords were used to replace and represent the original texts. This study aligns more closely with the work of Zhou [25], in particular with the strategy of expanding input features by retrieving supplementary knowledge from external open knowledge bases (e.g., DBpedia), which is then fused with the original text to enrich the representation. Our approach is framed within the scope of feature engineering, integrating keywords with primary product descriptions as input for training bilingual classification models. Transformer-based language models are equipped with the capability to recognise nuanced semantic differences in product descriptions, and a neural network-based architecture was employed to learn joint representations from both feature sources, producing semantically aligned product output.

The proposed method integrates product descriptions and extracted keywords into a deep learning architecture for bilingual trade product classification, as illustrated in Fig. 1. It

TABLE I
MODEL HYPERPARAMETERS AND TRAINING CONFIGURATION.

Hyperparameter	Value
Embedding dimension	768
Max length tokens	128
Epoch	3
Batch size	16
Learning rate	2×10^{-5}
Dropout	0.2
Optimizer	AdamW
Activation Function	ReLU

comprises three main components. The early phase addresses keyword extraction, keyword modelling, and text-keyword sample configuration. Next, a bilingual classification model is subsequently constructed through a dual encoder strategy, a feature fusion mechanism, and a deep classification layer.

A. Problem Formulation

We formulate the bilingual-text classification problem as follows. Let $L \in \{l_{\mathrm{EN}}, l_{\mathrm{ID}}\}$ be a set of languages where l_{EN} represents product description in English and l_{ID} represents Indonesian, $X = \{x_1^{g_1}, x_2^{g_2}, \ldots, x_z^{g_z}\}$ be a set of product descriptions where $g_i \in L$. Let $C = \{0, 1, \ldots, c\}$ be the categories (i.e., HS codes) of product descriptions. The dataset is denoted by $D = \{(x_1^{g_1}, y_1), (x_2^{g_2}, y_2), \ldots, (x_n^{g_n}, y_n)\}_{i=1}^{n}$ which consists of n samples and y_i an associated label where $y_i \in C$. Let $K_i = \{k_1^{g_1}, k_2^{g_2}\}$ represent the keywords for each product category y_i. Given a training dataset $D = \{(X_1, K_1, y_1), (X_2, K_2, y_2), \ldots, (X_N, K_N, y_N)\}_{i=1}^{N}$ the goal of our task is to train a model F that predicts HS code $\hat{y}_i = F((X_i, K_i), \theta)$ where θ is a set of model parameters, for a given product with a description X_i and the extracted keywords K_i. Implementation details and source code: https://github.com/anon-user-temp/mixed-keywords-hscode.

B. Keyword Extraction and Modelling

To extract keywords using supervised approaches, it relies on documents that are explicitly labelled with keywords [32], while unsupervised methods utilise the statistical and structural characteristics inherent within the text itself. Unsupervised methods are used to extract keywords, such as Term Frequency [33], TF-IDF [34], YAKE [35], and based on Transformer models, such as KeyBERT [36]. Our study employed a hybrid approach, utilising TF for unsupervised techniques to extract category-specific keywords. These frequently used terms in trade data are assumed to represent each product category best; these keywords were used to annotate the product transactions, rather than embedding them within the main text. Supervised learning was used to train the keywords classifier predicting the keywords from product descriptions needed for configuring training samples.

TF is considered relevant and important for keyword extraction in trade data, where specific terms are more common to represent theme of the products. These high-frequency terms are assumed to be the main product or representation of the declaration. To optimise obtaining the most relevant information in the product, the extraction information or keywords started with text preprocessing, which includes converting text to lowercase and removing duplicates, punctuation, numbers, mixed strings, and extra product details. Word frequency is then computed for each class, and the highest-ranking term is selected as the representative keyword. Following Zaraini's rationale [37], a keyword-per-language strategy was used to minimise semantic noise and ambiguity, resulting in one representative keyword for each product category in both English and Indonesian. These two keywords are combined to form what we refer to as the Mixed Keywords (MK) strategy. The MK strategy integrates the top keyword from each language, enabling cross-lingual alignment between product descriptions and their associated keywords. These combined keywords are then used to annotate the training data for the keyword prediction model. The detailed procedure is outlined in Algorithm 1. Having obtained all of the MK labels to represent the product categories, the keyword classifier is designed to predict the product descriptions and their representative keywords. This method used a discriminative modelling approach to assign each text instance a specific keyword label. The classifier employed DistilBERT fine-tuned on an annotated dataset of product descriptions paired with their corresponding keywords. Training followed standard sequence classification procedure [22], [38], and the model parameters are presented in Table I. During inference, predicted MK from the keyword model were combined with their corresponding product descriptions to create paired inputs. These pairs served as inputs for the next phase, in which a bilingual product classification model was trained. The integrated inputs are subsequently processed using deep learning methods to optimise classification performance.

