Computer Fraud and Security
ISSN (online): 1873-7056

Research on Vocabulary Optimization of Multilingual Models for Low-Resource Languages

First Author：**Zhenghang Tang**
Organization：Guangdong University of Technology
Email：13829007779@163.com
Address：Guangdong University of Technology, Xiaoguwei Street, Panyu District, Guangzhou City, Guangdong Province
Post Code：510000

Abstract:
To enhance the performance of multilingual models on low-resource languages, particularly in downstream tasks such as sentiment analysis, a framework for vocabulary expansion is proposed. This framework selects low-frequency but informative words using Zipf's Law and optimizes the vocabulary with weighted entropy analysis. Experimental results show improvements in accuracy and macro F1 scores by 3.85% and 5.22% respectively, particularly in tasks involving Hindi product reviews and Hindi-English code-switching. However, the study also notes limitations including performance fluctuations in intermediate stages of vocabulary expansion and a need for further exploration of the strategy's applicability to other NLP tasks. Despite these issues, the proposed framework provides a valuable method for enhancing the representation of low-resource languages in multilingual models.
Keywords: Sentiment Analysis, Zipf's law, Entropy, Low resources, Vocabulary

1 Introduction

Introduction When pre - training language models and fine - tuning them for different down- stream tasks, multilingual language models (MLLMs) such as mBERT or XLM - R [**?**] are often the top choices. These models have acquired rich language knowledge through learning from large- scale multilingual corpora. Research on multilingual language models has revealed that although they support numerous low - resource languages (LRLs), there is a significant order - of - magnitude difference in vocabulary allocation between low - resource languages and high - resource languages (HRLs) like English. Table 1 shows the quantitative differences between the vocabularies of various Indian languages in the mBERT vocabulary dictionary and those of English and Chinese. This difference may lead to a series of challenges in the use of low - resource languages.

Firstly, when words of LRLs cannot be effectively decomposed into word pieces by the vocabulary of MLLMs, they may be confused with the unknown (UNK) tokens, which will affect the model's comprehension and generation abilities. Secondly, although the word pieces of MLLMs may be fine - grained enough to combine into almost any LRL word, thus avoiding the direct appearance of unknown tokens, the embeddings of these word pieces may conflict due to irrelevant usage with HRLs or be too sparse during training. As a result, their integration in the context is insufficient to generate accurate LRL word embeddings. This situation may have an impact on the performance of tasks on LRLs. Although it is possible to invest a large amount of human and computational resources to establish large - scale LRL corpora for specific tasks to expand and train the vocabulary of MLLMs, this approach is often costly and not practical for researchers with limited resources. Therefore, exploring ways to enhance the support for LRLs during the pre - training stage and reduce the dependence on a large amount of specific corpora during fine - tuning has become a key direction for optimizing the performance of multilingual models.

In this study, we propose the "Zipf-Enhanced Lexicon Selection and Expansion Metric (ZELEM)". This method

combines Zipf's law [11] and the concept of weighted entropy of words and their con- stituent fragments [9]. ZELEM is specifically designed to guide the selection and expansion of low- resource language (LRL) lexicons. By implementing ZELEM, the adaptability of the pre-trained model to LRLs is enhanced, and the performance of the model in handling LRL sentiment analysis tasks is significantly improved.

Research shows that applying the "Zipf-Enhanced Lexicon Selection and Expansion Metric (ZELEM)" to existing multilingual large language models (MLLMs) during the fine-tuning stage can significantly affect the model's performance in various downstream classification tasks. These tasks cover multiple low-resource languages (LRLs) as well as code-mixed languages. Compared with the latest state-of-the-art research [5, 9], ZELEM's lexicon enhancement strategy brings more significant performance improvements to LRL tasks on most datasets.

2 Related Work

Continuous pre-training of existing multilingual models, such as mBERT [**?**] and XLM-R [**?**], with or without vocabulary expansion, can enhance the specific performance of the models in mul- tilingual tasks. This section will introduce two methods, namely pre-training without vocabulary expansion and pre-training with vocabulary expansion, and analyze their effects on the performance improvement of multilingual models.

2.1 Methods without Vocabulary Expansion

In the methods that do not involve vocabulary expansion, a large number of research works have been dedicated to solving the problems caused by rare or out-of-vocabulary (OOV) words, including the studies by Ruzzetti et al. [**?**] and Yu et al. [**?**]. Liu et al. [7] added an embedding generation module in the pre-training and fine-tuning processes, aiming to reduce the differences between words. In addition, the adapter-based methods explored by Sachidananda et al., Moon et al., and Hofmann et al. [4, 8, 14] avoided the initial stage of vocabulary expansion. Purkayastha et al. [13] found that transcribing from UTF-8 to Latin using the UROMAN tool can enhance the compatibility of multilingual large pre-trained models (mPLMs) with various low-resource languages. Perez et al. [10] aimed at the limitations of subword-based models in natural language processing (NLP) and sought solutions by aligning the word embedding layers of transformer models with fixed vocabularies with those of vocabulary-free models. Hofmann et al. [**?**] proposed a simple algorithm to maintain the integrity of the morphological structure of words by adjusting the word segmentation process. Although these methods have alleviated the problems to some extent, they have not fundamentally solved the domain and language-specific challenges of in-vocabulary token representation.

2.2 Methods of Vocabulary Expansion

In the methods involving vocabulary expansion, Beltagy et al. [1] trained a language model named SciBERT from scratch using a large amount of domain-specific corpora. This approach demonstrated that a vocabulary constructed based on specific corpora can significantly enhance the model's performance. Following this, Lee et al. [6] and Gururangan et al. [3] used large domain- specific corpora to conduct additional training on pre-trained language models to further optimize the models and prepare for the fine-tuning stage. Meanwhile, Poerner et al., Sato et al., and Tai et al. [12,15,16] made the models more closely meet the needs of specific domains by integrating domain- specific vocabularies into pre-trained models. In terms of multilingual tasks, Chung et al. [2] explored the method of creating multilingual vocabularies based on language groups, providing insights into understanding the diversity of languages. Notably, the entropy-based language model developed by Nag et al. [9] enhanced the vocabulary. However, these methods rely on word frequency for word selection, and may not fully consider the possible representation biases of the selected tokens within the model, which points to areas that require further exploration and improvement in future research.

