MLCL: MultiLingual Contrastive Learning Framework for Reducing Language Imbalance in Sino-Tibetan Languages

Jiajun Fang 2112205248@mail2.gdut.edu.cn
School of Computer Science and Technology,
Guangdong University of Technology, Guangzhou, 510006, Guangdong, China

Wentao Huang 3216958687@qq.com
School of Computer Science and Technology,
Guangdong University of Technology, Guangzhou, 510006, Guangdong, China

Aimin Yang amyang18@163.com
School of Computer Science and Intelligence Education,
Lingnan Normal University, Zhanjiang, 524000, Guangdong, China

Dong Zhou✉ dongzhou@gdufs.edu.cn
School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, 510006, Guangdong, China

Nankai Lin✉ neakail@outlook.com
School of Computer Science and Technology,
Guangdong University of Technology, Guangzhou, 510006, Guangdong, China

Abstract

Multilingual pre-trained models have been widely applied in natural language processing (NLP) tasks, including text classification. However, due to the varying amounts of language resources, these models exhibit performance imbalance across different languages, a phenomenon known as language imbalance. Existing research on mitigating language imbalance primarily harnesses text and image data, neglecting the auditory aspects of languages. This neglect results in an incomplete solution to language imbalance, as it fails to exploit the rich linguistic nuances conveyed through speech. To address these issues, this paper introduces a novel framework called **M**ulti**L**ingual **C**ontrastive **L**earning **(MLCL)** to reduce language imbalance. By incorporating concepts from comparative linguistics into neural networks, we utilize the phonetic similarities among languages within the Sino-Tibetan family to tackle the problem of language imbalance in multilingual pre-trained models. To evaluate our method's effectiveness, we conducted tests using two synthetic datasets derived from the Flores200 and mms datasets across various models. The experimental results show that, in terms of language imbalance metrics, our model surpasses all baseline models.

Keywords: Text Classification; Language Imbalance; Multilingual Pre-trained Model; Contrastive Learning; Comparative Linguistics; Sino-Tibetan Language Family

1. Introduction

With the development of artificial intelligence, current models (e.g., GPT (Brown et al., 2020)) are evolving to support multi-modal and multilingual capabilities. However, these multilingual pre-trained models primarily focus on high-resource languages and show weaker performance in low-resource languages, emphasizing the importance of enhancing model performance in these languages (Haddow et al., 2022).

However, existing research on enhancing the performance of multilingual pre-training models in low-resource languages often relies solely on text (Gou et al., 2023) or image data (Dutta Chowdhury et al., 2018), with few studies leveraging the phonetic similarities across different languages. The oversight of phonetic similarities between languages is significant, as these similarities can greatly enhance the understanding and performance of models in processing underrepresented languages.

The study by Wu and Dredze (2020) revealed that the multilingual pre-trained model, mBERT (Devlin et al., 2019), exhibits uneven performance across different languages, highlighting the persistent issue of language imbalance. In studies focusing on low-resource languages (Lee et al., 2023), enhancing performance often results in reduced effectiveness in languages with rich resources. This trade-off underscores the fundamental challenge of language imbalance, yet they lacks a comprehensive metric for assessing this imbalance effectively, pointing to a gap in evaluating the equitable performance of models across various languages.

Therefore, our work focuses on leveraging the inherent phonetic similarities among different languages within the Sino-Tibetan language family to address language imbalance issues when performing text classification across multiple languages. Additionally, to better measure the overall capability of models in multilingual contexts, we have incorporated the Gini coefficient, a metric from economics, into our work. The metric framework we proposed offers a single, standardized measure that facilitates the comparison of different models' overall performance on a specific group of languages. This framework helps in quantitatively assessing and benchmarking the level of language imbalance.

In this research, our primary contributions are as follows:

1. We integrate the concepts of contrastive linguistics into neural networks, focusing on the phonetic similarities among languages within the Sino-Tibetan family. By employing a multi-modal approach, we leverage these similarities to address the language imbalance issue in multilingual pre-trained models, presenting an innovative solution to the challenge of language imbalance.

2. We present a framework for evaluating language imbalance in multilingual text classification that encompasses two dimensions. This metric framework is designed to reflect the model's imbalance across different languages as well as the mean accuracy of the model across these languages.