C. Dual Encoder

A separate dual-encoder framework was used to generate representations from sample training consisting of paired product descriptions and MK, which were utilised to fine-tune DistilBERT [39] to generate 768-dimensional embeddings from each encoder. This process is formalised as follows: the product text sequence X is encoded into a latent representation H_T via the encoder function f_T. Similarly, the keyword sequence K is encoded by the keyword encoder function f_K, resulting in the representation H_K. Let n and m denote the maximum sequence lengths of the text and keyword, respectively, and let d represent the embedding dimension. The resulting embeddings are defined as: $H_T = f_T(X) \in \mathbb{R}^{n \times d}$ for product representation and $H_K = f_K(K) \in \mathbb{R}^{m \times d}$ for keyword representation.

D. Fusion Layer

For the bilingual classifier, this study employed an intermediate fusion strategy, where the concatenation of product description and MK was performed after the encoding phase, rather than at the raw text level [40]. This decision was motivated to produce richer and more expressive representations to integrate the distinct representations of product

Algorithm 1 Mixed Keywords Extraction Algorithm

Input: $Docs_{EN}$, $Docs_{ID}$, product_categories
Output: MK (Mixed keywords for each class)

for each c in product_categories **do**
  // Aggregate EN keyword candidate
  $V_{EN} \leftarrow \{\}$
  for d in $Docs_{EN}[c]$ **do**
    $v \leftarrow$ preprocess(d)
    $V_{EN} \leftarrow$ merge(V_{EN}, v)
  end for
  $K_{EN} \leftarrow$ sort(V_{EN})
  // Aggregate ID keyword candidate
  $V_{ID} \leftarrow \{\}$
  for d in $Docs_{ID}[c]$ **do**
    $v \leftarrow$ preprocess(d)
    $V_{ID} \leftarrow$ merge(V_{ID}, v)
  end for
  $K_{ID} \leftarrow$ sort(V_{ID})
  $k_{EN} \leftarrow$ top(K_{EN}, 1) *// EN keyword*
  $k_{ID} \leftarrow$ top(K_{ID}, 1) *// ID keyword*
  $MK[c] \leftarrow [k_{EN}, k_{ID}]$ *// Mixed keywords*
end for
return MK

text and keywords. To form a unified representation, the contextual embeddings of the $[CLS]$ token from both the text and keyword encoders are concatenated, resulting in a composite vector h_c. The operation is formalised as follows: $h_T = H_T\,[\text{CLS}]$ and $h_K = H_K\,[\text{CLS}]$ are the [CLS] representations from product descriptions and keywords respectively, and $h_c = \text{Concat}\,[h_T; h_K] \in \mathbb{R}^{2d}$ is their concatenated joint representation.

The 1536-dimensional concatenated embeddings were processed through a feed-forward layer that projected them to 768 dimensions. This process was followed by layer normalisation to stabilise training by mitigating gradient-related issues and a ReLU activation to introduce non-linearity to learn complex feature interactions, and dropout layers were used to prevent overfitting. The full process is formally described as follows: $f_{\text{Fusion}} : \mathbb{R}^{2d} \rightarrow \mathbb{R}^{d}$ and the output is $f_{\text{Fusion}}(h_c) = \text{ReLU}(\text{LayerNorm}(W_f h_c + b_f))$.

E. Deep Classification Layer

A deep classifier processed the combined text-keyword embedding through three fully connected layers, each activated by ReLU and followed by dropout to reduce overfitting. Training employed cross-entropy loss with label smoothing, replacing hard target labels with soft probability distributions to reduce overconfident predictions and improve robustness to noisy labels [41], with the parameter set to 0.1. Together, these elements produce a more generalisable and stable model. The full classification pipeline is described as follows: $a^{[0]} = f_{\text{Fusion}}(h_c)$. Subsequently, each layer computes $z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]}, \quad \text{for} \quad l = \{1, 2, 3\}$, followed by ReLU activation $a^{[l]} = \text{ReLU}(z^{[l]})$. The final output $P\,(y|X, K; \theta) = \text{Softmax}\left(z^{[3]}\right)$ provides class probabilities.

