A Hybrid Semantic-Statistical Feature Fusion Framework for Bilingual Text Classification on Multilingual Big Data Corpora

Kavitha M[1], Dr. Purohit Shrinivasacharya[2], Dr. Y S Nijagunarya[3]
Department of Computer Science and Engineering,
Siddaganga Institute of Technology, Tumkur-572103, Karnataka, India[1,3]
Visvesvaraya Technological University, Belagavi-590018, Karnataka, India[1,2,3]
Department of Information Science and Engineering, Siddaganga Institute of Technology,
Tumkur-572103, Karnataka, India[2]

***Abstract*—As the volume of scientific information is growing exponentially in several languages, there is a need for practical and scalable bilingual classification systems for large aligned scientific text corpora. To address this challenge, this study makes two key contributions - first, a large-scale bilingual English–Hindi aligned arXiv scientific text corpus, providing a structured resource for multilingual scientific text analytics and classification research is constructed. Second, the study proposes a bilingual scientific text classification framework and performs a rigorous experimental evaluation using three strong multilingual transformer models, namely Mini Language Model-12 Layers (MiniLM-L12), Multilingual Bidirectional Encoder Representations from Transformers (mBERT), and Cross-lingual Language Model-Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach (XLM-RoBERTa), on the developed English–Hindi aligned arXiv big data corpus. English and Hindi summaries are categorized independently to investigate the performance trade-offs in each language. The proposed hybrid MiniLM-L12 + Multi-Layer Perceptron (MLP) architecture enhance the classification capability through the integration of statistical feature design with contextual sentence embeddings. Empirical analysis indicates that the proposed bilingual classification framework consistently outperforms the baseline transformer-only models, achieving a higher accuracy of 95.56% and weighted F1-score of 95.31%, while maintaining computational efficiency. The findings emphasize the effectiveness of hybrid representation learning for bilingual big data corpora and provide practical insights for scalable multilingual scholarly text analytics.**

***Keywords*—Text classification (bilingual); aligned big data corpora; multilingual transformers; mini language model-12 layers; multilingual bidirectional encoder representations from transformers; cross-lingual language model-robustly optimized bidirectional encoder representations from transformers pretraining approach; multi-layer perceptron; hybrid deep learning**

I. Introduction

Due to the continued digitization of scientific literature, the amount of scholarly literature published in different languages has increased dramatically. Although English continues to predominate in scientific communication, much of the research is reported in regional and low resource languages, especially in the present research communities. For automated information organization, retrieval, and large-scale text analytics, where conventional monolingual classification systems are ineffective, this multilingual expansion presents serious challenges. Bilingual and cross-lingual techniques for text classification, capable of operating on aligned multilingual corpora with high precision and computational efficiency are thus highly essential. Recent developments highlight transformer-based language models as the dominant methodology in multilingual natural language processing. Multilingual models like mBERT, XLM-RoBERTa, and models like MiniLM are trained to learn common semantic representations across languages and have shown strong performance on diverse multilingual NLP tasks. However, these models exhibit notable differences in terms of model size, training and inference cost, and scalability. These computational considerations are essential in large-scale text classification with a bilingual corpus, particularly in a scientific context. Furthermore, relying solely on deep transformer architectures may limit the exploitation of complementary statistical characteristics inherent in scientific texts.

Hybrid architectures offer promising alternative which can enhance efficiency and yet provide competitive classification performance by jointly using rich semantic representations together with auxiliary features using lightweight neural classifiers. Despite their potential, such approaches have received limited attention in the context of bilingual scientific text classification, particularly for aligned English-Hindi corpora.

In the current work, a comprehensive bilingual experimental study on an aligned English-Hindi arXiv scientific summary dataset is conducted. Each language is used separately and text classification experiments are carried out so that fair language-wise assessment is achieved. This study benchmark widely used multilingual transformer models, including MiniLM-L12, mBERT, and XLM-RoBERTa, and propose a hybrid MiniLM-L12 + MLP architecture that fuses contextual sentence embeddings with statistical features for enhanced performance. Experimental findings indicate that the proposed hybrid model delivers enhanced efficiency and faster inference while achieving competitive or improved accuracy and weighted F1-scores relative to larger multilingual transformer models in both languages.

The key findings of this study are summarized as follows:

- Monolingual arXiv dataset sampled to Bilingual dataset through effective translation of title and summary attributes to Hindi to obtain an aligned multi-class arXiv big data corpus.

- A comprehensive bilingual test of MiniLM-L12, mBERT, and XLM-RoBERTa on aligned English-Hindi arXiv scientific data.
- A new hybrid MiniLM-L12 + MLP model that is effective to combine contextual embeddings with statistical features in bilingual text classification.
- Analysis of language-wise performance in terms of accuracy, macro precision, macro recall, macro F1, micro precision, micro recall, micro F1, weighted precision, weighted recall and weighted F1-score measures.

The study includes valuable information regarding the effective design of bilingual text classification and demonstrates that hybrid systems establish an effective tradeoff between the performance and the computational cost of large scale multilingual scientific collections.

II. Related Work

Research on text classification has evolved through interconnected methodological streams that collectively underpin modern multilingual and large-scale classification systems. Early investigations primarily emphasized algorithmic efficiency, feature engineering, and computational scalability. Xu suggested a KNN-optimized classification scheme that minimized the use of pre-defined parameter settings [1], whereas Mao et al. proposed parameter-free compression-based technologies that mitigated the use of explicit modeling assumptions [2]. Niranjan et al. investigated structured pipelines used to categorize questions [3]. In the context of Indic multilingual text classification, Pathak and Jain introduced the μBoost framework, which integrates CatBoost with MuRIL representations to improve abusive comment detection across Indian languages. Their ensemble approach demonstrated the benefits of combining multilingual language models with structured meta-features for robust multilingual classification [4]. These early developments put in place performance standards and computation limits that fueled the transition of neural representation learning.

The development of deep neural networks was a significant breakthrough as it made it possible to extract semantic features automatically. Agbesi et al. and Malik and Jain proposed ensemble convolutional structures and knowledge-infused architectures tailored to increase robustness [5], [6]. Jamshidi et al. suggested hybrid pipelines combining transformer embeddings and sequential models [7], while Padalko et al. and Maham et al. investigated sequential and adversarial modeling for misinformation detection [8], [9]. Gao et al. and Huang et al. studied graph-based structural modeling techniques [10], [11], and Ai et al. developed semi-supervised contrastive multigraph learning [12]. Alternative neural paradigms such as Kolmogorov Arnold networks and bi-attention fusion architectures were also explored to enable flexible and adaptive classification strategies [13], [14]. Together these advances demonstrated the effectiveness of representation learning over handcrafted features, but also increased computational demands.

Transformer architectures later emerged as a dominant paradigm in text classification due to their ability to provide contextualized and generalizable representations. Joshi and Joseph analyzed efficient fine-tuning strategies [15], whereas Zhang applied multilingual multi-feature fusion architectures [16]. Bekamiri et al. presented domain-adapted transformers for patent analytics [17], and Brack et al. studied cross-domain multi-task sequential classification [18]. Jain et al. suggested hierarchical and multimodal embedding strategies on crisis datasets [19]. Nevertheless, transformer-based systems still face challenges related to computational cost, scalability, and fairness, particularly in multilingual and low-resource scenarios.

Different pretrained transformer models have become standardized architectures of cross-lingual classification in multilingual settings. Multilingual BERT (mBERT), which was trained on Wikipedia corpora involving 104 languages, using a shared set of WordPiece vocabulary, showed that transfer across languages could start being observed even without explicit alignment supervision [20]. XLM-RoBERTa (XLM-R), which was pretrained on large-scale CommonCrawl corpora in 100 languages with dynamic masked language modeling, enhanced strength and low-resource generalization with high exposure to data and training strategies [21]. Later, small distilled models like MiniLM-L12 have been created that maintain semantic expressiveness using attention and representation distillation mechanisms [22]. These models represent respectively on the efficiency-capacity axis: mBERT is able to offer consistent multilingual contextualization, XLM-R focuses on large-scale robust representation learning, and MiniLM-L12 focuses on computational efficiency and faster inference. They are appropriate reference frameworks to analyse the trade-offs in multilingual performance given their architectural variety and variations in pretraining scale.

In addition to architectural developments, multilingual studies increasingly address fairness and variability in evaluation. Cross-lingual adaptability has been explored through hybrid multilingual models [23], cross-lingual topic detection [24], and prompt-based zero-shot classification frameworks [25]. Fairness-conscious multilingual assessment has been highlighted in recent studies [26], while empirical variability in low-resource classification scenarios has been emphasized [27]. Similarly, Mehra et al. proposed PhonoBiEmbedNet, a phoneme- and bigram-based embedding framework designed for low-resource spoken word recognition. By incorporating phonetic linguistic representations and pretrained multilingual models through a late-fusion neural architecture, the approach demonstrated improved recognition accuracy under limited training data conditions [28]. These findings highlight that multilingual classification remains a complex optimization task involving linguistic diversity, data bias, and computational constraints.

The applied use of text classification includes biomedical multi-label learning [29], [30], patent analytics using hybrid transformer-graph models [17], [31], [11], and crisis informatics and misinformation detection using sequential and adversarial models [8], [9], [19]. Dataset balancing strategies for multilingual software classification have also been explored [32]. More recently, classification approaches based on large language models have introduced hierarchical taxonomy refinement and causal classification paradigms [33], [34]. Although LLM-driven systems tend to achieve strong performance, they introduce significant computational overhead and

scalability challenges in multilingual large-scale environments.

Even though there is continued improvement, there are three gaps that remain critical in the literature. To begin with, efficient bilingual frameworks that are tuned to the aligned large-scale corpora are underutilized. Second, there are not many multilingual tests with systematic fairness analysis. Third, in the real-world scientific classification environment, empirical research on efficiency-performance trade-offs between compact and large-capacity transformer architectures is under-examined. To overcome these shortcomings, scalable learning of representation should be incorporated with considerations of computational efficiency in a principled manner. By placing small distilled transformer and large-capacity multilingual models in a single bilingual assessment system, the study under consideration is expected to connect the principles of foundational scalability with the contemporary contextual representation learning and to analyze the robustness, efficiency, and fairness between English and Hindi scientific texts classification in a systematic way.

III. Proposed Methodology

This section outlines the suggested hybrid semantic-statistical feature fusion framework for bilingual scientific text classification. First, the construction and preprocessing of an aligned English-Hindi arXiv corpus to provide the semantic consistency between languages is discussed, followed by the design of a computationally efficient classification architecture that combines contextual transformer-based embeddings with statistical lexical information. Then, the mathematical description of the semantic-statistical fusion mechanism and optimisation strategy is outlined. Finally, an algorithm description of the end to end text classification workflow is presented to show how the proposed structure will be implemented.