3. We propose a novel framework called **M**ulti**L**ingual **C**ontrastive **L**earning **(MLCL)**. Within this framework, we incorporate two distinct contrastive learning tasks: one focused on language and the other on modality. The overarching goal of these tasks is to minimize feature disparities within identical modalities and languages. These tasks capitalize on the inherent similarities in speech patterns to enhance text classification.

2. Related Work

2.1. Contrastive Learning

Contrastive learning, a growing technique, is extensively explored and applied in diverse fields to learn representations. Pan et al. (2021) enhanced translation quality by leveraging contrastive learning to fine-tune BERT, focusing on reducing the performance gap between English and non-English language pairs. Su et al. (2022) used contrastive learning for multi-label text classification with imbalanced and sparse labels, boosting the model's capacity to differentiate between similar labels through better representation learning. Pan et al. (2022) combined adversarial training with contrastive learning to enhance text classification models' robustness and generalization, showing how adversarial examples improve discriminative text feature learning.

Some studies have used contrastive learning for multimodal tasks. For example, CLIP (Radford et al., 2021) improved image-text semantic links by optimizing similarities of matched pairs and differences of unmatched ones, showing strong zero-shot learning results. Similarly, Lin and Hu (2022) combined single-mode coding and cross-modal prediction in a multimodal contrastive learning framework (MMCL) to enhance sentiment analysis accuracy by capturing both intra-modal and inter-modal dynamics.

2.2. Research On Boosting Low-Resource Language Performance

By boosting the performance of models on low-resource languages, the issue of language imbalance can be alleviated. Elbayad et al. (2023) tackled the problem of MoE models over-fitting in low-resource languages by introducing effective regularization strategies, significantly improving performance without negatively impacting high-resource tasks. Gou et al. (2023) proposed a novel framework that uses cross-lingual data augmentation to improve dialogue generation in low-resource languages, leveraging high-resource languages for enhanced performance. Hangya et al. (2022) introduced an unsupervised method to boost cross-lingual representations for low-resource languages, achieving performance enhancements in both word representation quality and downstream tasks using only non-parallel resources. Chronopoulou et al. (2023) proposed language-family adapters on mBART-50, significantly improving translation for low-resource languages and extending support to languages beyond the initial pretraining scope.

2.3. Research On Correlations Across Modalities and Languages

The composition of language is intricate, involving both textual and auditory components. CLAP (Elizalde et al., 2023) introduces a model that learns audio concepts from natural language supervision, integrating text and audio for improved multimodal understanding.

Researchers have also extensively explored the relevance of texts across diverse languages. For instance, Gey (2005) delves into the similarities and disparities between Chinese and Japanese within the realm of cross-lingual information retrieval (CLIR), illuminating both the practical applications and inherent challenges of CLIR technologies in transcending language barriers.

Auditory information is key for capturing emotional nuances in communication. It conveys emotions through intonation, stress, and rhythm, which text alone cannot fully

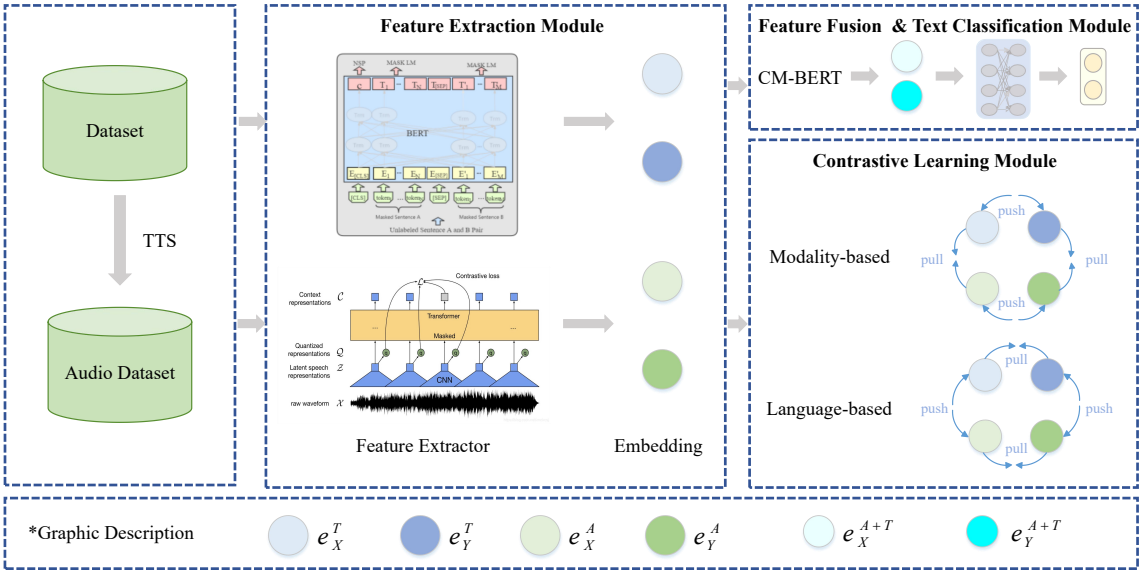

Figure 1: The MLCL framework.