TABLE II
DATASET SIZE IN ENGLISH AND INDONESIAN LANGUAGE.

Language	Rows
English	34,937
Indonesian	19,920

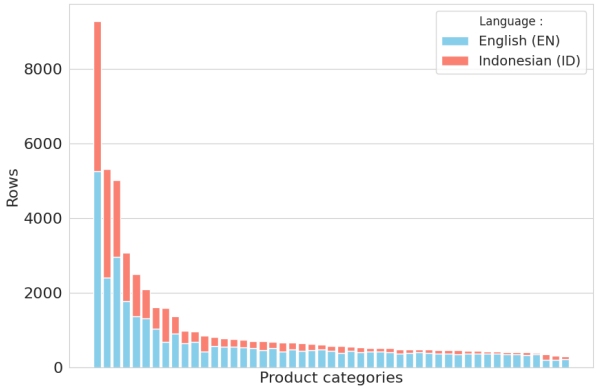

Fig. 2. Distribution of bilingual data across product categories

III. EXPERIMENTS

A. Datasets

This experiment utilised international trade datasets in English and Indonesian languages. The English dataset is publicly available and has been widely referenced in existing research [8]–[11]. These datasets are sourced from publicly available data on the Internet [42]. The Indonesian dataset was obtained from an Indonesian financial institution that has been cited in the literature [5], [43].

The integration of high-resource English datasets with available Indonesian datasets resulted in the matching of 49 distinct product categories, which were identified through the usage of six-digit HS codes. The overview of the dataset statistics, including the total number of samples, is provided in Table II, with the dataset distribution shown in Fig. 2. The keyword extraction and modelling method was applied to a random sample of half of each dataset, ensuring all class categories are represented. Table III shows examples of product descriptions and their predicted keywords. The performance of the models was evaluated using 10-fold cross-validation (CV-10) using the remaining split of datasets configured in pairs, consisting of product description and the predicted MK, comparing the performance of the proposed model to the baseline models. An ablation study was conducted to evaluate bilingual classification performance of the respective contributions of product descriptions and keywords individually. Additionally, an evaluation scenario was designed to reflect conditions of limited target-language data availability, particularly for non-English datasets. The low-resource setting limited the Indonesian training samples to four levels: 0, 1,000, 5,000,

TABLE III
AN EXAMPLE OF BILINGUAL PRODUCT DESCRIPTIONS WITH EXTRACTED MIXED KEYWORDS.

Text - EN	Text - ID	Mixed Keywords
gu06 umbrella alluminium with stone base 60 kgs for outdoor furniture	aksesoris payung taman britam 064 38249	['payung', 'umbrella']

and 10,000. A train-test-holdout split strategy was used for consistent and reliable evaluation.

B. Experimental Settings

1) Model Baselines: To evaluate the effectiveness of the proposed method, we conducted a comparative analysis using several multilingual language models trained with a similar configuration to the proposed method. The baseline models are as follows:

- Multilingual BERT [22]: mBERT was pretrained on corpora from 104 languages, enabling it to produce language-agnostic representations that generalise well across multilingual and bilingual settings.
- XLM-R [44]: XLM-RoBERTa was built on the RoBERTa architecture, an optimised variant of BERT to generate language-agnostic representations and demonstrates superior results on multiple cross-lingual benchmarks.
- Multilingual DeBERTaV3 [45]: mDeBERTaV3 is a multilingual extension of the DeBERTa architecture, outperforming XLM-R by 3.6% across 15 languages.
- Sentence BERT [46]: The model used in this experiment was paraphrase-multilingual-MiniLM-L12-v2, which supports cross-lingual semantic similarity tasks and can be fine-tuned on multilingual datasets.
- IndoBERT [47]: IndoBERT is a transformer-based model pretrained exclusively on Indonesian datasets and serves as a strong baseline for comparision and classifying using Indonesian text.

2) Experimental Setup: All experiments and data analysis procedures were executed on P100 GPUs with 16 GB of RAM. The pre-trained models were obtained via the HuggingFace library [48], which preferred the base version and uncased models. The embedding size of all models was set to 768, with the input text or keyword limited to a fixed sequence length of 128 tokens per sample, initiating the truncation and padding process to reach this length. The batch size for training was set to 16, balancing computational efficiency and model performance. The training process employed a learning rate of 2×10^{-5} and was conducted over three epochs, which are sufficient for fine-tuning in most tasks. Optimisation used the cross-entropy loss function in conjunction with the AdamW optimiser [22].