Table 1 Representation of the vocabulary of various Indian languages in mBERT's wordpiece dictionary. *Based on basic to extended Latin script Unicode range.

Language	Vocabulary Size	Percentage (%)
Bengali	946	0.79
Hindi	1852	1.55
Kannada	653	0.55
Tamil	832	0.7
Telugu	887	0.74
Chinese*	13542	11.32
English*	47464	39.70

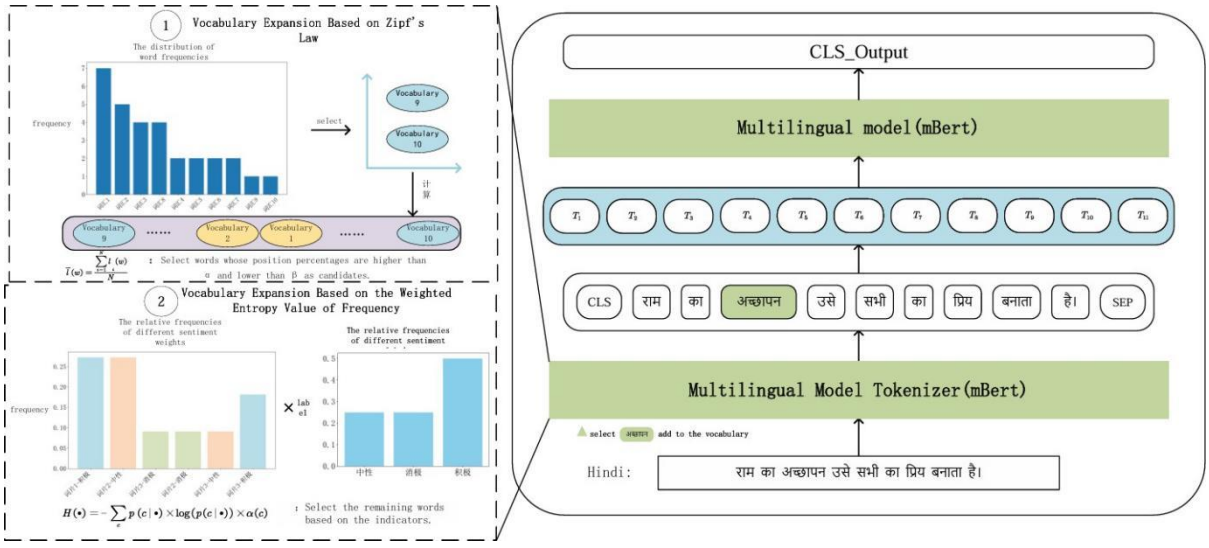


Figure 1 In the dashed box in the upper left corner, vocabulary added based on Zipf's Law is shown in the blue ellipse, representing rare nouns, and the orange ellipse indicates the results of these nouns being incorrectly segmented. For example, शांतिपुर usually conveys peace and tranquility, embodying positive emotions. If incorrectly segmented as "पर", this suffix might not carry any specific emotional connotation. If $\bar{p}(w)$ exceeds a certain threshold, then the word is chosen for inclusion in the vocabulary. The dashed box in the lower left corner displays vocabulary added based on the weighted entropy method, with the left bar chart calculating the frequency of word segments and labels, and the right showing label weights. After calculating weighted entropy, word segments with entropy values below the threshold are selected for the vocabulary. The right side shows the segmentation situation after expanding the vocabulary. For example, the sentence राम का अच्छापन उस सभी का प्रिय बनाता ह। conveys positive emotions, but if incorrectly segmented as अच्छा and पन, it might misinterpret its emotional polarity.

3 Method

3.1 Problem Definition

In this study, the vocabulary of the multilingual model is expanded by selecting and integrating appropriate vocabulary from low-resource languages. Denote V as the initial vocabulary of the multilingual model M， T as its tokenizer, and$|V|$ as the size of the vocabulary, that is, the total number of words. For a specific downstream task, let the number of classes be C， and the set of classes be $[C] = \{1, \ldots, C\}$， and use c to represent any one of these classes. The specific vocabulary set of the labels for the downstream task is denoted as $\boldsymbol{L} = \{l_1, l_2, \ldots, l_C\}$。

3.2 Overview of the Method

The proposed method consists of two main steps: Firstly, select vocabulary based on Zipf's law, briefly referred to as the "Zipf Selection Method" (ZSM); Secondly, calculate the frequency-based weighted entropy, briefly referred to as the "Frequency-Weighted Entropy Method" (FWEM). Finally, integrate the results of these two steps to determine the final selection of vocabulary. By analyzing the relationship between vocabulary and rare words, use ZSM to select these words and add them to the new vocabulary set. Subsequently, use FWEM to evaluate the weighted entropy of these words in downstream tasks, and select the remaining words based on this indicator. This method combines ZSM and FWEM to optimize and fine-tune the vocabulary. The specific frame- work is shown in Figure 1.