A. Aligned Bilingual Corpus Construction and Preprocessing Strategy

The experimental evaluation of the multilingual transformer models and the proposed hybrid model is performed on the arXiv scientific corpus, a publicly accessible dataset downloadable from Kaggle: https://www.kaggle.com/datasets/sumitm004/arxiv-scientific-research-papers-dataset. The arXiv corpus is a large collection of research papers from the arXiv preprint repository, which can be used for scientific text analysis, classification, citation prediction, and trend forecasting. The dataset covers a variety of research domains, including Computer Science, Mathematics, Physics, etc. Table I provides the features of the arXiv dataset downloaded from Kaggle.

TABLE I. arXiv Dataset Features Before Preprocessing, Translation and Balancing

Dataset Feature	Description
Size	177 MB
Attributes	id, title, category, category_code, published_date, updated_date, authors, first_author, summary, summary_word_count
Modality Addressed	Text
Linguality Addressed	English
Number of Samples	136,239
Number of Categories	73

The downloaded arXiv dataset's samples featuring all attributes mentioned in Table I before preprocessing, translation and balancing is showcased in Fig. 1.

id	title	category	category_code	published_date	updated_date	authors	first_author	summary	summary_word_count
abs-2501.02725v1	artificial intelligence	artificial_intelligence	cs.AI	06-01-2025	06-01-2025	['Nanthee	'Nantheera Anan	rapid advancer	192
abs-2501.10945v1	gradient base multi	machine_learning	cs.LG	19-01-2025	19-01-2025	['Weiyu Cl	'Weiyu Chen'	multi objective	129
abs-2501.16768v1	generalization multi	machine_learning_{s	stat.ML	28-01-2025	28-01-2025	['Wen We	'Wen Wen'	multiview learı	166
abs-2501.16790v1	exponential family a	machine_learning_{s	stat.ML	28-01-2025	28-01-2025	['Kevin Ch	'Kevin Christian	self attention r	172
abs-2501.16988v2	marginal conditiona	machine_learning_{s	stat.ML	28-01-2025	29-01-2025	['Mohamr	'Mohammad Ka	interpret black	137
abs-2501.17345v1	test conditional mei	machine_learning_{s	stat.ML	28-01-2025	28-01-2025	['Yi Zhang	'Yi Zhang'	conditional me	126
abs-2501.17512v1	survey cluster base	machine_learning_{s	stat.ML	29-01-2025	29-01-2025	['Omar El-	'Omar El-Rifai'	industrial comı	163
abs-2501.17513v1	sequential learning	machine_learning_{s	stat.ML	29-01-2025	29-01-2025	['Elise CrÃ	'Elise CrÃ©pon'	study problem	122

Fig. 1. Representative samples from the original arXiv dataset displaying the metadata and textual attributes used in this study before preprocessing, machine translation into Hindi and class balancing for bilingual scientific text classification.

The dataset is preprocessed using spaCy during which title and summary text fields are cleaned by removing noise, lowercased, lemmatized, and stopwords are eliminated. Extremely short summaries are filtered and category labels are normalized and label-encoded for supervised learning. To get the aligned English-Hindi corpus, Hindi translations are obtained for title and summary attributes using the deep-translator library, which leverages Google Translate's web-based translation service. Two bilingual experts manually evaluated 1000 randomly selected translations. Approximately 90% were judged semantically equivalent to the English source text. To compare the Hindi summaries generated by Helsinki-NLP/opus-mt-en-hi model with the Hindi summaries generated by Google Translate, a Corpus BLEU score was computed. The Corpus BLEU score obtained is 0.264, which is very low, reflecting very low lexical similarity between the two translation systems, and therefore, they produce very different Hindi translations for the same English summaries. Based on the experimental observations, the Hindi summaries generated using Google Translate were selected for the subsequent experiments, as they yielded more effective classification performance.

Algorithm 1 outlines the algorithm used to create aligned English-Hindi arXiv corpus.

Algorithm 1 English–Hindi arXiv Corpus Alignment

Require: arXiv English summaries $D = \{x_1, x_2, \ldots, x_N\}$
Ensure: Aligned bilingual arXiv corpus $D^{aligned}$
Load dataset and initialize GoogleTranslator $T_{(en \to hi)}$
for each summary $x_i \in D$ **do**
  if x_i is empty **then**
    $x_i^{hi} \Leftarrow$ ""
    continue
  end if
  Translate $x_i^{hi} \Leftarrow T(x_i)$ with retry limit R
  Apply delay Δt for rate control
  Construct aligned pair (x_i, x_i^{hi})
end for
Save updated dataset as aligned bilingual arXiv corpus $D^{aligned}$

Random oversampling and augmentation are performed on this aligned English-Hindi arXiv dataset in order to obtain balanced big data corpus. Table II outlines the dataset features after the preprocessing, translation and balancing.

The preview of the arXiv dataset samples after preprocessing, translation and balancing is depicted in Fig. 2.

TABLE II. ARXIV DATASET FEATURES AFTER PREPROCESSING, TRANSLATION AND BALANCING

Dataset Feature	Description
Size	895 MB
Attributes	id, title, title_hi_ref, category, category_code, published_date, updated_date, authors, first_author, summary, summary_word_count, year, month, day, category_label, summary_hi_ref
Modality Addressed	Text
Linguality Addressed	English, Hindi
Number of Samples	296,065
Number of Categories	73

id	title	title_hi_ref	category	category_code	published_date	updated_date	authors	first_author	summary	summary_word_count	year	month	day	category_label	summary_hi_ref
abs-2205.09519v1	design math	स्वदेक ओरगर	systems_a	eess.SY	19-05-2022	19-05-2022	['Aadhitiy	'Aadhitiya V	S neuromor	118	2022	5	19	72	न्यूरोमॉर्फिक कंप्यूटिंग ने
abs-2407.18523v1	dtformer tra	डीटीफोर्मेर ट्रां	machine_l	cs.LG	26-07-2024	26-07-2024	['Xi Chen',	'Xi Chen'	discrete ti	247	2024	7	26	43	असतत समय गतिशील
abs-1811.04115v1	adnet deep r	एडनेट डीप नेट	multimedi	cs.MM	09-11-2018	09-11-2018	['Murhaf H	'Murhaf Hoss	online vid	135	2018	11	9	50	ऑनलाइन वीडियो विज्ञा
abs-1803.10986v3	error analysi	त्रुटि विश्लेषण	numerical	cs.NA	29-03-2018	01-05-2019	['Barbara I	'Barbara Bar	popular di	191	2018	3	29	54	लोकप्रिय डीप न्यूरल नेट
abs-2212.04318v1	power consu	पावर खपत	networkin	cs.NI	08-12-2022	08-12-2022	['Nicola Pi	'Nicola Piove	fifth gener	186	2022	12	8	51	पांचवीं पीढ़ी के रेडियो ए
abs-1809.04951v1	valid simulta	उच्च आयामी	economet	econ.EM	13-09-2018	13-09-2018	['Philipp B	'Philipp Bach	increase a	146	2018	9	13	28	उपलब्धता में वृद्धि उच्च
abs-1809.03343v1	distribute dy	डायनेमिक मॉ	systems_a	cs.SY	07-09-2018	07-09-2018	['Wenqing	'Wenqing Li'	large scale	198	2018	9	7	71	बड़े पैमाने पर ऑप्टिमिक
abs-2410.18605v1	understand	खिंचाड़ी की बा	machine_l	cs.LG	24-10-2024	24-10-2024	['Tianze W	'Tianze Wang	pilot studi	78	2024	10	24	43	पायलट अध्ययन एल्गोरि

Fig. 2. Representative samples from the processed multilingual arXiv dataset after preprocessing, machine translation and class balancing, highlighting the retained metadata attributes, English abstracts, translated Hindi abstracts and category labels used for bilingual classification.

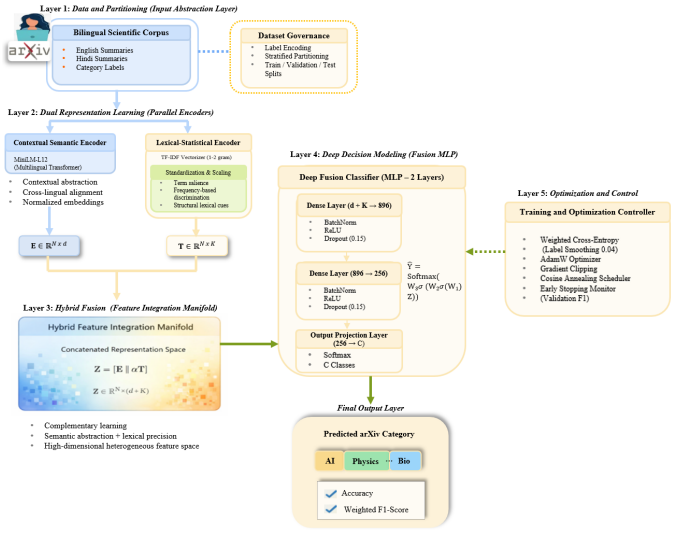


Fig. 3. Layered hybrid semantic-statistical feature fusion architecture showing multilingual context embeddings combined with statistical features, hybrid feature fusion and employing deep multiLayer perceptron to improve bilingual scientific text classification performance.

This aligned English-Hindi arXiv corpus is split using stratified sampling into 64% training, 16% validation, and 20% test sets to preserve the class distribution across all partitions. Experimental evaluation on this aligned big data corpus is performed separately for each language using multilingual transformer models and the proposed hybrid model to assess the language-specific performance.

B. Architectural Overview

To effectively address multilingual scientific summary classification, this study proposes a hybrid semantic-statistical fusion architecture that integrates contextual transformer-based representations with complementary lexical features within a unified deep learning framework. The proposed architecture, in contrast to purely embedding-driven models, further increases the discriminative power by explicitly modeling semantic abstraction as well as the statistical relevance of terms. This structure is specifically useful in scientific abstracts, where a mismatch in class representation, the variability of structure and vocabulary are significant challenges.

The proposed model depicted in Fig. 3 presents a hybrid representation learning model that works well to classify scientific literature by considering both deep contextual semantics and statistical lexical evidence. This system starts with a bilingual scientific corpus, i.e. English titles and summaries with category labels. A governance process of the data is necessary in order to guarantee consistency and reproducibility of the experimental pipeline whereby categorical labels are coded into numerical values and the dataset is stratified with stratified sampling. This stratification maintains the underlying distribution of classes in the training, validation, and test subsets. Hence, it ensures sound performance measurement and limits the chances of sampling bias in the process of training a model.