express. This makes auditory data crucial for text classification tasks. In Multimodal Sentiment Recognition (MSR) (Zhu et al., 2023), combining audio and text data provides a fuller understanding of emotional context. Additionally, comparative linguistics studies the similarities and differences between languages. It helps understand their development and relationships, particularly within the Sino-Tibetan languages.

Compared to existing research, our work uniquely combines comparative linguistics and contrastive learning to address the language imbalance issue in multilingual pre-trained models for the Sino-Tibetan language family. Different from the existing research mentioned above, our study focuses on exploiting phonetic links and the integration of auditory and textual information to improve the model's performance on low-resource languages, thereby mitigating the language imbalance issue.

3. Method

3.1. Problem Definition

In our task, the input consists of textual and acoustic modalities, where the input modality $m \in \{T, A\}$. The sequences of these two modalities are represented as pair(T, A), including the textual modality $T \in \mathbb{R}^{N_t \times d_t}$ and acoustic modality $A \in \mathbb{R}^{N_a \times d_a}$ where N_m denotes the sequence length of corresponding modality and d_m denotes the dimensionality. The objective of this task is to learn a mapping $f(T, A)$ for inferring the classification score $\hat{y} \in \mathbb{R}$.

3.2. Overall Architecture

In Figure 1, we present the **M**ulti**L**ingual **C**ontrastive **L**earning **(MLCL)** framework, which is specifically designed to enhance classification accuracy and reduce imbalances in the Sino-Tibetan language family, focusing on three languages: Chinese, Tibetan, and Burmese. This framework consists of four principal components: feature extraction, contrastive learning, feature fusion, and classification. For feature extraction, we utilize a multilingual pre-trained model (mBERT) to capture textual features. We employ two Text-to-Speech (TTS) tools for audio generation: gTTS[1] for Chinese and Burmese, and MaryTTS[2] for Tibetan. Additionally, in the feature extraction phase, we have used the wav2vec 2.0 model, a multilingual audio pre-trained model, by incorporating an adapter structure to extract features from the raw audio signals generated by these tools. Through contrastive learning, the framework adeptly minimizes the distance between differing modalities within the same language and similar modalities across different languages. This method, further enhanced by the CM-BERT (Yang et al., 2020) technique for feature fusion, ensures a deep integration of textual and audio features.

Given a sample set $D = \{(T_1^X, T_1^Y, A_1^X, A_1^Y, y_1), (T_2^X, T_2^Y, A_2^X, A_2^Y, y_2), ..., (T_n^X, T_n^Y, A_n^X, A_n^Y, y_n)\}$, where T_i^X and A_i^X are the representations of sample i in the textual modality (T) and acoustic modality (A) with language X, T_i^Y and A_i^Y in modality T and A with language Y. y_i is the desired label of sample i, where $y_i \in Y$. Y denotes the category set of the text classification task.

3.3. Feature Extraction Module

The mBERT model, using Transformer architecture and trained on 102 languages, enables cross-lingual text classification and sentiment analysis through unified semantic encoding. Therefore, our feature extraction module utilizes mBERT as the foundational model to obtain sentence vector features of different languages. For the i-th text input T_i^X and T_i^Y of languages X and Y respectively, their corresponding feature vector representations are as follows:

$$e_X^{T_i} = mBERT(T_i^X), \tag{1}$$

$$e_Y^{T_i} = mBERT(T_i^Y). \tag{2}$$

The wav2vec 2.0 (Baevski et al., 2020) model, based on the Transformer architecture, is pre-trained on a large amount of unlabeled audio data to learn general audio representations and is commonly used for audio processing and Automatic Speech Recognition (ASR) tasks. Therefore, our feature extraction module utilizes wav2vec 2.0 as the foundational model to obtain speech vector features of different languages.