3.3 Vocabulary Enhancement Based on Zipf's Law

This section applies Zipf's Law to identify and enhance the vocabulary of low-resource languages in multilingual models. Zipf's Law states that in a specific corpus, the frequency of occurrence of a word is inversely proportional to its rank: words with higher ranks occur more frequently, while those with lower ranks occur less frequently. The mathematical expression of Zipf's Law is as follows:

$$f\ (w) = \frac{1}{rank(w)} \tag{1}$$

where rank(w) represents the position of the word w when all words are arranged in descending order of their frequencies of occurrence. By identifying words whose frequencies are lower than a certain proportion, we can uncover low-frequency words that might otherwise be overlooked. Typically, rare nouns with emotional polarity tend to appear at the beginning or the end of a sentence. Due to the low-frequency nature of these words, they may not be included in the vocabulary of multilingual models.

In order to discover these candidate words, first calculate the percentage of the relative position of each word in the sentence

$$l(w) = \left(\frac{pos(w)}{len(s)}\right) \times\ 100 \tag{2}$$

where pos(w) is the position of the word w in the sentence, and len(s) is the total length of the sentence. Next, calculate the average percentage of the position for each word:

$$\overline{l}(w) = \frac{\sum_{i=1}^{N} l_i(w)}{N} \tag{3}$$

where N is the total number of occurrences of the word w.

Based on the above analysis, select a set of words whose position percentages fall within a specific threshold range. These thresholds are defined by the parameters α and β, which represent the highest and lowest position percentages respectively. Select the words that are in the top α% and the bottom β% of the position percentages as candidates. These words usually play important roles in the structure of the sentence and the expression of emotions.

3.4 Vocabulary Enhancement Based on the Weighted Entropy Method

This section describes the vocabulary enhancement method based on the weighted entropy method, which is used to evaluate the feasibility of introducing low-resource language vocabulary into the multilingual vocabulary. In order to avoid incorrect splitting of words during tokenization, the complete low-resource language vocabulary is directly added to the vocabulary. If a word w is split into a sequence of tokens $s_1, \ldots, s_T$ by the tokenizer T, then the frequencies of its occurrence in different classes of the downstream task need to be considered.

Given the set of classes C for the downstream task, the normalized frequencies of the word w and the token s_t in each class $c \in C$ can be calculated:

$$p(c \mid \bullet) = \frac{n(\bullet, c)}{\sum_{c'} n(\bullet, c')}, \tag{4}$$

where • can be either the word w or the token s_t。

Next, use the weights of each class to adjust the calculation of the entropy. The weights are based on the reciprocals of the class frequencies and are normalized:

$$\alpha(c) = \frac{\frac{1}{f(c)}}{\sum_{c'} \frac{1}{f(c')}}, \tag{5}$$

where $\alpha(c)$ represents the weight ratio of each class, and $f(c)$ is the frequency of each class among all classes. With these weights, the entropy $H(\bullet)$is defined as:

$$H(\bullet) = -\sum_{c} p(c \mid \bullet) \times \log\big(p(c \mid \bullet)\big) \times \alpha(c) \tag{6}$$

Table 2 Key statistical information of task datasets, note the limited number of Low Resource Language (LRL) datasets.

Task Name	LRL(s)	Training Set
Sentiment Analysis of IITP Product Reviews	Hindi	4182
Sentiment Analysis of Bengali	Bengali	12576
Sentiment Analysis of GLUECos	Hi,En,code-mix	10079

This entropy value not only reflects the distribution of vocabulary across different classes but also takes into account the importance of different classes. If the entropy value of the word w is very low, it means that the word is highly correlated with a specific class and may be useful for a specific task. On the contrary, if the entropy value of the token s_t is very high, it indicates that the token is widely used in multiple tasks and may not be task-specific. Through this weighted entropy method, we can better evaluate the impact of low-resource language vocabulary on the multilingual vocabulary and its effectiveness in specific tasks.

4 Dataset

In this paper, experiments on three low-resource multilingual text sentiment classification tasks are carried out, covering two Indian languages (IITP product reviews, Bengali sentiment analysis) and a Hindi-English code-mixed dataset (GLUECos sentiment analysis). In addition, the impact of ZELEM's vocabulary expansion on the original model's tokenizer is verified on a Chinese dataset. The detailed information of the low-resource multilingual datasets is presented in Table 2.

5 Experiments and Analysis

5.1 Experimental Setup

The weights of the multilingual BERT base model are initialized using a truncated normal distribution with a standard deviation of 0.02, and the biases are set to 0. The experiment maintains a constant learning rate of 2e-5 and a maximum sequence length of 128 tokens. The training process consists of a total of 15 epochs, with a batch size of 16. The entire process is carried out on an NVIDIA A100 40-GB GPU. The details are shown in Table 3.

To accelerate the convergence of the model during training, the embeddings of the newly added low-resource language tokens are initialized. For the three language tasks, the initialization method described by Nag et al. [9] is adopted, that is, the existing low-resource language tokens in the MLLM vocabulary and their corresponding English translation tokens are used to initialize the embeddings of the new low-resource language tokens.

Hyperparameters	**Values**
mBERT version	bert-base-multilingual-cased
Batch Size	16, 32
Epoch	15
Learning Rate	2×10^{-5}, 5×10^{-5}
Maximum Sequence Length	128
θ	1
γ	25

Table 3 Hyperparameter settings for the mBERT model.

5.2 Metrics

The macro F1 score and accuracy are selected as the evaluation metrics for the three down- stream tasks, referring to the work of Nag et al. (2023) [9]. Accuracy is a quantitative metric for measuring the accuracy of a prediction or classification model, reflecting the proportion of correctly predicted instances in the dataset. The macro F1 score is the average of the F1 scores of all classes, with each class being assigned equal weight regardless of their frequencies. This approach ensures the balance of the evaluation, which is particularly important for dealing with imbalanced datasets. All experimental results are based on three random seeds.

5.3 Baseline Systems

Three methods are selected as the control benchmarks in this study:

Fine-tuning：Different from other control groups, fine-tuning adopts a direct strategy. Instead of expanding the model's understanding of language by adding extra words to the vocabulary, this method relies on the existing vocabulary and fine-tunes the parameters on the corpus of small low-resource language (LRL) tasks to better adapt to specific downstream tasks.