After the preparation of data, the architecture uses a dual representation learning approach where two complementary encoders are used simultaneously to represent various aspects of a textual information. The former route involves the use of a multilingual MiniLM-L12 transformer encoder, which produces contextualized representations of each document in the form of embeddings. This encoder represents deep semantic connections between tokens through self-attention mechanisms to enable the representation to absorb contextual dependency, semantic abstraction, and cross-lingual alignment that may exist within scientific text. The resulting dense embedding captures the overall semantic structure of the document, allowing the model to identify conceptual similarities despite the use of different lexical expressions.

Simultaneously, a lexical-statistical encoder is built to create a complementary representation with TF-IDF vectorization. The component uses discriminative lexical signals to extract the terms according to their significance in relation to their frequency in the corpus. In this way, domain-specific terminology which strongly defines specific scientific domains is emphasized more. TF-IDF, in contrast to contextual embeddings, which learn meaning by interacting with the context, emphasizes statistically salient keywords and structural lexical patterns that are often good predictors of types of documents. With this lexical view, the architecture conserves important information that would otherwise be diluted in neural embeddings.

The transformer encoder produces contextual semantic embeddings, while the TF-IDF module generates lexical-statistical feature vectors that encode term importance and document-specific vocabulary patterns. These complementary representations are integrated through a hybrid feature fusion layer utilizing feature concatenation, thereby constructing a unified high dimensional representation space. In particular, the semantic embedding vector and lexical-statistical feature vector are concatenated along the feature dimension, maintaining both the contextual semantic knowledge and discriminative lexical evidence. The resulting fused representation preserves

heterogeneous information sources that jointly characterize the conceptual content and terminological specificity of scientific documents. Such integration is particularly useful for scientific document classification, where semantic understanding and domain specific vocabulary play important roles in accurate category discrimination.

The fused representation is thereafter fed into a deep decision modeling component that is a multi-layer perception-based classifier. This type of fusion classifier also learns non-linear interactions among the semantic and lexical representations that allow the model to detect complex patterns that might not be discovered when these heterogeneous cues are used individually. The classifier gradually optimizes the fused representation to a discriminative feature space that is appropriate to predict categories by successive dense transformations, with the aid of the batch normalization and rectified linear activation functions. The dropout regularization is used to reduce the effect of overfitting by promoting feature learning that is robust throughout the learning process. The classifier effectively fills the gap between neural semantic representations and traditional lexical indicators by modeling interactions between contextual and statistical signals.

The architecture is coupled with a specific optimization and training control mechanism that is in charge of the learning dynamics of the model. This controller controls updating of parameters by adaptive optimization and training stabilization by gradient management and training scheduling strategies. Weighted cross-entropy is used to deal with possible class imbalance in the scientific data, so that no minority categories have a disproportionate influence on the learning goal. Gradient clipping ensures that deep networks do not update unstably, whereas cosine-based learning rate scheduling brings about a slow adaptation of optimization dynamics that encourages smooth convergence. The early stopping methodology which relies on F1-score of validation further ensures that the training process stops as soon as the generalization performance stops improving and thus avoiding overfitting and maintaining the robustness of the model.

Lastly, the trained classifier is used to generate the inferred scientific category of each document by mapping the learned representation into the target label space. The accuracy and weighted F1-score are used to measure the performance of the model which can give a holistic analysis of the effectiveness of classifications under imbalanced categories. Through the integration of transformer-based semantic modeling with statistical lexical representations and their combination by the means of a deep fusion-classifier, the given architecture creates a powerful framework of the scientific document categorization that exploits the strengths of both neural and traditional text representation paradigms.

C. Mathematical Formulation of the Proposed Semantic-Statistical Fusion Model

The suggested classification framework combines contextual semantics representations and statistical lexical representations to achieve a hybrid feature space of scientific text classification. Although transformer-based encoders can provide deep contextual semantics, they might not be able to maintain the significance of domain-specific lexical clues which can be very important in scientific texts. The proposed model overcomes this weakness by integrating multilingual MiniLM semantic embeddings with normalized TF-IDF statistical representations and learns their interactions by a simple FusionMLP classifier. This composite formulation allows the model to take advantage of both contextual meaning and term-level significance in a unit representation space. The following mathematical framework formally describes the construction of the semantic-statistical representation, the feature fusion mechanism, the classification function, and the optimization strategy used to learn the proposed model.

Let x_i denote the i-th document in the dataset. The proposed framework constructs a hybrid representation by combining contextual semantic embeddings obtained from a multilingual MiniLM encoder with normalized statistical TF–IDF features.

1) Semantic representation: Each document is first encoded using the MiniLM transformer to obtain contextual semantic embeddings:

$$e_i = f_{\text{MiniLM}}(x_i)$$

where, $f_{\text{MiniLM}}(\cdot)$ denotes the pretrained MiniLM encoder and $e_i \in \mathbb{R}^d$ represents the semantic embedding of dimension d.

2) Statistical feature extraction: To capture lexical importance signals, TF–IDF features are extracted for each document:

$$t_i = \text{TFIDF}(x_i), \qquad t_i \in \mathbb{R}^K$$

The statistical features are standardized using training-set statistics:

$$\tilde{t}_i = \frac{t_i - \mu}{\sigma}$$

where, $\mu = \mathbb{E}[t_{\text{train}}]$ and $\sigma = \text{Std}[t_{\text{train}}]$ denote the mean and standard deviation computed from the training corpus.

3) Controlled statistical feature scaling: To regulate the contribution of statistical features relative to semantic embeddings, a scaling coefficient is applied:

$$s_i = \alpha \tilde{t}_i \tag{1}$$

where, $\alpha \in [0, 1]$ controls the influence of the statistical component.

This transformation ensures that statistical features remain normalized and proportionally integrated with semantic embeddings.

4) Hybrid semantic-statistical feature fusion: The semantic and statistical representations are concatenated to construct the hybrid feature vector:

$$z_i = F(x_i) = [e_i \,;\, s_i] \tag{2}$$

where, $z_i \in \mathbb{R}^{d+K}$ represents the fused feature vector integrating contextual semantic information and lexical statistical signals. Using Eq. (1), the hybrid representation combines semantic embeddings with scaled statistical features.

5) FusionMLP decision function (two-layer architecture): The fused representation is passed through a two-layer multilayer perceptron (MLP) classifier to capture interactions between semantic and statistical components:

$$h_1 = \sigma(W_1 z_i + b_1)$$

$$h_2 = \sigma(W_2 h_1 + b_2)$$

$$\hat{y}_i = \text{Softmax}(W_3 h_2 + b_3)$$

where, z_i is the fused semantic–statistical feature vector defined in Eq. (2); W_1, W_2, W_3 and b_1, b_2, b_3 denote the learnable weight matrices and bias vectors of the MLP layers; $\sigma(\cdot)$ represents a nonlinear activation function (e.g., ReLU); h_1 and h_2 correspond to intermediate hidden representations; and $\hat{y}_i$ denotes the predicted class probability distribution obtained through the Softmax layer which can be compactly expressed as

$$\hat{y}_i = g_\theta(F(x_i))$$

where, $\theta = \{W_1, W_2, W_3, b_1, b_2, b_3\}$ denotes the model parameters.

6) Hybrid representation learning objective: The model parameters are optimized using cross-entropy loss as defined in Eq. (3):

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} y_i \log g_\theta(F(x_i)) \tag{3}$$

where, N denotes the total number of training samples, y_i represents the ground-truth label of the i-th sample, $g_\theta(\cdot)$ denotes the classifier parameterized by θ, $F(x_i)$ is the fused semantic–statistical representation defined in Eq. (2), and $\mathcal{L}(\theta)$ represents the cross-entropy loss used to optimize the model parameters.

This objective encourages the classifier to learn discriminative decision boundaries within the hybrid feature space defined in Eq. (2).

7) Optimization strategy: Training is performed using cosine annealing with warm restarts for adaptive learning-rate scheduling which is expressed in Eq. (4):

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos\left(\frac{\pi T_{\text{cur}}}{T_i}\right)\right) \tag{4}$$

where, η_t denotes the learning rate at iteration t, $\eta_{\min}$ and $\eta_{\max}$ represent the minimum and maximum learning rates respectively, T_{cur} denotes the number of iterations since the last restart, and T_i denotes the length of the i-th cosine annealing restart cycle, representing the number of iterations over which the learning rate decays from $\eta_{\max}$ to $\eta_{\min}$ before a restart occurs, computed as

$$T_i = T_0 \cdot T_{\text{mult}}^k$$

here, T_0 denotes the initial restart period, T_{mult} is the restart cycle multiplier controlling the growth of subsequent cycles, and k indicates the number of completed restart cycles.

8) Optimal model selection criterion: The final classifier is selected by maximizing validation performance which is defined in Eq. (5):

$$M^* = \arg\max_M F1_{\text{val}}(M) \tag{5}$$

where, $F1_{\text{val}}$ denotes the validation F1-score.

9) Compact unified formulation of the proposed model: The complete semantic–statistical fusion learning framework can be summarized as in Eq. (6):

$$M^* = \arg\max_M F1_{\text{val}}\left(g_\theta\left[f_{\text{MiniLM}}(x_i) \,;\, \alpha\left(\frac{\text{TFIDF}(x_i) - \mu}{\sigma}\right)\right]\right) \tag{6}$$

This formulation integrates multilingual contextual embeddings and normalized statistical features into a unified hybrid representation for efficient bilingual scientific document classification. The suggested model has three methodological contributions:

- A structured semantic-statistical fusion formulation that combines contextual embedding with features of lexical importance into a shared representation space.
- A controlled statistical scaling mechanism that regulates the influence of TF-IDF features to maintain distributional balance with semantic embeddings.
- An efficient hybrid representation learning strategy, which captures nonlinear interactions between semantic and statistical signals using a lightweight neural classifier.

D. Algorithm for Text Classification

The overall workflow of the proposed semantic–statistical feature fusion framework is summarized in Algorithm 2. It outlines the end-to-end pipeline for aligned bilingual text classification, including data preprocessing, hybrid feature construction using MiniLM embeddings and TF–IDF statistics, classifier training, and final evaluation.