Adapter modules are compact modules added to a pre-trained Transformer model and fine-tuned for downstream tasks, with the original model parameters remaining unchanged. The architecture of the adapter module is shown in Figure 2.

Inspired by Hou et al. (2021b), we incorporate the adapter structure into the transformer architecture of the wav2vec2 model. In our experiments, we employed two types of

1. https://gtts.readthedocs.io/
2. https://marytts.github.io/

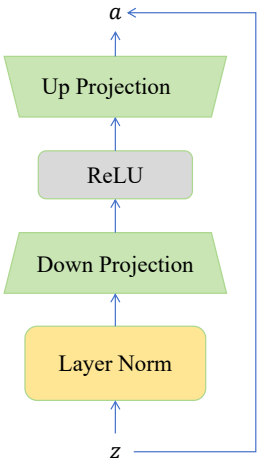

Figure 2: Architecture of the adapter module.

structures: MetaAdapter (Hou et al., 2021a) and SimAdapter (Hou et al., 2021b). For the i-th audio input A_i^X and A_i^Y of languages X and Y respectively, their corresponding feature vector representations are as follows:

$$e_X^{A_i} = wav2vec2(A_i^X), \tag{3}$$

$$e_Y^{A_i} = wav2vec2(A_i^Y). \tag{4}$$

3.4. Contrastive Learning Module

In the contrastive learning module, we introduced two sub-tasks: First, "Language-based Contrastive Learning (LCL)", which aims to reduce the feature distance between different languages within the same modality; Secondly, "Modality-based Contrastive Learning (MCL)", designed to diminish the feature disparities across different modalities within the same language. By reducing linguistic distances in LCL, the model achieves a more generalized representation that effectively bridges linguistic gaps, enhancing performance across diverse languages. Similarly, by aligning features from different modalities in MCL, such as text and speech, the model leverages complementary information, increasing its robustness and accuracy in processing complex linguistic inputs. These efforts allow our model to capture shared linguistic features more effectively, aiding in the recognition of low-resource languages and mitigating language imbalance.

In each training batch, we randomly select N pairs of samples to construct a mini-batch. For the i-th input, the language-based contrastive learning loss L_{LCL} can be given by:

$$\text{sim}(e_X, e_Y) = e_X \cdot e_Y, \tag{5}$$

$$\text{L}(e_X, e_Y) = -\log \frac{\exp(\text{sim}(e_X, e_Y)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(e_X, e_Y^i)/\tau)}, \tag{6}$$

$$\mathrm{L}_{LCL} = L(e_X^T, e_Y^T) + L(e_X^A, e_Y^A), \tag{7}$$

where e_X and e_Y can be e_X^T and e_Y^T , or e_X^A and e_Y^A . τ controls the temperature and $\mathrm{sim}(\cdot)$ denotes the cosine similarity.

Much like L_{LCL} ,the modality-based contrastive learning loss L_{MCL} can be given by:

$$\mathrm{sim}(e^A, e^T) = e^A \cdot e^T, \tag{8}$$

$$\mathrm{L}(e^A, e^T) = -\log \frac{\exp(\mathrm{sim}(e^A, e^T)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(e^A, e^{T_i})/\tau)}, \tag{9}$$

$$\mathrm{L}_{MCL} = L(e_X^A, e_X^T) + L(e_Y^A, e_Y^T), \tag{10}$$

where e^A and e^T can be e_X^A and e_X^T , or e_Y^A and e_Y^T .

3.5. Framework's Overall Loss

At last, we integrate text and audio representations using the CM-BERT fusion method, capturing effective interactions between different modalities, to attain the final multimodal representation F_M. Together with the ground truth classification label y, our regression task loss L_{CE} is calculated using the weights W and bias b, which are learnable parameters of the model, to map the fused representation to the prediction space.

$$F_M = CM\text{-}BERT(\{e^A; e^T\}), \tag{11}$$

$$L_{CE} = CrossEntropy(W \cdot F_M + b, y). \tag{12}$$

Combined with the two contrastive learning losses above, the total loss for the training can be formulated as:

$$L_{total} = (1 - \alpha - \beta) \cdot L_{CE} + \alpha \cdot L_{MCL} + \beta \cdot L_{LCL}, \tag{13}$$

where α and β are weighted hyperparameters, which are used to balance the learning density of each module.