FLOTA：FLOTAintroduces an innovative tokenization strategy that aims to improve the per- formance of the tokenizer in pre - trained language models, rather than simply expanding the token library. Different from conventional tokenization methods, by preferentially selecting the longest possible tokens during the tokenization process, FLOTA effectively preserves the original morpho- logical structure of words and significantly reduces the information loss caused by over - tokeniza- tion. In addition, by favoring long tokens, FLOTA significantly enhances the tokenization process's resistance to whitespace noise and effectively reduces the occurrence of incorrect tokenization.

EVALM：EVALM implements an innovative task - aware metric method specifically for selecting the most vulnerable words in low - resource languages (LRLs). The core of this method is to use entropy as a measurement tool to accurately identify the vulnerable words in LRLs. By calculating the entropy value of words, it evaluates the information content and prediction difficulty of the vocabulary. A lower entropy value indicates that the word may be particularly crucial for LRL tasks. A higher entropy value of a token indicates a more balanced distribution across all classes, which may lead to excessive fragmentation of the vocabulary. EVALM calculates the average entropy value of all tokens by aggregating their data and uses the increase in entropy from LRL words to tokens as the basis for determining the risk of word decomposition. Subsequently, these words will be assigned appropriate initial embedding values and fine - tuned on the corpus of small - scale LRL tasks.

5.4 Analysis of Dataset Fragmentation

Dataset fragmentation refers to the situation in the text processing, where words are subdivided into smaller token units due to the mismatch of the vocabulary. This phenomenon is quantitatively characterized by the Word/Token Ratio in Table 4. The ratio is calculated by dividing the number of unique words by the total number of tokens. A lower ratio indicates a higher degree of fragmentation, meaning that the default vocabulary fails to fully cover the words in the text, resulting in an excessive number of tokens generated for each word. As shown in Table 4, the Hindi language content of the IITP product review dataset exhibits a moderate level of fragmentation, with the Word/Token Ratio slightly lower than 0.5. The GLUECos sentiment analysis dataset of Hindi-English Code-mix shows a slightly higher level of fragmentation, with the ratio being approximately 0.6. For the Bengali sentiment analysis dataset, the ratio is approximately 0.48, indicating that the fragmentation level is similar to that of the IITP product reviews. These ratios reflect the degree of dependence of each dataset on the native vocabulary, as well as the possible vocabulary optimization measures required when performing natural language processing tasks.

Dataset	Language	Word/Token Ratio		
		Training Set	Validation Set	Test Set
IITP Product Reviews	Hindi	0.50	0.49	0.50
GLUECos Sentiment Analysis	Hindi-English Code-mix	0.59	0.60	0.60
Bengali Sentiment Analysis	Bengali	0.48	0.48	0.48

Table 4 The fragmentation statistics of each dataset.

5.5 Main Experimental Results

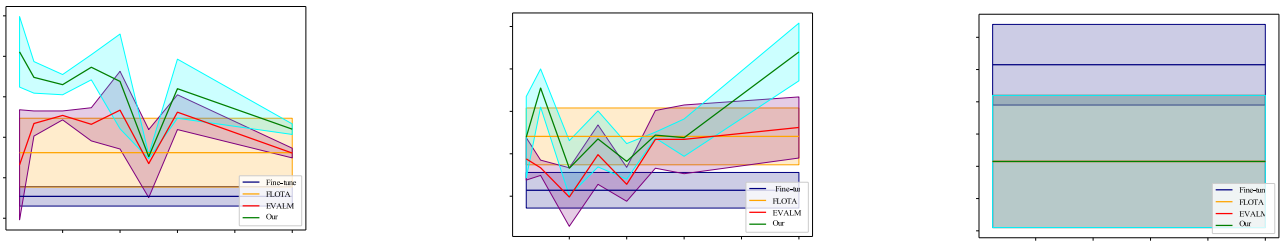
Through experimental analysis, it can be seen that ZELEM has been extensively compared and analyzed with three benchmark methods in three multi-class text classification tasks. As shown in Table 5, it is found that the average performance of ZELEM surpasses all baselines, achiev- ing the state-of-the-art (SOTA) level. When analyzing the performance improvements of different languages in multilingual tasks, the data in Table 5 reveals the performance differences of various methods, among which the ZELEM method stands out in some key tasks. ZELEM performs par- ticularly well in the IITP product review (Hindi) task, with the accuracy rate and macro F1 score reaching 76.06(±0.11) and 73.11(±0.87) respectively. This shows that ZELEM's vocabulary en- hancement strategy and fine model adjustment provide effective support for processing Hindi texts. In comparison, the accuracy rate and macro F1 score of the fine-tuning method are 72.21(±0.39) and 69.54(±0.24) respectively, indicating that ZELEM has obvious advantages in enhancing the vocabulary and optimizing the model's ability to understand specific texts. In the GLUECoS Hindi- English code-mixing task, the accuracy rate and macro F1 score of ZELEM are 61.92(±0.27) and 61.40(±0.68) respectively, demonstrating its effective processing ability for mixed-language data. Although ZELEM's performance in the Bengali sentiment analysis task is similar to that of other methods, it still shows certain advantages in handling highly specific and diverse language envi- ronments. This indicates that although the ZELEM method does not always perform the best, its overall strategy has broad applicability and effectiveness in optimizing multilingual models and dealing with low-resource languages. By integrating vocabulary optimization and weighted entropy analysis, ZELEM not only enhances the model's performance on low-resource languages but also improves the overall accuracy and adaptability in multilingual tasks. ZELEM has demonstrated significant performance improvements in multilingual multi-class text classification tasks. The ex- perimental results not only prove the efficiency of the ZELEM method but also showcase its broad potential in multilingual natural language processing applications.