Algorithm 2 Semantic–Statistical Feature Fusion Framework for Aligned Bilingual Text Classification

Require: Aligned document summaries $D = \{x_1, x_2, \ldots, x_N\}$
Ensure: Optimal classifier M^* and evaluation metrics $\{\text{Accuracy}, \text{Weighted F1-score}\}$
1: Load dataset D and corresponding class labels $Y = \{y_1, y_2, \ldots, y_N\}$
2: Encode labels and perform stratified partitioning to obtain training, validation, and test sets
3: **for** each document $x_i \in D$ **do**
4: Generate contextual semantic embedding e_i using the MiniLM encoder
5: Compute TF–IDF vector t_i using unigram and bigram statistics and normalize using training-set statistics
6: Construct fused representation $f_i = [e_i, t_i]$
7: **end for**
8: Feed fused features f_i into the FusionMLP classifier M
9: Train M using cross-entropy loss with AdamW optimization and cosine learning-rate scheduling
10: Monitor validation F1-score and apply early stopping to obtain optimal model M^*
11: Evaluate M^* on the test set and compute Accuracy and Weighted F1-score

IV. Empirical Evaluation and Comparative Performance Analysis

A comprehensive empirical study of the suggested semantic-statistical feature fusion model on English-Hindi aligned corpus of arXiv is presented in this section. Model specification and hyperparameter configurations are outlined in detail. The quantitative performance of the proposed model, its' convergence behavior, class-level robustness, overall classification performance, architectural contribution via ablation, computational efficiency, and bilingual alignment properties are systematically analyzed. Multiple evaluation metrics are used to report comparative measurements against known multilingual transformer baselines in order to provide a rigorous and balanced performance evaluation.

A. Model Specification and Training Parameters

Using the common data splits and hyperparameter configuration, both the baseline and proposed models were trained to carry out comprehensive testing of the proposed framework. In this subsection, the architectural considerations of the base and hybrid models are presented as well as the training environments such as the learning rates, the batch size and the optimization methods. The setup allows studying convergence processes and performance efficiency and ensures that the proposed MiniLM-L12+MLP fusion architecture and the multilingual baselines can be fairly compared. Table III summarizes the entire experimental arrangement used for classification.

TABLE III. Model Design, Training Protocol and Hyperparameter Configuration for Hybrid Feature Fusion Classification Model Applied to an Aligned English–Hindi arXiv Corpus

Category	Hyperparameter	Value	Purpose
Embedding Model	Sentence Transformer	paraphrase-multilingual-MiniLM-L12-v2	Lightweight multilingual semantic encoder.
	Embedding Normalization	Enabled	Ensures cosine-consistent representations.
Batching	Batch Size	96	Balances GPU memory & convergence stability.
Training	Epochs (max)	65	Allows full convergence with early stopping.
	Learning Rate	1.5×10^{-4}	Stable for fused semantic + lexical features.
	Optimizer	AdamW	Handles sparse & dense feature fusion.
	Weight Decay	3×10^{-5}	Prevents overfitting.
	Gradient Clipping	1.0	Stabilizes training.
Regularization	Dropout	0.15	Controls co-adaptation in fusion layers.
	Label Smoothing	0.04	Improves generalization.
Scheduler	LR Scheduler	CosineAnnealing Warm Restarts	Escapes sharp minima.
	Initial Restart Period (T_0)	8	Fast warm restarts.
	Restart Multiplier (T_{mult})	2	Gradually longer cycles.
Early Stopping	Patience	8 epochs	Prevents overfitting.
	Monitoring Metric	Weighted F1-score	Robust for class imbalance.
Classifier	Architecture	MLP ($896 \to 256 \to C$)	Deep fusion classifier.
	Batch Normalization	Yes	Training stability.
Evaluation	Metrics	Accuracy, Weighted F1-score	Balanced performance reporting.
	Hardware	GPU	Efficient computation.

The multilingual MiniLM-L12 encoder computes a semantic embedding vector of 384 dimensions for each document in English text classification. Meanwhile, a 512-dimensional TF-IDF representation with the help of unigram and bigram features is obtained after removing English stop-words. Lexical information is also preserved by scaling the TF-IDF features by a factor of 0.50, while semantic embeddings dominate the features. The output of the semantic embedding vector is concatenated with the output of the scaled TF-IDF feature vector, forming an 896-dimensional representation (384 + 512), and then input to the residual MLP classifier.

The multilingual MiniLM-L12 encoder also generates a 384-dimensional vector embedding of the text in Hindi for the task of Hindi text classification. No stop-word filtering is done to retain the linguistic cues for Hindi and a 512 dimensional unigram–bigram TF-IDF representation is extracted. After

obtaining the semantic embedding vector, the features in the TF-IDF are scaled by a factor of 0.50 and then fused at the feature level. The fused representation from this process is a 896 dimensional (384+512) vector, fed into the residual MLP classifier.

B. Quantitative Performance Evaluation of the Proposed Framework

To evaluate the efficiency of the proposed semantic-statistical hybrid structure, quantitative test of English and Hindi summary classification is done which is outlined in this section. Performance is evaluated using standard classification measures derived from True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). For balanced assessment of the potentially imbalanced scientific categories, Accuracy and Weighted F1-score are used.

Classification accuracy is computed as in Eq. (7):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

For class-wise evaluation, precision, recall, and F1-score are defined as in Eq. (8), (9) and (10) respectively:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{8}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{9}$$

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{10}$$

To account for class imbalance, the weighted F1-score is calculated as in Eq. (11):

$$\text{Weighted-F1} = \sum_{c=1}^{C} \frac{n_c}{N} \cdot \text{F1}_c \tag{11}$$

where, C denotes the total number of classes, n_c represents the number of samples in class c, and N denotes the total number of samples.

The findings in Table IV indicate the general categorization ability of the proposed MiniLM-L12 + MLP (2 Layers) + TF-IDF fusion model in the bilingual environment. The proposed framework delivers high and consistent results between the two languages, reaching an accuracy of 95.56% and a weighted F1-score of 95.31% in English, and 94.85% in accuracy with a weighted F1-score of 94.53% in Hindi. The cost of marginal difference in performance between English and Hindi represents the cross-lingual generalization and strong semantic correspondence. The high weighted F1-scores also prove that the model has balanced predictive accuracy across various scientific categories indicating the success of the semantic-statistical feature fusion strategy in the classification of bilingual scientific texts.

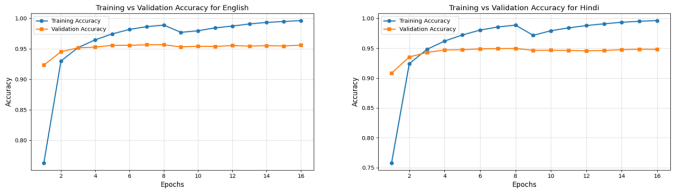
TABLE IV. ACCURACY AND WEIGHTED F1-SCORE ANALYSIS OF THE PROPOSED MODEL ON ENGLISH AND HINDI SUMMARY CLASSIFICATION

	Accuracy		Weighted F1-Score	
Proposed Model	**English**	**Hindi**	**English**	**Hindi**
MiniLM-L12 + MLP (2 Layers) + TF-IDF	0.9556	0.9485	0.9531	0.9453

C. Training Dynamics and Convergence Behaviour of the Proposed Model

This subsection analyses the learning stability and convergence characteristics of the proposed bilingual MiniLM-L12 + MLP framework using Training Accuracy curves, Validation F1 curves, and Training-Validation Loss curves across epochs. All these metrics give a clue of the efficiency in optimization, generalization, and overfitting behavior of both English and Hindi datasets.

The training dynamics are characterized by a quick early convergence and a slow stabilization in subsequent epochs, which is evidence of efficient parameter adaptation and stable optimization. Fig. 4(a) and 4(b) shows that the training accuracy is steep in the initial few epochs, which is indicative of learning features effectively and using a strong gradient-driven optimization. Following the initial phase, the curve switches to smooth plateau towards almost saturation (∼99%), which proves to be stable convergence without oscillations. The fact that the two languages are consistent demonstrates the strength of the bilingual representation learning.

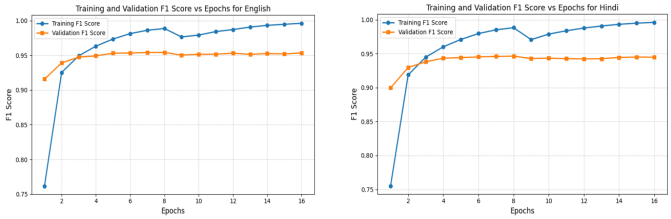


(a) For English Language (b) For Hindi Language

Fig. 4. Accuracy of the proposed model with progression of training epochs, which showed rapid convergence, low overfitting, and stable generalization in English and Hindi languages.

The validation F1 score increases consistently through the initial epochs and reaches the high level (∼95%), meaning that the model has a high-performance in terms of generalization as discussed in Fig. 5(a) and 5(b). The narrow discrepancy between training and validation curves indicates regulated model capability and restricted overfitting. The trend of the validation F1 also confirms that the classifier has a balanced precision recall trade-offs across classes.

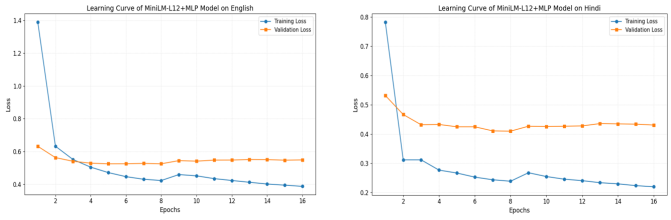
Fig. 6(a) and 6(b) show the loss-versus-epochs curves of both English and Hindi datasets, which approach zero very quickly in early epochs, which is evidence of an effective gradient based learning and fast parameter adjustment. After this initial convergence stage, loss curves begin to flatten which is an indication of a stabilization in the process of optimization. Both the validation loss and the training loss are stable and have slight variation indicating model capacity is controlled and regularization is successful. The lack of a substantial sep-



(a) For English Language (b) For Hindi Language

Fig. 5. Validation F1-score of the proposed model with training epochs, showing the rapid start with a high convergence, followed by a level with stabilization, and then high performance in generalization in both English and Hindi language.

aration between training and validation paths is a verification of the low overfitting and high-quality generalization both in languages.



(a) For English Language (b) For Hindi Language

Fig. 6. Training and validation loss traces per epoch of the proposed model in English and Hindi summary classification, showing early rapid convergence and stable optimization behavior.