4. Experiment

4.1. Datasets

To evaluate the effectiveness of our proposed framework, we carry out experiments on two datasets: Flores200 and mms. Flores200 is used for domain classification, while mms is employed for sentiment classification. Flores200 (Costa-jussà et al., 2022) is a machine translation (MT) benchmark dataset connecting English with low-resource languages. mms (Augustyniak et al., 2024) is a multilingual corpus for sentiment analysis, encompassing 27 languages from 79 selected datasets, aimed at enhancing sentiment analysis in lower-resource languages with its extensive, culturally nuanced data.

We extracted parallel Chinese, Tibetan, and Burmese text from Flores200 and converted them to audio via TTS. Similarly, we processed Chinese data from mms, translating it into Burmese and Tibetan with GPT-4, and then used TTS for parallel audio generation. The detailed information of two synthetic datasets is shown in Table 1.

Table 1: Dataset Information

Dataset	Number of Texts	Average Length	Total Speech Length
Flores200	2008	23 words	20.22 hours
mms	6162	25 words	60.18 hours

4.2. Baseline

In this section, we outline the baseline methods utilized in our study. Given the challenges of language limitations in multilingual pre-trained models, we've chosen the following four methods for comparison.

mBERT-Unique: Similar to Lin et al. (2023), our research has been applied to BERT, where we fine-tune BERT to measure its performance on tasks of domain classification and sentiment classification. For each language, we conduct "Unique training" using BERT.

mBERT-Combined: To fully learn the information of different languages, we employ a "Combined Training" fine-tuning approach with BERT. Through this method, we obtain a single model, on which we conduct tests to measure the model's performance in tasks of topic classification and sentiment classification.

mBERT+PGD: We utilize adversarial training, notably Projected Gradient Descent (PGD) (Madry et al., 2018), to enhance the robustness of the BERT model, tackling language imbalance. Through parallel corpora and strategic perturbations, we obscure sensitive language attributes, leading to a more equitable treatment of languages in the model's learning process.

mBERT+FGM: Another adversarial training technique, FGM (Miyato et al., 2017), a single-step method, boosts BERT's robustness and mitigates language imbalance by obstructing precise identification of sensitive features, leading to more balanced language processing.

4.3. Experimental Setting

All experiments were conducted using the PyTorch[3] framework on an NVIDIA A100 GPU with 40GB of VRAM. Our framework is built upon a transformer architecture, incorporating pre-trained models such as "bert-base-multilingual-uncased"[4] and "wav2vec2-base-960h"[5]. To ensure a consistent hyperparameter space across runs, we fixed the random seed to 42. The training configuration includes a batch size of 8, a feature dimension of 768, a total of 5 epochs, and a maximum input length of 128. For the PGD method, we set epsilon to 0.01, alpha to 0.01, and the number of steps to 10. In contrast, for FGM, epsilon was set to 0.03. The learning rate for BERT model training was set to 3e-5, utilizing the Adam optimizer for adjustments. For the tasks of LCL and MCL, τ is tuned over the range {0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.3, 0.5}. The framework employs Hyperopt to optimize the weights of loss parameters α and β, as well as the temperature τ in the contrastive learning process.

3. https://pytorch.org/
4. https://huggingface.co/google-bert/bert-base-multilingual-uncased
5. https://huggingface.co/facebook/wav2vec2-base-960h

4.4. Language Balance Metrics

Despite extensive research on enhancing models for low-resource languages, there remains a performance trade-off, with a notable lack of metrics for overall efficacy in language balance. We propose a dual-dimension framework for multilingual text classification evaluation, targeting both the language balance across different languages (G) and the mean classification accuracy (M), to provide a comprehensive assessment of language performance equity.

The Gini coefficient (G), developed by Gini (1912) and widely used in economics, quantifies inequality in a distribution. It ranges from 0, representing perfect equality where all entities have equal shares, to 1, indicating maximum inequality where one entity holds all resources. Adapting this concept to multilingual classification systems, we have applied G, and its pseudocode is provided in Algorithm 1.