5.6 Impact of Vocabulary Size

The experimental analysis includes the impact of the vocabulary size on performance. Figure-1 shows the correlation between the macro F1 score and the degree of vocabulary increase for the three tasks, ensuring that the vocabulary increases are comparable in scale[1)]. The yellow, blue, red, and green lines represent the performance of FLOTA, Fine-tune, EVALM, and ZELEM respectively, and the corresponding colored bands indicate the standard error. In the experimental settings of fine-tuning and FLOTA, the vocabulary size remains unchanged because fine-tuning and FLOTA are not affected by the increase in size. In contrast, the performance of EVALM and ZELEM is expected to improve with the increase in vocabulary size. However, the actual data 1) For detailed information on the relationship between accuracy and the degree of vocabulary increase in the three tasks, please refer to Appendix A.1.

Table 5 Main results.

Method	Accuracy	MacroF1
	IITP Product Reviews (Hindi)	
Fine-tuning	72.21(±0.39)	69.54(±0.24)
FLOTA	74.19(±0.90)	70.62(±0.85)
EVALM	75.21(±0.46)	71.67(±0.96)
ZELEM	**76.06(±0.11)**	**73.11(±0.87)**
	GLUECoS (Hindi-English code-mixing)	
Fine-tuning	59.74(± 0.60)	58.14(±0.42)
FLOTA	60.87(±0.66)	59.41(±0.67)
EVALM	61.30(±1.09)	59.62(±0.72)
ZELEM	**61.92(±0.27)**	**61.40(±0.68)**
	Bengali Sentiment Analysis (Bengali)	
Fine-tuning	**69.76(±0.21)**	**67.43(±0.25)**
FLOTA	69.21(±0.34)	66.83(±0.41)
EVALM	69.21(±0.34)	66.83(±0.41)
ZELEM	69.21(±0.34)	66.83(±0.41)
	average	
Fine-tuning	67.24(±0.25)	65.04(±0.18)
FLOTA	68.09(±0.39)	65.62(±0.39)
EVALM	68.57(±0.41)	66.04(±0.42)
ZELEM	**69.06(±0.15)**	**67.11(±0.39)**



(a) IITP Product Reviews (b) GLUECos Sentiment Analysis (c) Bengali Sentiment Analysis
Figure -1 Macro F1 vs. increasing low-resource language words added to MLLM dictionary.

shows that the performance does not always show a monotonically increasing trend. Especially in the IITP product review and GLUECoS tasks, there is no obvious positive correlation between the macro F1 score of EVALM and the expansion of the vocabulary size. This may indicate that the words selected to be added to the vocabulary do not have a sufficient direct correlation with the downstream tasks, resulting in performance fluctuations. For ZELEM, although the performance shows an improvement at the initial and final stages of vocabulary expansion, in the intermediate stage (especially in the range of vocabulary size from 1500 to 2500), the macro F1 score actually

decreases. This fluctuation may reflect that as the vocabulary size increases, some less relevant words are introduced, temporarily reducing the model's adaptability to specific tasks. Subsequently, when the vocabulary size continues to increase and more relevant words are added, the model performance recovers and improves. To gain a deeper understanding of the impact of vocabulary size on model performance, by analyzing the data in Tables 6, 7, and 8, we can have an in-depth understanding of the specific impact of vocabulary size on model performance. In these tables, the relationship between the macro F1 score and the degree of vocabulary increase shows various dynamics for different languages and tasks. In particular, for the IITP product review (Hindi), Table 6 shows that ZELEM shows performance improvement at the initial and final stages of vocabulary expansion, but there is a performance decline in the vocabulary range of 1500 to 2500. This fluctuation may imply that some words that are not closely related to the task may have been added during this stage. Subsequently, as the vocabulary size further increases, the model performance recovers and improves, demonstrating the importance of selecting the right words for performance optimization. For the GLUECoS Hindi-English code-mixing task, the data in Table 7 also reveals a similar pattern. The performance of ZELEM outperforms that of EVALM at certain vocabulary size points, especially when the vocabulary size is high, indicating the potential of its strategy in precisely matching task requirements. However, there is also a performance decline in the intermediate stage, emphasizing the need for continuous optimization of the vocabulary selection strategy. In the Bengali sentiment analysis task, Table 8 shows that there is no significant positive correlation between the macro F1 scores of EVALM and ZELEM and the expansion of the vocabulary size. This indicates that these methods may not have been fully optimized in terms of vocabulary selection, resulting in no obvious performance improvement.

These findings highlight the potential of significantly improving model performance through vocabulary enhancement strategies. At the same time, they also reveal the fluctuations that need to be noted during the process of vocabulary expansion, so as to ensure that the added words are closely related to the downstream tasks, thereby continuously improving the accuracy indicators across multiple tasks.

5.7 Performance on the Chinese Dataset

This section analyzes the performance of ZELEM and EVALM under different vocabulary sizes on the ChnSentiCorp hotel review dataset, with a focus on exploring the impact of these two vocabulary expansion methods on model performance. The main purpose of the experiment is to evaluate the impact of ZELEM's vocabulary expansion on the original model and compare it with EVALM. Especially when expanding the vocabulary, it examines how ZELEM can more

Table 6 IITP Product review (Hindi) under the influence of vocabulary size.

Method	Expanded Vocabulary Size	250	500	1000	1500	2000	2500	3000	5000
Fine-tuning	Macro F1						69.54		
	Standard Error						0.24		
FLOTA	Macro F1						70.62		
	Standard Error						0.85		
	Macro F1	70.32	71.34	71.54	71.32	71.67	70.35	71.62	70.61
	Standard Error	1.36	0.31	0.11	0.41	0.96	0.84	0.43	0.12

ZELEM	Macro F1	73.11	72.48	72.30	72.73	72.38	70.52	72.20	71.20
Standard Error		0.87	0.39	0.25	0.31	1.17	0.05	0.73	0.13

Table 7 GLUECoS (Hindi-English code-mix) under the influence of vocabulary size.