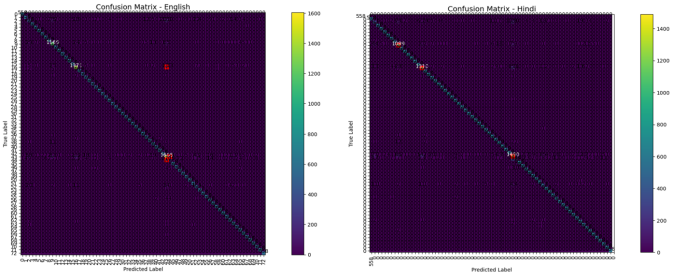
In general, the dynamics of the losses observed support the stability and robustness of the proposed framework. The fast convergence and a low validation loss show that the model does not only learn discriminative features effectively but also retains the performance of cross-lingual generalization. Such a consistent optimization pattern is also additional evidence supporting the power of the MiniLM-L12 + MLP model in the task of bilingual scientific text classification.

D. Class-Level Performance Analysis of the Proposed Model

In this subsection, a comprehensive analysis of the proposed MiniLM-L12 + MLP model on a per-class level is conducted to determine its discriminative ability when dealing with individual categories. Although overall accuracy and macro-F1 give a comprehensive picture of the performance, a more detailed analysis of the results by classes gives a better understanding of the label-specific behavior, the patterns of inter-class confusion, and the cross-lingual consistency.

Here, confusion matrices of English and Hindi are studied and per-class F1-scores are investigated and a comparative visual representation of English and Hindi class-wise F1 performance is given. These analyses collectively demonstrate that the model ensures a balanced precision-recall trade-offs between categories, detects the possible confusion-prone classes, and confirms the strength of the bilingual representation learning framework.

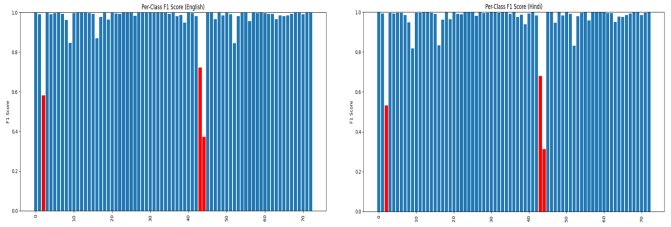
The confusion matrices provided in Fig. 7(a) and 7(b) of both English and Hindi reveal high degree of concentration



(a) For English Language (b) For Hindi Language

Fig. 7. Confusion matrix analysis of the proposed model for English and Hindi text classification, illustrating the distribution of true and predicted labels, highlighting strong category discrimination with reduced inter-class confusion.

of all the values along the diagonal which implies that a majority of the instances are being correctly classified. The low values of off-diagonals indicate low levels of misclassification of classes. The pattern indicates high accuracy, and the performance of the model in the two languages and also shows that multilingual text classification is well learned and capable of good generalization.



(a) For English Language (b) For Hindi Language

Fig. 8. Per-class F1 score analysis of the proposed model for English and Hindi text classification, demonstrating balanced classification performance, confirming the effectiveness of hybrid feature representation.

In order to further justify this performance at a more detailed level, the per-class F1 scores are examined. Although the confusion matrices give a general picture of the correct and incorrect predictions, the per-class F1 score gives a clear picture of the ability of the model to strike the right balance between precision and recall on each category selected. Based on the Fig. 8(a) and 8(b), English and Hindi models have consistently high F1 scores on most of the classes, which shows that the models have high and consistent classification performance. There are only a few classes with slightly lower values, indicating that there is a slight room to improve. In general, the findings indicate that the model is well generalized in both languages, and the performance is strong across categories.

After analyzing the individual per-class version of the F1 score, the comparative graph in Fig. 9 gives a direct class-based analysis of English and Hindi. According to the visualization, the proposed model demonstrates the same high F1 scores on the majority of the classes in both languages, with the performance trends being slightly different. Small differences identified in some classes suggest some slight language-specific differences, but in general, the outcomes prove that the model is highly multilingual and has equal

classification power when working in both English and Hindi datasets.

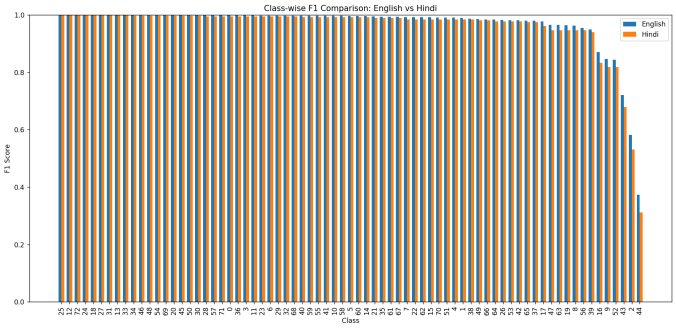


Fig. 9. Class-wise F1 score comparison of the proposed model for English and Hindi arXiv summary classification, highlighting strong cross-lingual generalization.

E. Overall Comparative Performance Analysis with Baselines

This subsection presents a comprehensive quantitative comparison between the proposed MiniLM-L12 + MLP fusion framework and established multilingual transformer baselines (MiniLM-L12, mBERT, and XLM-RoBERTa) on the English-Hindi aligned arXiv corpus. In order to ensure a balanced assessment across class distributions and languages, performance is evaluated using accuracy, macro, micro, and weighted precision, recall, and F1-scores The comparison will focus both on the absolute performance increase and also on the consistency and strength of the proposed approach in bilingual environments.

The macro F1-score, which treats all classes equally, is computed as in Eq. (12):

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} F1_c \tag{12}$$

where, C denotes the total number of classes, and $F1_c$ represents the F1-score of class c.

The micro F1-score, which aggregates global counts across classes, is given by Eq. (13):

$$\text{Micro-F1} = \frac{2\sum_c TP_c}{2\sum_c TP_c + \sum_c FP_c + \sum_c FN_c} \tag{13}$$

As demonstrated in Table V, the suggested MiniLM-L12 + MLP architecture yields the best performance in almost all the evaluation metrics of both the English and Hindi languages. The macro and micro F1-scores also validate the stability of performance in categories illustrating that there is no dominance in the improvement but it is always distributed throughout the spectrum of classes. The improvements to the strongest baseline (XLM-RoBERTa) are also consistent across all languages and all aggregated measures, indicating that semantic-statistical feature fusion can be successfully used to reinforce multilingual scientific text classification. In general, the findings indicate excellent, balanced, and statistically strong bilingual performance of the suggested framework in comparison to the state-of-the-art transformer baselines.

TABLE V. COMPARATIVE PERFORMANCE ANALYSIS OF BASELINE TRANSFORMER MODELS AND THE PROPOSED MINILM-L12+MLP FUSION FRAMEWORK FOR ENGLISH–HINDI ALIGNED ARXIV CORPUS

Performance	Model							
	MiniLM		mBERT		XLM-R		Proposed MiniLM+MLP	
Metric	E	H	E	H	E	H	E	H
Accuracy	0.9183	0.8983	0.9449	0.9369	0.9422	0.9377	**0.9556**	**0.9485**
Macro Prec.	0.9245	0.9058	0.9561	0.9476	0.9530	0.9480	**0.9670**	**0.9603**
Macro Recall	0.9413	0.9242	0.9703	0.9659	0.9684	0.9661	**0.9714**	**0.9678**
Macro F1	0.9305	0.9122	0.9623	0.9554	0.9594	0.9558	**0.9684**	**0.9634**
Micro Prec.	0.9183	0.8983	0.9449	0.9369	0.9422	0.9377	**0.9556**	**0.9485**
Micro Recall	0.9183	0.8983	0.9449	0.9369	0.9422	0.9377	**0.9556**	**0.9485**
Micro F1	0.9183	0.8983	0.9449	0.9369	0.9422	0.9377	**0.9556**	**0.9485**
Weighted Prec.	0.9087	0.8872	0.9443	0.9351	0.9417	0.9361	**0.9522**	**0.9437**
Weighted Recall	0.9183	0.8983	0.9449	0.9369	0.9422	0.9377	**0.9556**	**0.9485**
Weighted F1	0.9109	0.8899	0.9424	0.9329	0.9389	0.9337	**0.9531**	**0.9453**

To further substantiate the comparative findings reported in Table V, a percentage gain analysis is conducted to quantify the relative improvement of the proposed MiniLM-L12 + MLP framework over established multilingual baselines. Although absolute performance measures reveal unwavering superiority in the English and Hindi tasks, relative gain calculation offers some normalized view of the magnitude of improvement in models with different baseline levels.

The percentage gain is computed as in Eq. (14)

$$\text{Gain}(\%) = \frac{M_{\text{proposed}} - M_{\text{baseline}}}{M_{\text{baseline}}} \times 100 \tag{14}$$

where, M denotes the corresponding evaluation metric. This formulation ensures that improvements are measured relative to the baseline strength rather than in absolute terms. These are the percentage-based improvements in both accuracy and weighted F1-score, which are represented in the heatmap in Fig. 10, which gives a more accurate view of differences in performance between the proposed framework and MiniLM-L12, mBERT, and XLM-RoBERTa.

As indicated in the heatmap, the suggested framework has the highest relative gains over MiniLM-L12, with the improvements of 4.06% (English accuracy), 5.59% (Hindi accuracy), 4.63% (English F1), and 6.23% (Hindi F1). Such large gains signify that TF-IDF-based statistical features combined with contextual embeddings are powerful boosters of representational capabilities of small transformer architecture. The improvement when compared with mBERT is more moderate, and it is about 1.13-1.33 % showing the growth in form of improvements to already strong multilingual benchmark. On the XLM-RoBERTa, the strongest baseline, the improvements are, nevertheless, positive, ranging between 1.15 and 1.51%, which proves the idea that large-capacity pretrained models can still be influenced by semantic-statistical fusion.

Significantly, not only accuracy, but also weighted F1-score is being shown to be continuously improved. The consistent improvement in both languages and both measures proves that the given framework improves the overall predictive ability and its strength even in the case of uneven distributions of

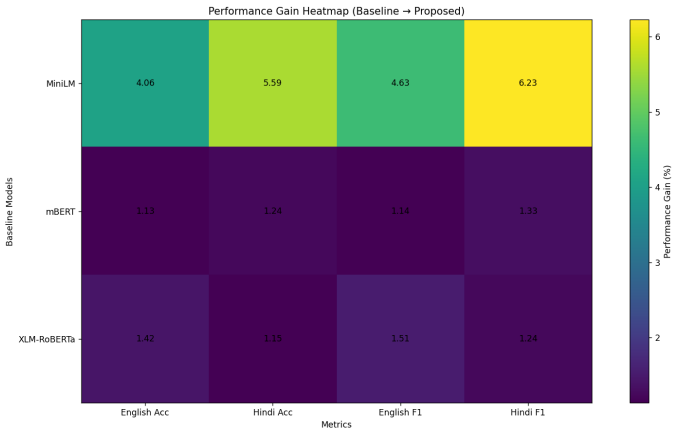


Fig. 10. Heatmap of percentage performance gains of the proposed model over MiniLM-L12, mBERT, and XLM-RoBERTa, highlighting consistent improvements across English and Hindi classification accuracy and weighted F1-scores.