3) Evaluation Metrics: The experiment evaluated the model performance using accuracy, weighted precision, recall, and F1-score to handle class imbalance. Additionally, cosine similarity was used to assess semantic alignment between embeddings of cross-lingual product descriptions.

IV. RESULTS AND DISCUSSION

A. Model Performance

The proposed method achieved the highest performance, with an accuracy of 97% and an F1-score of 96.8%, as shown in Table IV. The method also demonstrates performance gain to achieve an average improvement of 17.50% in accuracy and significantly outperforming all baselines; in particular, the best baseline model was achieved by the mBERT model. The F1-score difference was statistically significant as shown in Table IV.

The empirical results are consistent with previous research [25], [31], which highlights the importance of using keyword features in classification tasks. Furthermore, our approach advances previous studies employed in bilingual models by configuring cross-lingual keywords paired with the main text to support model understanding, which in turn improves the bilingual classification task.

B. Embeddings Alignment

PCA was used to visualise in low-dimensional embeddings to provide insight into the cross-lingual alignment. We only compared the embeddings generated by the XLM-R model and the proposed method. Fig. 3.a shows a scattered fine-tuned XLM-R, where both points representing across languages were unclustered. On the other hand, the proposed method generated data points that present a structure overlap between languages, implying improved cross-lingual alignment, illustrated in Fig. 3.

In order to further examine the result, a quantitative evaluation was performed the cross-lingual alignment by calculating the average similarity between embeddings in different languages across all product categories. Fig. 3.b visualises these scores as a heatmap, where diagonal values indicate intra-class similarity and darker blue shades indicate higher similarity. The metric is calculated as: $\sigma_c^{\text{mean}} = \frac{1}{n_c m_c}\sum_{i=1}^{n_c}\sum_{j=1}^{m_c}\text{sim}_{\cos}{}_{ij}^{c}$. Our proposed method achieved better cross-lingual alignment than XLM-R, with an intra-class similarity of 0.92. This result demonstrates that integrating keywords improves semantic alignment, encouraging product descriptions within the same class to map closely in the embedding space across languages.

We also performed a case study to evaluate the bilingual model by submitting a query of "rattan table". The result shows that the model retrieved the two most relevant products in both languages, correctly mapped to HS code 940381 for furniture made of rattan or bamboo [3], as seen in Fig. 4. The bilingual product search experiment has resulted in cross-

TABLE IV
COMPLETE 10-CV TEST RESULTS OF ALL MODELS. THE PROPOSED MODELS CONSISTENTLY OUTPERFORM BASELINE MODELS ACROSS EVALUATION METRICS.

Models	Accuracy	Precision	Recall	F1-Score
Baseline:				
Multilingual BERT	0.820 ± 0.006	0.819 ± 0.007	0.820 ± 0.006	0.817 ± 0.007
XLM-R	0.802 ± 0.007	0.801 ± 0.008	0.802 ± 0.007	0.798 ± 0.007
Multilingual DeBERTa	0.794 ± 0.008	0.793 ± 0.008	0.794 ± 0.008	0.788 ± 0.008
SBERT	0.794 ± 0.004	0.792 ± 0.005	0.794 ± 0.004	0.788 ± 0.005
Indo BERT	0.791 ± 0.005	0.790 ± 0.005	0.791 ± 0.005	0.787 ± 0.006
Proposed method:				
MK + DistilBERT	0.970 ± 0.003	0.969 ± 0.003	0.970 ± 0.003	0.968 ± 0.003