Method	Expanded Vocabulary Size	250	500	1000	1500	2000	2500	3000	5000
Fine-tuning	Macro F1						58.14		
Standard Error							0.42		
FLOTA	Macro F1						59.41		
Standard Error							0.67		
Error	Macro F1 Standard	58.88	58.67	57.98	58.98	58.28	59.34	59.34	59.62
		0.50	0.18	0.69	0.70	0.40	0.68	0.81	0.72
ZELEM	Macro F1	59.39	60.55	58.66	59.35	58.82	59.44	59.38	61.40
Standard Error		0.96	0.45	0.65	0.66	0.42	0.07	0.44	0.68

Table 8 Bengali Sentiment Analysis under the influence of vocabulary size.

Method	Expanded Vocabulary Size	250	500	1000	1500	2000	2500	3000	5000
Fine-tuning	Macro F1						67.43		
Standard Error							0.25		
FLOTA	Macro F1						66.83		
Standard Error							0.41		
Error	Macro F1 Standard	66.83	66.83	66.83	66.83	66.83	66.83	66.83	66.83
		0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41
ZELEM	Macro F1	66.83	66.83	66.83	66.83	66.83	66.83	66.83	66.83
Standard Error		0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41

Table 9 ChnSentiCorp hotel reviews under the influence of vocabulary size.

Method	**Expanded Vocabulary Size**	**250**	**500**	**1000**	**1500**	**2000**	**2500**
EVALM	Accuracy	69.17	71.00	74.57	73.23	73.30	68.47
Standard Error		0.54	0.50	0.11	0.14	0.34	0.59
ZELEM	Accuracy	69.40	72.13	75.80	69.27	67.40	72.27

Standard Error		0.31	0.11	0.35	0.59	0.54	0.43

Table 10 ChnSentiCorp hotel reviews under the influence of vocabulary size (Macro-F1).

Method	**Expanded Vocabulary Size**	**250**	**500**	**1000**	**1500**	**2000**	**2500**
EVALM	F score	56.13	58.03	66.87	58.20	57.67	57.57
Standard Error		0.13	0.19	0.05	0.06	0.15	0.12
ZELEM	F score	43.80	59.03	54.63	56.77	57.60	67.40
Standard Error		0.03	0.12	0.12	0.14	0.19	0.04

effectively maintain or improve the model performance. The experimental results show that under most vocabulary size settings, ZELEM exhibits more stable or superior performance compared to EVALM. Particularly when the vocabulary size is large, ZELEM often outperforms EVALM, demonstrating its precision in selecting words highly relevant to the task, thus having less impact on the core performance of the original model. As shown in Table 9, when the vocabulary size is expanded to 1000 and 2500, the accuracy rates of ZELEM reach 75.80These results highlight the advantage of ZELEM, which has less impact on the model performance during vocabulary expansion. ZELEM's strategy performs more outstandingly in maintaining model stability and improving performance, especially when dealing with Chinese sentiment analysis tasks that require a large vocabulary to understand complex contexts. The effectiveness of this strategy is not only reflected in improving the model's accuracy and macro F1 score, but also in its ability to adapt to a wider range of vocabulary and context changes without sacrificing the performance of the original model, demonstrating its potential application value in the field of multilingual processing.

5.8 Case Study

In this section, through a practical case study, the effectiveness of the ZELEM method in recognizing the sentiment of complex texts in different Chinese and Hindi datasets is demonstrated. Table 11 below lists in detail several reviews selected from three datasets and their sentiment tendencies. All entries are the results of correct recognition by ZELEM. From the ChnSentiCorp hotel reviews, to the Hindi-English code-mixed texts in GLUECoS, and then to the pure Hindi texts in the IITP product reviews, ZELEM can accurately identify reviews with different sentiment tendencies. This ability demonstrates its high adaptability to complex contexts and multilingual environments, especially when dealing with challenging code-mixing and languages using complex scripts.

The ZELEM method is not only applicable to sentiment analysis in a single language, but also performs excellently in multilingual and multicultural contexts. This cross-cultural and cross-lingual adaptability makes it an ideal choice for processing internationalized data, providing an effective sentiment recognition tool for future language processing technologies.

5.9 Ablation Experiments

This section explores two key techniques, the Frequency-Weighted Entropy Method (FWEM) and the Zipf Selection Method (ZSM) through ablation experiments.

In the IITP product review (Hindi) task, according to the data in Table 12, the configuration with the application of FWEM (w/o FWEM) shows relatively stable performance, highlighting the importance of FWEM in adjusting and optimizing the vocabulary to enhance the sensitivity to the characteristics of the task. At the same time, the

configuration with the application of ZSM (w/o ZSM) shows a gradually increasing performance trend as the vocabulary expands, indicating that

Table 11 Emotion recognition cases of ZELEM method on different datasets.

Review Content	Sentiment Tendency	Dataset Source
The hotel has a great location, convenient transportation, good environment, excellent service and fast laundry service.	Positive	ChnSentiCorp Hotel Reviews
Convenient transportation, beautiful environment, very quiet, the air is also fresh. Additional comments May 4, 2008: Reasonable price	Positive	ChnSentiCorp Hotel Reviews
What a shady shop!!! Checking out at 12:30 will charge you half a day's room rate. The room is very dark!!! It's quite old!!!	Negative	ChnSentiCorp Hotel Reviews
not funny -_-	Negative	GLUECoS (Hindi-English Code-mixed)
just like when k was a kid saturday morning and i'm watching saved by the bell ! might even watch soul train @ 12	Neutral	GLUECoS (Hindi-English Code-mixed)
Camera making company Nikon has added to its 1 series cameras and has come up with the best camera of this series.	Positive	IITP Product Reviews (Hindi)
Both these tablets will work on Android 4.2.2 version operating system.	Neutral	IITP Product Reviews (Hindi)
Popular singer Ankit Tiwari is also disappointed.	Negative	IITP Product Reviews (Hindi)

Table 12 Ablation experiment of IITP Product review (Hindi) under the influence of vocabulary size.