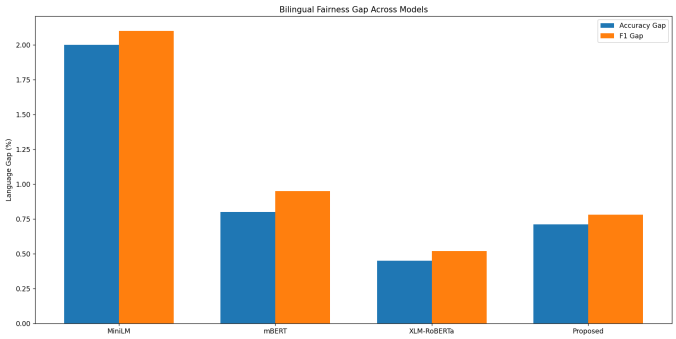


Fig. 11. Bilingual fairness gap comparison that demonstrates the performance disparity between English and Hindi languages in baseline and proposed models.

scientific categories. Taken together, the heatmap confirms the excellence and consistency of the proposed bilingual fusion model at different transformer capacities.

F. Cross-Language Alignment and Fairness Results in Comparison with Baseline and Proposed Models

In order to measure the strength and impartiality of the suggested scheme of bilingual classification, the cross-language alignment of English and Hindi is studied through performance gap analysis. Instead of measuring absolute performance, this subsection measures the consistency of individual model performance across languages. A lower performance difference shows a more robust cross-lingual representation balance and enhanced fairness, which ensures that a gain in one language does not affect the other in a more inappropriate way. This assessment is especially significant in multilingual scientific text analytics, where balanced cross-language performance is a direct measure of representation quality and transferability.

The bilingual fairness gap is computed as the absolute difference between English and Hindi performance metrics which is as defined in Eq. (15):

$$\text{Gap}_{metric} = |M_{EN} - M_{HI}| \tag{15}$$

where, M_{EN} and M_{HI} denote the evaluation metric computed for the English and Hindi text respectively, and M can represent either Accuracy or F1-score. The absolute operator ensures non-negativity and measures deviation irrespective of direction.

Based on the results obtained as depicted in Fig. 11, the proposed MiniLM-L12 + MLP fusion model proves to have a significantly small bilingual gap and, at the same time, achieves the highest overall Accuracy and F1 scores. This is a good sign of cross language embedding compatibility and good features fusion. The proposed architecture does not only provide competitive multilingual consistency when compared to bigger models like mBERT and xLM-RoBERTa, but also enhances efficiency. The decreased fairness disparity puts to rest the fact that the lightweight hybrid modeling can gain balanced and fair bilingual performance without declining predictive strength, which lays down the suitability of lightweight hybrid modeling as an approach to scale-oriented multilingual scientific document classification.

According to the bilingual alignment scatter plot in Fig. 12, it is evident that all models have high-levels of cross-lingual consistency because all their coordinate pairs are consistent. $(x_i, y_i) = (\text{Accuracy}_{EN}, \text{Accuracy}_{HI})$ lie close to the ideal alignment line $y = x$. The deviation from perfect alignment is quantified as $\Delta_i = |x_i - y_i|$, where smaller Δ_i indicates better bilingual stability.

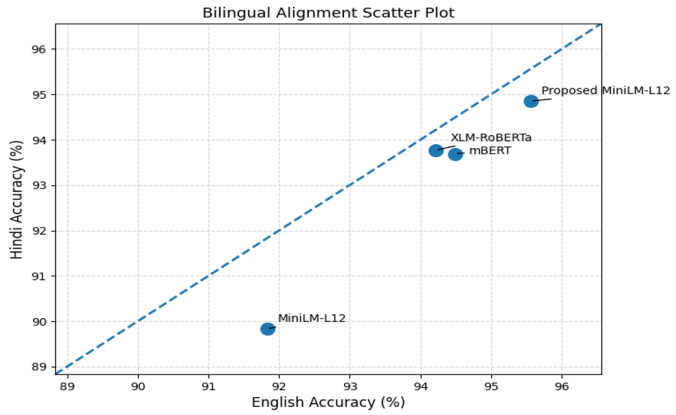


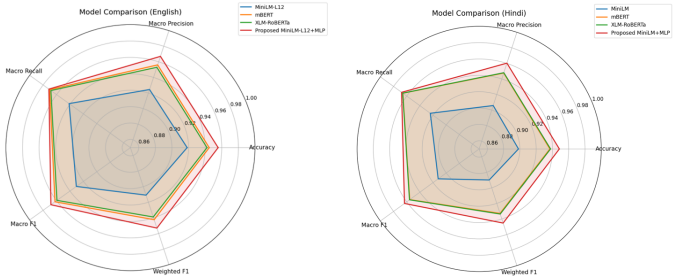
Fig. 12. Bilingual alignment scatter plot comparing the English and Hindi classification accuracy of baseline models and the proposed framework with the diagonal line representing an ideal cross-language performance balance.

The proposed MiniLM-L12 + MLP model is positioned closest to the top-right region, meaning it maximizes $\max(x_i, y_i)$ simultaneously while maintaining minimal alignment gap, demonstrating strong balanced performance across English and Hindi. Regarding the baselines, MiniLM-L12 is one of the weakly distilled transformer baselines with fast inference and relatively lower accuracy; mBERT and XLM-RoBERTa are powerful multilingual transformer baselines that demonstrate high accuracies because of the deep contextual representations; however, their points, although having a tendency to be close to the diagonal, have more significant differences in the alignment and do not exceed the suggested model in terms of combined bilingual capabilities. It means

that despite such large pretrained multilingual models being strong, the suggested hybrid structure demonstrates an even better cross lingual efficiency accuracy ratio and much lower processing cost.

The radar plots depicted in Fig. 13 provide a holistic multi-metric comparison by representing each model as a vector in a 5-dimensional performance space:

$$M_i = [Acc,\ P_{\text{macro}},\ R_{\text{macro}},\ F1_{\text{macro}},\ F1_{\text{weighted}}]$$



(a) For English Language (b) For Hindi Language

Fig. 13. Dual-language radar chart comparing MiniLM-L12, mBERT, XLM-RoBERTa, and the proposed MiniLM-L12+MLP model across English and Hindi summary classification accuracy, macro precision, macro recall, macro F1 and weighted F1-scores, illustrating the superior and balanced bilingual performance achieved by the proposed framework.

Each metric is mapped to a polar coordinate axis with equally spaced angles as defined in Eq. (16):

$$\theta_k = \frac{2\pi k}{K}, \quad k = 0, 1, \ldots, K-1 \tag{16}$$

where, θ_k denotes the angular position of the k-th metric on the circular axis, with $K = 5$ evaluation metrics. The radial value r_{ik} corresponds to the metric score of model i for metric k. where $K = 5$ metrics. The radial value r_{ik} corresponds to the metric score of model i for metric k. By closing the polygon (repeating the first value), the area enclosed by each model can be interpreted as an approximate aggregate performance measure, defined in Eq. (17):

$$\text{Area}_i \propto \sum_{k=1}^{K} r_{ik}\, r_{i(k+1)} \sin\left(\frac{2\pi}{K}\right) \tag{17}$$

A larger enclosed area indicates stronger overall balanced performance across metrics. In both English and Hindi plots depicted in Fig. 13, the proposed MiniLM-L12 + MLP constantly optimizes the outermost and most homogenous polygon implying concurrent maximization of accuracy, precision, recall, and F1 values. All the baselines are also characterized by high and stationary performance where, mBERT and XLM-RoBERTa have competitive macro-level results which are provided by deep multilingual representations. Nevertheless, it is possible to state that the offered hybrid model will provide the most balanced and extended radar coverage in both languages, which is the justification of its high multi-metric strength and cross-linguistic generalization power.

G. Efficiency Analysis of the Baseline Versus Proposed Model

Computational efficiency is a decisive factor in multilingual scientific text classification especially when the model is expected to be deployed on a scale or to make real-time inferences. Despite offering effective contextual representations of large-scale multilingual models, including mBERT and XLM-RoBERTa, large transformer-based models have large parameter sizes that lead to high training and inference expenses. By comparison, distilled transformer architectures, such as MiniLM-L12, are designed to maintain the quality of semantic representations and simplify computation. The hybrid scheme also adds an additional lightweight layer of two-layers MLP and TF-IDF feature fusion to MiniLM-L12 with a slight increment in the number of parameters. Thus, the systematic efficiency analysis is performed through the comparison of parameters count, inference time on a single sample, and total training time between English and Hindi data to measure both the computational scalability and predictive performance.

Inference time per sample is computed as in Eq. (18)

$$T_{\text{inf}} = \frac{T_{\text{total inference}}}{N} \tag{18}$$

where, N denotes the number of test samples. Training time is computed as in Eq. (19)

$$T_{\text{train}} = \frac{E \times N}{B} \times T_{\text{step}} \tag{19}$$

where, E denotes the number of epochs, B represents the batch size, and T_{step} denotes the time required for one training step, which depends on the parameter count and the backpropagation cost $O(P)$.

TABLE VI. COMPARATIVE EFFICIENCY ANALYSIS OF BASELINE TRANSFORMER MODELS AND THE PROPOSED MINILM-L12+MLP FUSION FRAMEWORK ACROSS ENGLISH AND HINDI SUMMARY CLASSIFICATION

Model	Architecture	No. of Params	Inference Time (ms)		Training Time (sec)	
			E	H	E	H
MiniLM-L12	Transformer encoder (distilled)	∼117M	0.0060	0.0058	166.25	209.38
mBERT	BERT-base (12 layers)	∼110M	2.5013	2.6770	8020.74	8442.66
XLM-RoBERTa	RoBERTa transformer	∼270M (base)	2.4997	2.5200	8802.94	8887.52
Proposed	**MiniLM + 2-Layer MLP with TF-IDF fusion**	∼118M (117M + ∼1.3M)	**0.0074**	**0.0075**	**101.94**	**77.63**

The results outlined in Table VI clearly demonstrate the computational efficiency and scalability of the proposed MiniLM-based hybrid model compared to heavy multilingual transformers like mBERT and XLM-RoBERTa. Since transformer complexity scales approximately as $O(L \cdot H^2 + L^2 \cdot H)$ (where L denotes the sequence length and H represents the hidden dimension), larger multilingual models such as XLM-RoBERTa (∼270M parameters) incur higher computational

latency. Conversely, MiniLM-L12 employs knowledge distillation to reduce attention computation, achieving an almost constant ultra-low inference latency ($\sim$0.006$-$0.007 ms/sample). Although the proposed model introduces a lightweight MLP classifier ($896 \rightarrow 256 \rightarrow C$) along with TF–IDF feature fusion, the additional parameters ($\sim$1.3M) minimally affect the overall computational complexity, maintaining a compact total parameter count of approximately $\sim$118M. close to MiniLM while significantly reducing training time due to efficient convergence and feature fusion. Altogether, hybrid architecture demonstrates a very good efficiency-performance ratio, which proves the appropriateness to scalable bilingual scientific text analytics of English and Hindi.