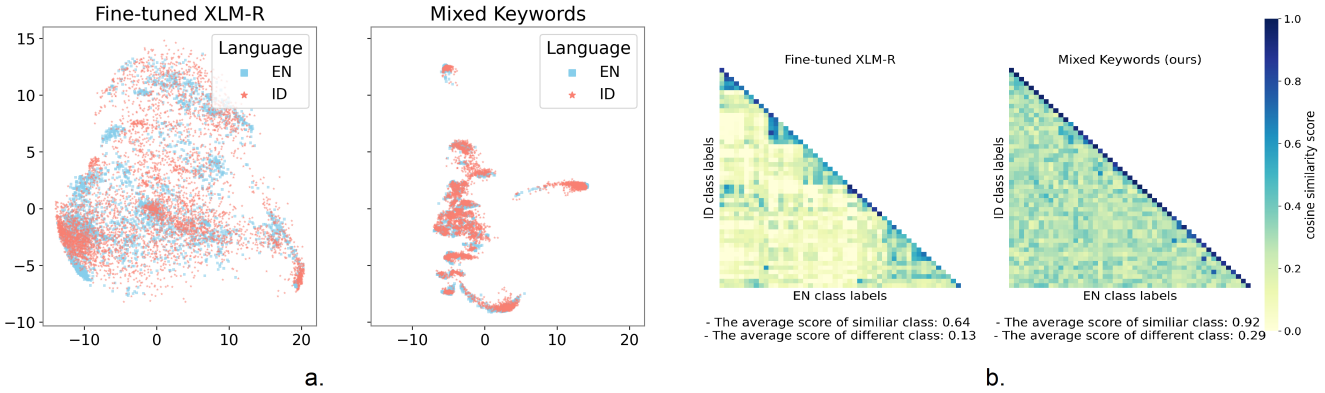

Fig. 3. a. PCA visualisation presents cross-lingual alignment in of English (EN) and Indonesian (ID) generated by fine-tuned XLM-R and our proposed method. b. Heatmap shows similarity scores between inter-class and intra-class labels (the diagonal) across ID (y-axis) and EN (x-axis) generated by the fine-tuned XLM-R and the proposed method.

lingual product results and captured material information embedded in the queries.

C. Model Performance in Low-resource Setting

Given the scarcity of trade data in non-English languages, and in this study, the Indonesian language was considered a low-resource language. In this experiment, we assumed only to train the model on English training samples and the predicted keyword, or MK. The result shows that the model achieved a notable F1-score of 94.79%. Additional Indonesian training samples used for model training improved the model's performance, achieving a significant F1-score of 96.22%, as shown in Fig. 5. This finding demonstrates that combining high-resource languages, such as English, with cross-lingual keywords is sufficient to achieve strong model performance, particularly when the target language is limited or non-existent.

Our implementation of the DistilBERT model [39] in dealing with low-resource language expands on previous research using BERT-based models in low-resource settings [49], [50]. DistilBERT is a language model trained on monolingual corpora in English; yet it demonstrates effectiveness on domain-specific data, such as bilingual trade datasets. The model also offers lower computation due to its language model size compared to other multilingual models.

TABLE V
THE ABLATION RESULTS COMPARE PRODUCT ONLY (PO) AND KEYWORD ONLY (KO) MODELS, WITH THE LOWER PANEL SHOWING THE PERFORMANCE GAINS OF THE MIXED KEYWORD (MK) METHOD OVER BOTH BASELINES.

	Accuracy	F1-Score
Models:		
PO	0.816 ± 0.005	0.813 ± 0.005
KO	0.904 ± 0.000	0.865 ± 0.001
MK	0.970 ± 0.003	0.968 ± 0.003
Performance Gains:		
MK vs. PO	+15.88%	+16.04%*
MK vs. KO	+6.83%	+10.65%*

D. Ablation Study

In this ablation study, comparisons were performed between the proposed model and two variant models when trained only with the product description, specifically Product Only (PO), and the model trained only with predicted keywords, specifically Keyword Only (KO). The results show that the proposed method surpasses both PO and KO by 15.88%–16.04% and KO by 6.83%–10.65%, respectively, evaluated on 10-CV, as shown in Table V. These results reinforce the concept that relying solely on texts or keywords is insufficient to achieve a high-performance model, and underscore the value of combining keywords with the main product text.

V. Conclusion

In conclusion, our bilingual models have addressed the issue of bilingual language presented in trade product declarations by developing a bilingual classification model that incorporates cross-lingual keyword features used in the bilingual model training. Mixed keyword strategies facilitated the capture of product terms, addressed language variations, and aligned bilingual trade data. The model successfully enhances the embedding alignment of bilingual languages, demonstrates strong effectiveness in low-resource target language settings, and surpasses the other baseline models, indicating potential applicability in real-world applications. This study was conducted using English and Indonesian data, and the limited availability of the other languages trade datasets presents a generalisation challenge. However, this also suggests a promising avenue for future research to expand the approach to broader multilingual settings.

Acknowledgment

Yuhua Li was partially supported by the EPSRC through the AI Hub in Generative Models [grant number EP/Y028805/1].