Method	Expanded Vocabulary Size	250	500	1000	1500	2000	2500	3000	5000
w/o ZSM	Accuracy	73.30	73.12	74.12	74.25	75.21	73.23	74.19	74.51
Macro F1		70.32	71.34	71.54	71.32	71.67	70.35	71.62	70.61
w/o FWEM	Accuracy	72.21	72.21	72.21	72.21	72.21	72.21	72.21	72.21
Macro F1		69.54	69.54	69.54	69.54	69.54	69.54	69.54	69.54
ZELEM	Accuracy	75.27	74.82	75.08	74.80	76.06	73.93	74.66	74.68
Macro F1		73.11	72.48	72.30	72.73	72.38	70.52	72.20	71.20

Table 13 Ablation experiment of GLUECoS (Hindi-English code-mix) under the influence of vocabulary size.

Method	Expanded Vocabulary Size	250	500	1000	1500	2000	2500	3000	5000
w/o ZSM	Accuracy	60.16	60.12	60.27	60.37	59.60	61.30	60.66	61.08
Macro F1		58.88	58.67	57.98	58.98	58.28	59.34	59.34	59.62
w/o FWEM	Accuracy	59.74	59.74	59.74	59.74	59.74	59.74	59.74	59.74
Macro F1		58.14	58.14	58.14	58.14	58.14	58.14	58.14	58.14
ZELEM	Accuracy	61.84	61.05	61.53	60.61	59.84	61.92	61.47	61.22
Macro F1		59.39	60.55	58.66	59.35	58.82	59.44	59.38	61.40

ZSM has a positive effect on improving the model performance. Similar trends are also observed in the ablation experiments of the GLUECoS (Hindi-English code-mixing) task. The configuration without the application of FWEM generally shows lower performance, once again emphasizing the key role of FWEM in enhancing the model's adaptability. While the configuration without the application of ZSM reaches an accuracy close to that of the complete ZELEM method at some vocabulary size points, the overall performance is still lower than that of the configuration using the complete ZELEM method.

As can be seen from the data in Table 12 and Table 13, both FWEM and ZSM significantly improve the performance of the model. By combining these two techniques, the ZELEM method can optimize performance under different vocabulary sizes and dynamically adjust the vocabulary according to specific tasks, thus achieving higher accuracy and macro F1 scores. These results highlight the comprehensive role of FWEM and ZSM in improving the performance of multilingual models and demonstrate the effectiveness and necessity of these methods when dealing with di- verse language data. By precisely adjusting the vocabulary, ZELEM can adapt to various complex language environments, which is crucial for improving the applicability and accuracy of language models. These ablation experiments not only show the importance of FWEM and ZSM respectively but also emphasize their complementary roles in improving the overall model performance.

6 Conclusions and Prospects

In this study, through the ZELEM method that combines ZSM and FWEM, the relevant vo- cabulary of low-resource

languages has been successfully incorporated into the vocabulary of the multilingual model. In all tasks, ZELEM, with an average macro F1 score of 67.11 and an accuracy of 69.06, proves its effectiveness as a vocabulary enhancement strategy in low-resource language envi- ronments. The achievements of ZELEM highlight the key role of semantic and frequency indicators in the vocabulary selection process. The results of this work not only demonstrate the practicality of the ZELEM method but also highlight its great potential in improving the capabilities of multilingual models, especially for languages with limited resources.

Limitations

Although the framework has been proven effective, its limitations need to be recognized when conducting evaluations and interpretations. The scope of this study is limited to pre-trained models and does not cover the application of large language models.

References

1. Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, 2019.
2. Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. Improving multilingual models with language-clustered vocabularies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4536–4546, 2020.
3. Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.
4. Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. Superbizarre is not superb: Derivational morphology improves bert's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, 2021.
5. Valentin Hofmann, Hinrich Schuetze, and Janet B Pierrehumbert. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. Association for Computational Linguistics, 2022.
6. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
7. Xin Liu, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Min Zhang, Haiying Zhang, and Jinsong Su. Bridging subword gaps in pretrain-finetune paradigm for natural language generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6001–6011, 2021.
8. Sangwhan Moon and Naoaki Okazaki. Patchbert: Just-in-time, out-of-vocabulary patching. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7846–7852, 2020.
9. Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. Entropy-guided vocabulary augmentation of multilingual language models for low-resource tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8619–8629, 2023.
10. Alejandro Rodriguez Perez, Korn Sooksatra, Pablo Rivas, Ernesto Quevedo Caballero, Javier S. Turek,

Gisela Bichler, Tomas Cerny, Laurie Giddens, and Stacie Petter. An empirical analysis towards replacing vocabulary-rigid embeddings by a vocabulary-free mechanism. In *LatinX in AI Workshop at ICML 2023 (Regular Deadline)*, 2023.

11. Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130, 2014.
12. Nina Poerner, Ulli Waltinger, and Hinrich Schütze. Inexpensive domain adaptation of pretrained language models: Case studies on biomedical ner and covid-19 qa. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1482–1490, 2020.
13. Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. Romanization-based large-scale adaptation of multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7996–8005, 2023.
14. Vin Sachidananda, Jason Kessler, and Yi-An Lai. Efficient domain adaptation of language models via adaptive tokenization. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 155–165, 2021.
15. Shoetsu Sato, Jin Sakuma, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. Vocabulary adaptation for domain adaptation in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4269–4279, 2020.
16. Wen Tai, HT Kung, Xin Luna Dong, Marcus Comiter, and Chang-Fu Kuo. exbert: Extending pre-trained models with domain- specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, 2020.