To further validate the competitiveness of the proposed framework, additional experiments were conducted using recent multilingual sentence embedding models including Language-agnostic BERT Sentence Embedding (LaBSE) and multilingual E5. Results in Table VII indicate that the proposed semantic-statistical fusion framework consistently achieves superior classification performance.

TABLE VII. COMPARATIVE PERFORMANCE ANALYSIS OF THE PROPOSED MINILM-L12+MLP FUSION FRAMEWORK AND STATE-OF-THE-ART MODELS FOR ENGLISH–HINDI ALIGNED ARXIV CORPUS

Performance	Model					
	LaBSE		E5-Base		Proposed MiniLM+MLP	
Metric	E	H	E	H	E	H
Accuracy	0.9065	0.9012	0.9055	0.8883	**0.9556**	**0.9485**
Macro Prec.	0.9079	0.9039	0.9095	0.8951	**0.9670**	**0.9603**
Macro Recall	0.9445	0.9419	0.9421	0.9294	**0.9714**	**0.9678**
Macro F1	0.9220	0.9184	0.9221	0.9084	**0.9684**	**0.9634**
Micro Prec.	0.9065	0.9012	0.9055	0.8883	**0.9556**	**0.9485**
Micro Recall	0.9065	0.9012	0.9055	0.8883	**0.9556**	**0.9485**
Micro F1	0.9065	0.9012	0.9055	0.8883	**0.9556**	**0.9485**
Weighted Prec.	0.8948	0.8898	0.8970	0.8782	**0.9522**	**0.9437**
Weighted Recall	0.9065	0.9012	0.9055	0.8883	**0.9556**	**0.9485**
Weighted F1	0.8931	0.8874	0.8939	0.8752	**0.9531**	**0.9453**

H. Ablation Study

In order to systematically analyze the role played by each architectural component, an ablation study of the base MiniLM-L12 encoder is performed and following are gradually added: 1) a 1-layer MLP, 2) a 2-layer MLP, 3) TF-IDF feature fusion and (iv) summary length features. The aim is to measure the influence of each additional module on classification performance, computational performance, and resource consumption in the English and Hindi languages. Accuracy and Weighted F1-score are used to determine performance, whereas inference time per sample, total training time, and GPU memory allocated are used to measure efficiency.

The underlying MiniLM-L12 as given in Table VIII, offers a high level of efficiency at low inference time ($\sim$ 0.006 ms/sample) and low GPU memory ($\sim$ 0.46 GB), but at a medium level of accuracy. The performance increases with the addition of 1-layer and 2-layer MLP sequentially, which demonstrates that the increased depth of nonlinear transformation increases the separability of the features. This however, adds parameters to the training time slightly. The highest Accuracy and the Weighted F1 are shown in the proposed

TABLE VIII. ABLATION EXPERIMENT ON HYBRID ARCHITECTURE OF MINILM-L12 FOR ENGLISH–HINDI CLASSIFICATION

	Acc.		Wt. F1		Infer.(ms)		Train (sec)		GPU Mem.	
Model	**E**	**H**	**E**	**H**	**E**	**H**	**E**	**H**	**E**	**H**
MiniLM-L12	0.9183	0.8983	0.9109	0.8899	0.0060	0.0058	166.25	209.38	0.460	0.467
MiniLM + 1-L MLP	0.9189	0.9118	0.9146	0.9068	0.0066	0.0064	284.39	327.52	0.470	0.487
MiniLM + 2-L MLP	0.9195	0.9124	0.9167	0.9089	0.0076	0.0067	306.54	349.67	0.470	0.500
Proposed MiniLM + 2-L MLP + TF–IDF	**0.9556**	**0.9485**	**0.9531**	**0.9453**	**0.0074**	**0.0075**	**101.94**	**77.63**	**0.480**	**0.510**
MiniLM + 2-L MLP + TF–IDF + Sum. Len.	0.9367	0.9296	0.9344	0.9266	0.0074	0.0078	279.52	322.65	0.480	0.510

model (MiniLM-L12 + 2-layer MLP + TF-IDF fusion) in both languages. A feature fusion enhances semantic embeddings by adding statistical features, strengthening the discriminative capacity of the models and surprisingly minimizing the amount of time needed to train, because of accelerated convergence.

Interestingly, summary length as an added feature does not enhance the performance any further which shows that TF-IDF already has enough statistical variance. Comprehensively, the ablation suggests that the significant improvement of the performance is caused by semantic-statistical fusion and not simply by the increase in the depth of the network, providing optimal performance and computational efficiency.

I. External Benchmark Generalization

A comprehensive comparative analysis of recent text classification studies across diverse domains, datasets, and linguistic settings is presented in Table IX.

According to the survey, the tasks of multi-class classification are largely dominant and include reviews of drugs, news, and patent analysis, classification of scientific documents, and biomedical sentence labeling, whereas binary classification is found mostly in sentiment and hate speech classification tasks. The vast majority of the methods are used in monolingual environments, and relatively few multilingual implementations are made. Models and datasets performance differ with the highest accuracy reported at 94.1%. It is worth noting that the proposed MiniLM-L12 + MLP model is the most accurate (95.56%) in a multilingual scientific classification environment, which proves to be more effective than the currently available baselines.

V. DISCUSSION

The success of the suggested model could be explained by the fact that its architecture incorporates the complementary representations of textual information and ensures the efficient learning dynamics. The model provides both deep semantic representations and surface-level discriminative patterns that exist in the dataset by embedding contextual semantic meanings, in addition to statistically based textual features. Contextual embeddings give rich representations to encode syntactic relationships, long-range dependencies, and domain specific semantics in the text enabling the model to interpret the underlying intent and contextual subtlety of every instance.

TABLE IX. COMPARATIVE PERFORMANCE ANALYSIS OF RECENT STATE-OF-THE-ART MODELS AND THE PROPOSED MODEL ACROSS DIVERSE BENCHMARK DATASETS

Cit.	Model	Task	Dataset	Size	Lang.	Acc.(%)
[7]	BERT+MLM LSTM+DT	Multi-class	Drug Review	65 MB	Mono.	71.26
[5]	MuTCELM	Binary+ Multi	AJGT Corpus	12 MB	Multi.	93
[9]	Longformer	Binary	Fake News	8 MB	Mono.	78.4
[1]	KNN+CMD	Multi-class	20 News Groups	14 MB	Mono.	94
[2]	GZ Classifier	Multi-class	20 News Groups	14 MB	Mono.	76.7
[3]	FAGB	Multi-class	Kannada Corpus	30 MB	Multi.	94.1
[11]	ResGAT	Multi-class	Aminer Dataset	200 MB	Mono.	61
[12]	ConNHS	Multi-class	20 News Groups	14 MB	Mono.	87.43
[13]	KAN	Multi-class	European Patent	500 MB	Mono.	75.12
[17]	PatentSBERTa	Multi-class	Patents View	1 GB	Mono.	58
[16]	MFFMP-ETC	Binary	Stanford Sentiment	5 MB	Mono.	91.5
[23]	XGBoost+ KNN+LR	Multi-class	Security News	20 MB	Mono.	92.8
[27]	Qwen 2.5	Multi-class	Bangla News	60 MB	Mono.	72
[26]	SentBias	Binary	Hate Speech	25 MB	Multi.	89.84
[18]	SciBERT +MTL	Multi-class	PubMed 20K	50 MB	Mono.	93
[30]	PubMedBERT	Multi-class	RCT Corpus	45 MB	Mono.	92.9
[19]	TF-IDF +EVM	Multi-class	Crisis Dataset	2 GB	Mono.	87.93
Proposed Model	**MiniLM +MLP**	Multi-class	Aligned arXiv	895 MB	Multi.	**95.56**

Meanwhile, the addition of the statistical textual features provides further signals that emphasize the importance of terms and distributional aspects of documents. The combination of these heterogeneous representations forms a more informative feature space that allows the model to identify subtle differences between categories that would not be clear when the model uses a single method of representation.

The other important aspect of the model that makes it efficient is its balanced design in architecture which focuses on representational strength but maintains the computational complexity. The architecture is able to use fine-tuning of large-scale transformers but instead of doing it directly, it uses pre-computed semantic representations and blends them with a carefully designed neural classification system. This method has a huge impact on training overhead, but maintains the expressiveness of transformer-based embeddings. The layers that perform feature fusion and classification are optimized to learn the non-linear features between the combination of features to allow the system to bias the most informative signal per class. Consequently, the model is presented with a high generalization without a significant increase in the number of parameters or long training time.

The training strategy used in the model is also an important factor in its effectiveness. Balanced data handling, proper regularization, and optimized learning schedules are some of the techniques used to stabilize the training process, and avoid overfitting. Such mechanisms enable the classifier to acquire strong decision boundaries despite the changes in class distributions or overlapping textual patterns across classes. Moreover, the convergence of the model is guaranteed by the application of current optimization techniques that identify valuable patterns in the fused feature representations within a limited number of training epochs.

Practically, the architecture shows that high classification performance is not always achieved with models of very deep or very high computational complexity. The proposed method provides a high accuracy, efficiency, and scalability through a combination of semantic embeddings and the classical textual representations and a simplified neural network. The model is able to capture effectively both contextual and discriminative lexical information, and hence has a higher level of classification when it comes to a variety of textual inputs. Therefore, the findings suggest that the suggested design offers a computationally effective and reliable method of handling large-scale text classification processes, which explains the high results that were expressed throughout the evaluation process.

The discussion below emphasizes the key factors to the success of the proposed hybrid framework: the complementary nature of TF-IDF features, the minimal performance difference between Hindi models, and the benefits of the proposed framework over larger transformer-based models like XLM-RoBERTa.