References

[1] H. Tenzer, S. Terjesen, and A.-W. Harzing, "Language in International Business: A Review and Agenda for Future Research," Manag Int Rev, vol. 57, no. 6, pp. 815–854, Dec. 2017, doi: 10.1007/s11575-017-0319-x.

[2] J. Lohmann, "Do language barriers affect trade?," Economics Letters, vol. 110, no. 2, pp. 159–162, Feb. 2011, doi: 10.1016/j.econlet.2010.10.023.

[3] World Customs Organization, "Harmonized System Compendium – 30 Years On," 2018. [Online]. Available: https://www.wcoomd.org/-/media/wco/public/global/pdf/topics/nomenclature/activities-and-programmes/30-years-hs/hs-compendium.pdf

[4] D. C. Novith, "Harmonized System Code Recommendation: A Multi-Class Classification Model," Jurnal BPPK: Badan Pendidikan dan Pelatihan Keuangan, vol. 17, no. 3, pp. 1–11, 2024.

[5] I. G. Y. Paramartha, I. Ardiyanto, and R. Hidayat, "Developing Machine Learning Framework to Classify Harmonized System Code. Case Study: Indonesian Customs," in 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), Apr. 2021, pp. 254–259. doi: 10.1109/EIConCIT50028.2021.9431888.

[6] DGCE. Pemberitahuan Pabean Impor, May 2009.

[7] A. Grainger, "Customs Tariff Classification and the Use of Assistive Technologies," World Customs Journal, vol. 18, no. 1, pp. 3–32, Mar. 2024, doi: 10.55596/001c.116525.

[8] D. Ruder, "Application of Machine Learning for Automated HS-6 Code Assignment," Master Thesis, Tallinn University of Technology, Tallinn, 2020.

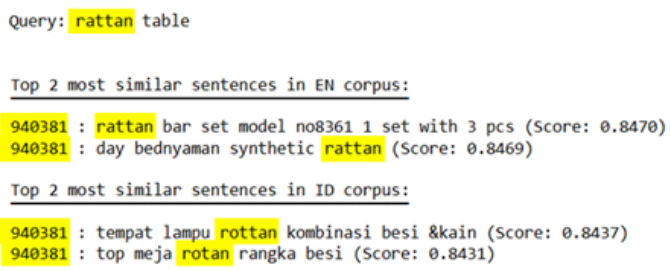

Fig. 4. Retrieved product descriptions using bilingual queries.

[9] J. Luppes, "Classifying Short Text for the Harmonized System with Convolutional Neural Networks," Radboud University, 2019.

[10] Shubham, A. Arya, S. Roy, and S. Jonnala, "An Ensemble-Based Approach for Assigning Text to Correct Harmonized System Code," in 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Jan. 2023, pp. 35–41. doi: 10.1109/AISC56616.2023.10085512.

[11] M. Spichakova and H.-M. Haav, "Application of Machine Learning for Assessment of HS Code Correctness," Balt. J. Mod. Comput., vol. 8, no. 4, pp. 698–718, 2020, doi: 10.22364/bjmc.2020.8.4.13.

[12] R. R. de Lima, A. M. R. Fernandes, J. R. Bombasar, B. A. da Silva, P. Crocker, and V. R. Q. Leithardt, "An Empirical Comparison of Portuguese and Multilingual BERT Models for Auto-Classification of NCM Codes in International Trade," Big Data and Cognitive Computing, vol. 6, no. 1, 2022, doi: 10.3390/bdcc6010008.

[13] E. Lee, S. Kim, S. Kim, S. Jung, H. Kim, and M. Cha, "Explainable Product Classification for Customs," ACM Trans. Intell. Syst. Technol., vol. 15, no. 2, p. 25:1-25:24, Feb. 2024, doi: 10.1145/3635158.

[14] I. Marra de Artiñano, F. Riottini Depetris, and C. Volpe Martincus, "Automatic Product Classification in International Trade: Machine Learning and Large Language Models," Inter-American Development Bank, Working Paper IDB-WP-01494, 2023. doi: 10.18235/0005012.

[15] M. Liao, L. Huang, J. Zhang, L. Song, and B. Li, "Enhanced HS Code Classification for Import and Export Goods via Multiscale Attention and ERNIE-BiLSTM," Applied Sciences (Switzerland), vol. 14, no. 22, Nov. 2024, doi: 10.3390/app142210267.