A Appendix

A.1 Impact of Vocabulary Size

This section discusses the impact of the vocabulary size on the accuracy performance of the model. As shown in Figure A-2, there is no obvious positive correlation between the accuracy of EVALM and the expansion of the vocabulary size, which further confirms the instability of its performance. In contrast, although the performance of ZELEM is generally positively correlated with the vocabulary expansion, it does not increase completely monotonically throughout the process. Especially in certain vocabulary size intervals, the performance first rises, then drops, and finally rises again, surpassing other methods and reaching the state-of- the-art (SOTA) level. This dynamic indicates that ZELEM demonstrates high efficiency in accurately identifying words that are directly related to downstream tasks, and its vocabulary enhancement strategy significantly improves the overall performance of the model.

In the experiment of IITP product reviews (Hindi), as shown in Table A1, the performance of the ZELEM method improves during the initial expansion of the vocabulary size, but experiences a performance decline in the intermediate stage (especially in the interval where the vocabulary size is between 1500 and 2500). Later, as the vocabulary size further increases, the accuracy rate rises again. This fluctuation may be related to the introduction of words with lower relevance within a specific vocabulary size interval, which temporarily reduces the model's adaptability to the task. The data of the GLUECoS Hindi-English code-mixing task, as shown in Table A2, shows that the accuracy rate of ZELEM experiences similar fluctuations, indicating that the processing of complex datasets requires precise and highly relevant word selection to maintain performance stability. For the Bengali sentiment analysis, as shown in Table A3, the accuracy performance of ZELEM is relatively stable and does not show significant fluctuations. This indicates that for certain specific tasks, a single vocabulary expansion strategy may not be sufficient to significantly improve performance.
These results emphasize that when expanding the vocabulary, in addition to increasing the quantity, it is equally important to ensure the quality of the vocabulary and its close relevance to the task. The ZELEM

method can effectively utilize vocabulary expansion to improve performance in most cases through its strategy, but it also reveals the necessity of making meticulous adjustments to word selection to avoid performance degradation. This highlights the crucial role of precise word selection and timely strategy adjustment in improving the performance of multilingual models.

Table A1 IITP Product review (Hindi) under the influence of vocabulary size.

Method	Expanded Vocabulary Size	250	500	1000	1500	2000	2500	3000	5000
Fine-tuning	Accuracy						72.21		
Standard Error							0.39		
FLOTA	Accuracy						74.19		
Standard Error							0.90		
	Accuracy	Standard	73.30	73.12	74.12	74.25	75.21	73.23	74.19	74.51
Error			1.38	0.37	0.17	0.17	0.81	0.61	0.44	0.13
ZELEM	Accuracy		75.27	74.82	75.08	74.80	76.06	73.93	74.66	74.68
Standard Error			0.34	0.51	0.13	0.29	0.11	0.23	0.06	0.44

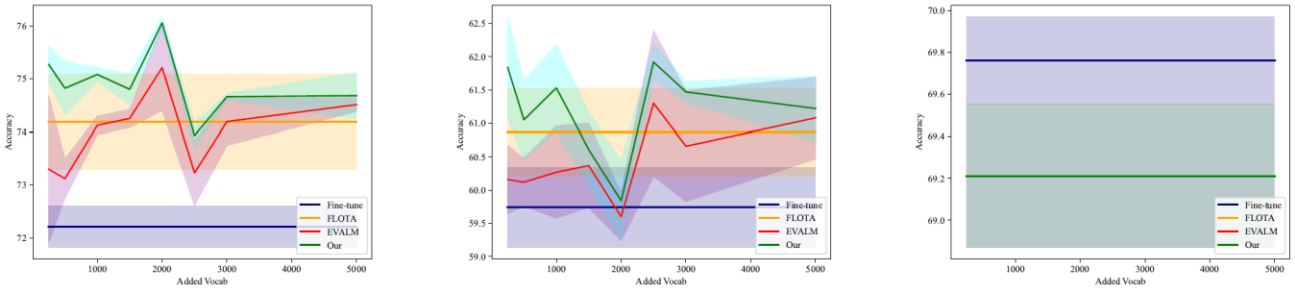
Table A2 GLUECoS (Hindi-English code-mix) under the influence of vocabulary size.

Method	Expanded Vocabulary Size	250	500	1000	1500	2000	2500	3000	5000
Fine-tuning	Accuracy						59.74		
Standard Error							0.60		
FLOTA	Accuracy						60.87		
Standard Error							0.66		
	Accuracy	Standard	60.16	60.12	60.27	60.37	59.60	61.30	60.66	61.08
Error			0.51	0.36	0.69	0.63	0.36	1.09	0.83	0.61
ZELEM	Accuracy		61.84	61.05	61.53	60.61	59.84	61.92	61.47	61.22
Standard error			0.76	0.59	0.65	0.55	0.60	0.27	0.16	0.49

Table A3 Bengali Sentiment Analysis under the influence of vocabulary size.

Method	Expanded Vocabulary Size	250	500	1000	1500	2000	2500	3000	5000
Fine-tuning	Accuracy						69.76		
Standard Error							0.21		

FLOTA	Accuracy				69.21				
Standard Error					0.34				
	Accuracy	Standard69.21	69.21	69.21	69.21	69.21	69.21	69.21	69.21
Error		0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34
ZELEM	Accuracy	69.21	69.21	69.21	69.21	69.21	69.21	69.21	69.21
Standard Error		0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.34



(a) IITP Product Review (Hindi) **(b)** GLUECoS (Hindi-English code-mix) **(c)** Bengali Sentiment Analysis

Figure A-2 Accuracy vs. increasing low-resource language words added to MLLM dictionary.