Why TF-IDF Helps?
The improvements seen suggest that the embeddings from the context transformer are not sufficient to capture all the lexical cues of the domain. Scientific abstracts can include very discriminative terminology like algorithm names, mathematical terms and domainspecific terms. TF-IDF explicitly emphasizes these semantic abstracts, which are learned by MiniLM, as lexical indicators.

Why Hindi Slightly Lower?
The performance trend is similar in both English and Hindi, but a slight decrease in accuracy is seen in Hindi. This may be due to the artifacts of machine translation, the morphological differences and vocabulary normalization issues arising during the construction of the corpus.

Why Proposed Outperforms XLM-R?
While XLM-RoBERTa is larger and can process more parameters, it is still advantageous to include complementary lexical-statistical information in the proposed model, which allows to process more information.

Overall, the proposed hybrid framework is shown to be successful in producing robust, efficient, and consistently better results for bilingual scientific text classification tasks by leveraging semantic and lexical information.

VI. CONCLUSION

The present research offered an effective hybrid architecture of MiniLM-L12 + MLP for bilingual text classification task on aligned arXiv English-Hindi big data corpus. The results show that feature fusion of semantic and statistical features may be more effective than the bigger transformer

models and is also scalable. The proposed work enables decoupled bilingual semantic learning without expensive transformer fine-tuning and achieves comparable and language-agnostic performance in English and Hindi. The framework combines contextual embedding and statistical information to provide a good efficiency-accuracy trade-off in big data settings. Its' empirical superiority to strong baselines of multilingual models like mBERT and XLM-RoBERTa is also affirmed by empirical comparisons using fewer parameters. Collectively, the studies bridge a broad gap between the study of multilingual NLP and bilingual applied corpus analytics.

The future work will extend the suggested framework to multimodal scientific document comprehension, incorporating visual representations (figures, diagrams, and tables) into textual representations. Studies on the adaptive feature fusion process can also contribute to better learning of cross-modal representations and increase the strength of classification in massive academic collections. Moreover, the framework can be extended to facilitate more low-resource languages, which may enhance its role in scientific knowledge analysis in more languages. The investigation of parameters efficient transformer adaptation methods could also enhance the scalability further with the efficiency-accuracy ratio illustrated in this study.

ACKNOWLEDGMENT

The authors acknowledge the support of institutional research facilities used for conducting this study.

REFERENCES

[1] Q. Xu, "Application of an Intelligent English Text Classification Model with Improved KNN Algorithm in the Context of Big Data in Libraries," *Systems and Soft Computing*, p. 200186, 2025, doi: 10.1016/j.sasc.2025.200186.

[2] Y. Mao, Y. Ding, and T. Cui, "A parameter-free text classification method based on dual compressors," *Knowledge and Information Systems*, pp. 1–31, 2025, doi: 10.1007/s10115-024-02335-9.

[3] P. Y. Niranjan, V. S. Rajpurohit, and S. S. Sannakki, "Classification of Questions Using Machine Learning Techniques," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1–8, 2024.

[4] M. Pathak and A. Jain, "μboost: An effective method for solving indic multilingual text classification problem," in *Proc. IEEE Int. Conf. Multimedia Big Data (BigMM)*, 2022, pp. 96–100.

[5] V. K. Agbesi, W. Chen, S. B. Yussif, C. C. Ukwuoma, Y. H. Gu, and M. A. Al-Antari, "MuTCELM: An optimal multi-TextCNN-based ensemble learning for text classification," *Heliyon*, vol. 10, no. 19, 2024, doi: 10.1016/j.heliyon.2024.e38515.

[6] S. Malik and S. Jain, "Deep Convolutional Neural Network for Knowledge-Infused Text Classification," *New Generation Computing*, vol. 42, no. 1, pp. 157–176, 2024, doi: 10.1007/s00354-024-00245-6.

[7] S. Jamshidi, M. Mohammadi, S. Bagheri, H. E. Najafabadi, A. Rezvanian, M. Gheisari, M. Ghaderzadeh, A. S. Shahabi, and Z. Wu, "Effective text classification using BERT, MTM LSTM, and DT," *Data & Knowledge Engineering*, vol. 151, p. 102306, 2024, doi: 10.1016/j.datak.2024.102306.

[8] H. Padalko, V. Chomko, and D. Chumachenko, "A novel approach to fake news classification using LSTM-based deep learning models," *Frontiers in Big Data*, vol. 6, p. 1320800, 2024, doi: 10.3389/fdata.2023.1320800.

[9] S. Maham, A. Tariq, M. U. G. Khan, F. S. Alamri, A. Rehman, and T. Saba, "ANN: adversarial news net for robust fake news classification," *Scientific Reports*, vol. 14, no. 1, p. 7897, 2024, doi: 10.1038/s41598-024-56567-4.

[10] E. Gao, H. Yang, D. Sun, H. Xia, Y. Ma, and Y. Zhu, "Text classification optimization algorithm based on graph neural network," in *Proc. 2024 IEEE 6th Int. Conf. Power, Intelligent Computing and Systems (ICPICS)*, 2024, pp. 814–822.

[11] X. Huang, Z. Wu, G. Wang, Z. Li, Y. Luo, and X. Wu, "ResGAT: an improved graph neural network based on multi-head attention mechanism and residual network for paper classification," *Scientometrics*, vol. 129, no. 2, pp. 1015–1036, 2024, doi: 10.1007/s11192-023-04898-w.

[12] W. Ai, J. Li, Z. Wang, Y. Wei, T. Meng, and K. Li, "Contrastive multi-graph learning with neighbor hierarchical sifting for semi-supervised text classification," *Expert Systems with Applications*, vol. 266, p. 125952, 2025, doi: 10.1016/j.eswa.2024.125952.

[13] M. Cheon and C. Mun, "Towards Efficient Patent Classification: Kolmogorov Arnold Networks as an Alternative To MLP," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 21, 2024.

[14] A. Li and L. Zhang, "Multi-Label Text Classification Based on Label-Sentence Bi-Attention Fusion Network with Multi-Level Feature Extraction," *Electronics*, vol. 14, no. 1, p. 185, 2025, doi: 10.3390/electronics14010185.

[15] H. Joshi and S. Joseph, "ULMFIT: Universal Language Model Fine-Tuning for Text Classification," *International Journal of Advanced Medical Sciences and Technology*, vol. 4, no. 6, pp. 1–9, 2024, doi: 10.54105/ijamst.e3049.04061024.

[16] R. Zhang, "Multilingual pretrained based multi-feature fusion model for English text classification," *Computer Science and Information Systems*, pp. 1–10, 2025, doi: 10.2298/CSIS123456789X.

[17] H. Bekamiri, D. S. Hain, and R. Jurowetzki, "Patentsberta: A deep NLP based hybrid model for patent distance and classification using augmented SBERT," *Technological Forecasting and Social Change*, vol. 206, p. 123536, 2024.

[18] A. Brack, E. Entrup, M. Stamatakis, P. Buschermöhle, A. Hoppe, and R. Ewerth, "Sequential sentence classification in research papers using cross-domain multi-task learning," *International Journal on Digital Libraries*, vol. 25, no. 2, pp. 377–400, 2024, doi: 10.1007/s00799-023-00392-z.

[19] T. Jain, D. Gopalani, and Y. Kumar Meena, "Informative task classification with concatenated embeddings using deep learning on crisisMMD," *International Journal of Computers and Applications*, pp. 1–18, 2025, doi: 10.1080/1206212X.2024.2447066.

[20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019.

[21] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proc. ACL*, 2020.

[22] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-trained Transformers," in *Advances in Neural Information Processing Systems*, 2020.

[23] C. M. Liapis, K. Kyritsis, I. Perikos, N. Spatiotis, and M. Paraskevas, "A Hybrid Ensemble Approach for Greek Text Classification Based on Multilingual Models," *Big Data and Cognitive Computing*, vol. 8, no. 10, p. 137, 2024, doi: 10.3390/bdcc8100137.

[24] J. Kapočiūtė-Dzikienė and A. Ungulaitis, "Towards Media Monitoring: Detecting Known and Emerging Topics through Multilingual and Crosslingual Text Classification," *Applied Sciences*, vol. 14, no. 10, p. 4320, 2024, doi: 10.3390/app14104320.

[25] K. Feng, L. Huang, K. Wang, W. Wei, and R. Zhang, "Prompt-based learning framework for zero-shot cross-lingual text classification," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108481, 2024, doi: 10.1016/j.engappai.2024.108481.

[26] N. Lin, J. He, Z. Tang, J. Fang, D. Zhou, and A. Yang, "Model and evaluation: Towards fairness in multilingual text classification," *International Journal of Machine Learning and Cybernetics*, vol. 17, no. 2, p. 59, 2026.

[27] M. M. Hoque, M. M. Hassain, M. H. Tanvir, and R. Nandy, "Bengali text classification: An evaluation of large language model approaches," *arXiv preprint arXiv:2601.12132*, 2026.

[28] S. Mehra, V. Ranga, and R. Agarwal, "PhonoBiEmbedNet: A phoneme and bigram embedding framework for low-resource spoken word recognition," *Circuits, Systems, and Signal Processing*, pp. 1–35, 2026.

[29] S. P. K. Veeranki, A. Abdulnazar, D. Kramer, M. Kreuzthaler, and D. B. Lumenta, "Multi-label text classification via secondary use of large clinical real-world data sets," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-76424-8.

[30] M. Lan, M. Cheng, L. Hoang, G. Ter Riet, and H. Kilicoglu, "Automatic categorization of self-acknowledged limitations in randomized controlled trial publications," *Journal of Biomedical Informatics*, vol. 152, p. 104628, 2024, doi: 10.1016/j.jbi.2024.104628.

[31] N. Haupka, "Presenting a classifier to improve the identification of research journal publications in OpenAlex," *Scientometrics*, pp. 1–17, 2026.

[32] M. Limaylla-Lunarejo, N. Condori-Fernandez, M. Rodríguez Luaces, and O. Karras, "Improving the multi-class classification of non-functional requirements in Spanish," *Empirical Software Engineering*, vol. 31, no. 1, p. 6, 2026.

[33] J. Golde, N. Jedema, R. Krishnan, and P. Le, "Hierarchical text classification with LLM-refined taxonomies," *arXiv preprint arXiv:2601.18375*, 2026.

[34] R. Song, Y. Li, M. Tian, H. Wang, F. Giunchiglia, and H. Xu, "Causal keyword driven reliable text classification with large language model feedback," *Information Processing & Management*, vol. 62, no. 2, p. 103964, 2024, doi: 10.1016/j.ipm.2024.103964.