[16] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[17] A. King, "Using Machine Translation to Augment Multilingual Classification," May 09, 2024, arXiv: arXiv:2405.05478. doi: 10.48550/arXiv.2405.05478.

[18] S. Ruder, I. Vulić, and A. Søgaard, "A Survey Of Cross-lingual Word Embedding Models," jair, vol. 65, pp. 569–631, Aug. 2019, doi: 10.1613/jair.1.11640.

[19] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, "Unsupervised Machine Translation Using Monolingual Corpora Only," Apr. 13, 2018, arXiv: arXiv:1711.00043. Accessed: Oct. 11, 2024. [Online]. Available: http://arxiv.org/abs/1711.00043

[20] E. Lehmann, A. Simonyi, L. Henkel, and J. Franke, "Bilingual Transfer Learning for Online Product Classification," in Proceedings of Workshop on Natural Language Processing in E-Commerce, H. Zhao, P. Sondhi, N. Bach, S. Hewavitharana, Y. He, L. Si, and H. Ji, Eds., Barcelona, Spain: Association for Computational Linguistics, Dec. 2020, pp. 21–31. Accessed: Sep. 28, 2024. [Online]. Available: https://aclanthology.org/2020.ecomnlp-1.3

[21] M. Artetxe and H. Schwenk, "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond," Transactions of the Association for Computational Linguistics, vol. 7, pp. 597–610, Nov. 2019, doi: 10.1162/tacl_a_00288.

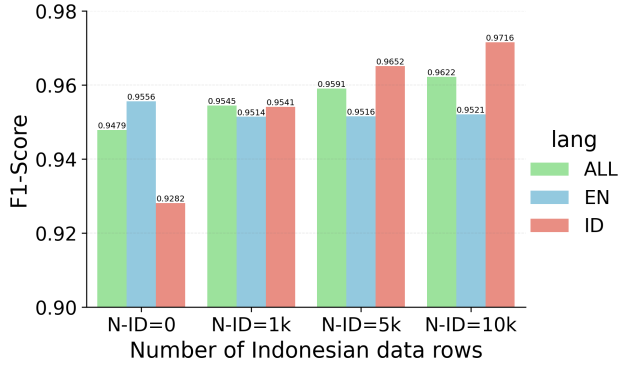

Fig. 5. The model's performance on bilingual classification in an incremental training sample setting.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in 2019 Conference of The North American Chapter of The Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019), VOL. 1, Stroudsburg: Assoc Computational Linguistics-Acl, 2019, pp. 4171–4186. Accessed: Dec. 21, 2023. [Online]. Available: https://www.webofscience.com/wos/woscc/full-record/WOS:000900116904035

[23] P. Chuanakrud, T. Leelanupab, C. Damrongrat, and N. Kanungsukkasem, "Keyword-Text Graph Representation for Short Text Classification," in 2021 13th International Conference on Information Technology and Electrical Engineering (ICITEE), Oct. 2021, pp. 24–29. doi: 10.1109/ICITEE53064.2021.9611935.

[24] Z. Tang, D. Dai, Z. Chen, and T. Chen, "Short Text Classification Combining Keywords and Knowledge," in 2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Jan. 2022, pp. 662–665. doi: 10.1109/ICCECE54139.2022.9712673.

[25] X. Zhou et al., "A hybrid classification method via keywords screening and attention mechanisms in extreme short text," Intell. Data Anal., vol. 27, no. 5, pp. 1331–1345, Jan. 2023, doi: 10.3233/IDA-220417.

[26] A. E. Blanchard et al., "A Keyword-Enhanced Approach to Handle Class Imbalance in Clinical Text Classification," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 6, pp. 2796–2803, Jun. 2022, doi: 10.1109/JBHI.2022.3141976.

[27] M. Abulaish, M. Fazil, and M. J. Zaki, "Domain-Specific Keyword Extraction Using Joint Modeling of Local and Global Contextual Semantics," ACM Trans. Knowl. Discov. Data, vol. 16, no. 4, p. 70:1-70:30, Jan. 2022, doi: 10.1145/3494560.

[28] L.-C. Chen, "An extended TF-IDF method for improving keyword extraction in traditional corpus-based research: An example of a climate change corpus," Data & Knowledge Engineering, vol. 153, p. 102322, Sep. 2024, doi: 10.1016/j.datak.2024.102322.

[29] T. Nomoto, "Keyword Extraction: A Modern Perspective," SN COMPUT. SCI., vol. 4, no. 1, p. 92, Dec. 2022, doi: 10.1007/s42979-022-01481-7.

[30] W. Zhang, Y. Lu, B. Dubrov, Z. Xu, S. Shang, and E. Maldonado, "Deep Hierarchical Product Classification Based on Pre-Trained Multilingual Knowledge," Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2021.

[31] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," Expert Systems with Applications, vol. 57, pp. 232–247, Sep. 2016, doi: 10.1016/j.eswa.2016.03.045.

[32] S. Siddiqi and A. Sharan, "Keyword and Keyphrase Extraction Techniques: A Literature Review," International Journal of Computer Applications, vol. 109, no. 2, pp. 18–23, Jan. 2015.

[33] N. Firoozeh, A. Nazarenko, F. Alizon, and B. Daille, "Keyword extraction: Issues and methods," Natural Language Engineering, vol. 26, no. 3, pp. 259–291, May 2020, doi: 10.1017/S1351324919000457.

[34] B. Hashemzahde and M. Abdolrazzagh-Nezhad, "Improving keyword extraction in multilingual texts," IJECE, vol. 10, no. 6, p. 5909, Dec. 2020, doi: 10.11591/ijece.v10i6.pp5909-5916.

[35] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," Information Sciences, vol. 509, pp. 257–289, Jan. 2020, doi: 10.1016/j.ins.2019.09.013.

[36] M. Grootendorst et al., MaartenGr/KeyBERT: v0.8. (Sep. 29, 2023). Zenodo. doi: 10.5281/ZENODO.8388690.

[37] M. N. Zaraini and N. Yusop, "How Many Keywords are Enough? Determining the Optimal Top-K for Educational Website Classification," IJRISS, vol. IX, no. VI, pp. 1574–1586, Jul. 2025, doi: 10.47772/ijriss.2025.906000126.

[38] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," in Chinese Computational Linguistics, vol. 11856, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds., in Lecture Notes in Computer Science, vol. 11856. , Cham: Springer International Publishing, 2019, pp. 194–206. doi: 10.1007/978-3-030-32381-3_16.

[39] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a Distilled Version of Bert: Smaller, Faster, Cheaper and Lighter," Feb. 29, 2020, arXiv: arXiv:1910.01108. doi: 10.48550/arXiv.1910.01108.

[40] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," Machine Vision and Applications, vol. 32, no. 6, p. 121, Nov. 2021, doi: 10.1007/s00138-021-01249-8.

[41] M. Lukasik, S. Bhojanapalli, A. K. Menon, and S. Kumar, "Does label smoothing mitigate label noise?," Mar. 05, 2020, arXiv: arXiv:2003.02819. doi: 10.48550/arXiv.2003.02819.

[42] Enigma, "US Imports - Automated Manifest System (AMS) Shipments 2018." Jan. 01, 2018. [Online]. Available: https://aws.amazon.com/

[43] P. Harsani, A. Suhendra, L. Wulandari, and W. C. Wibowo, "A Study Using Machine Learning with Ngram Model in Harmonized System Classification," Journal of Advanced Research in Dynamical and Control Systems, vol. 12, no. 6 Special Issue, pp. 145–153, 2020, doi: 10.5373/JARDCS/V12SP6/SP20201018.

[44] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[45] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," presented at the International Conference on Learning Representations, arXiv, Mar. 2023. doi: 10.48550/arXiv.2111.09543.

[46] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2019, pp. 3982–3992. doi: 10.18653/v1/d19-1410.

[47] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in Proceedings of the 28th International Conference on Computational Linguistics, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[48] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Q. Liu and D. Schlangen, Eds., Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[49] J. C. B. Cruz and C. Cheng, "Establishing Baselines for Text Classification in Low-Resource Languages," May 05, 2020, arXiv: arXiv:2005.02068. doi: 10.48550/arXiv.2005.02068.

[50] X. Li, Z. Li, J. Sheng, and W. Slamu, "Low-Resource Text Classification via Cross-lingual Language Model Fine-tuning," in Proceedings of the 19th Chinese National Conference on Computational Linguistics, M. Sun, S. Li, Y. Zhang, and Y. Liu, Eds., Haikou, China: Chinese Information Processing Society of China, Oct. 2020, pp. 994–1005. Accessed: Jul. 23, 2025. [Online]. Available: https://aclanthology.org/2020.ccl-1.92/