Algorithm 1 Calculate Gini Coefficient & Mean Value

Input: A list of classification scores S_1, S_2, S_3
Calculate total sum: $total = \sum(S_1, S_2, S_3)$
Compute mean value: $M = total/\text{count}(S_1, S_2, S_3)$
Normalize each value: $normalized\_values = [v/total$ for each v in $(S_1, S_2, S_3)]$
Sort the normalized values in ascending order
Construct the Lorenz curve with a starting point of 0
Use the trapezoidal rule to calculate the area under the Lorenz curve, denoted as B
Calculate $A = 0.5 - B$
Calculate the Gini coefficient $G = A/(A + B)$
Output: The Gini coefficient G and the mean value M of the input list

The mean classification accuracy (M) represents the mean value of classification accuracies for all languages in a given dataset. When combining the two metrics (G) and (M), we can comprehensively assess the performance of multi-language classification systems, not only focusing on average performance but also identifying and quantifying the uneven distribution of performance. This evaluation approach is particularly suitable for scenarios requiring balancing and optimizing support for multiple languages, ensuring that no language is unequally treated due to imbalanced system performance.

4.5. Results

Table 2: Experimental Results of Different Models on the Flores200 Dataset

Flores200	Accuracy ↑			G ↓	M ↑
	Chinese	Tibetan	Burmese		
mBERT-Unique	78.61%	50.25%	28.86%	0.2103	0.5257
mBERT-Combined	**87.24%**	58.38%	46.19%	0.1427	0.6394
mBERT+PGD	85.26%	32.68%	33.65%	0.2312	0.5053
mBERT+FGM	84.74%	38.05%	40.87%	0.1902	0.5455
OUR MODEL	84.67%	**59.21%**	**48.71%**	**0.1245**	**0.6420**

On the Flores200 dataset, mBERT-Combined achieved higher accuracies than mBERT-Unique across Chinese, Tibetan, and Burmese languages. Its accuracy on Chinese reached the maximum value among all baselines, reaching 87.24%. For Tibetan and Burmese, the accuracies were 58.38% and 46.19%, respectively, slightly lower than our model's 59.21% and 48.71%. Upon adversarial training using PGD and FGM on top of mBERT-Combined, the model's accuracies decreased across all three languages, the M value of the model dropped from 0.6394 to 0.5053 and 0.5455, and the G value increased from 0.1427 to 0.2312 and 0.1901, respectively. Our model achieved the minimum G value among all baselines, at 0.1245, and the highest M value, at 0.6420, which indicated that our model possessed superior language balance performance.

Table 3: Experimental Results of Different Models on the mms Dataset

mms	Accuracy ↑			G ↓	M ↑
	Chinese	Tibetan	Burmese		
mBERT-Unique	71.91%	42.17%	25.54%	0.2216	0.4656
mBERT-Combined	**80.25%**	45.15%	39.01%	0.1672	0.5480
mBERT+PGD	79.67%	44.76%	37.07%	0.1759	0.5383
mBERT+FGM	77.82%	44.23%	35.58%	0.1786	0.5254
OUR MODEL	79.70%	**48.42%**	**42.01%**	**0.1477**	**0.5671**

We conducted experiments on the second dataset, mms, using the same approach as the Flores200 dataset, where SimAdapter was used to initialize the adapter parameters. Similar to the results on the Flores200 dataset, mBERT-Combined achieved the best performance on the Chinese task, reaching 80.25. Our model obtained the best performance on Tibetan and Burmese tasks, with accuracies of 48.42% and 42.01%, respectively. Additionally, our model also demonstrated superior language balance performance on the mms dataset, with both G and M values outperforming other models, while maintaining the lowest G value when achieving the highest M value.

Table 4: Experimental Results of Different Adapters on the Flores200 Dataset

Flores200	Accuracy ↑			G ↓	M ↑
	Chinese	Tibetan	Burmese		
SimAdapter	85.01%	55.61%	46.65%	0.1366	0.6242
MetaAdapter	84.67%	59.21%	48.71%	**0.1245**	**0.6420**

To evaluate the impact of two adapter initialization techniques, SimAdapter and MetaAdapter, on multilingual classification, we utilized these methods to set the adapter parameters and then proceeded to train our model using the Flores200 dataset. The experimental outcomes revealed that, with MetaAdapter, the model's classification accuracy in Burmese and Tibetan significantly improved, outperforming SimAdapter by 2.06% and 3.60%, respectively. Conversely, in the Chinese language, MetaAdapter's performance slightly declined by 0.34% compared to SimAdapter. More importantly, the overall evaluation demonstrated that MetaAdapter enhanced the model's performance, as evidenced by improved G and M values of 0.1245 and 0.6420, respectively.

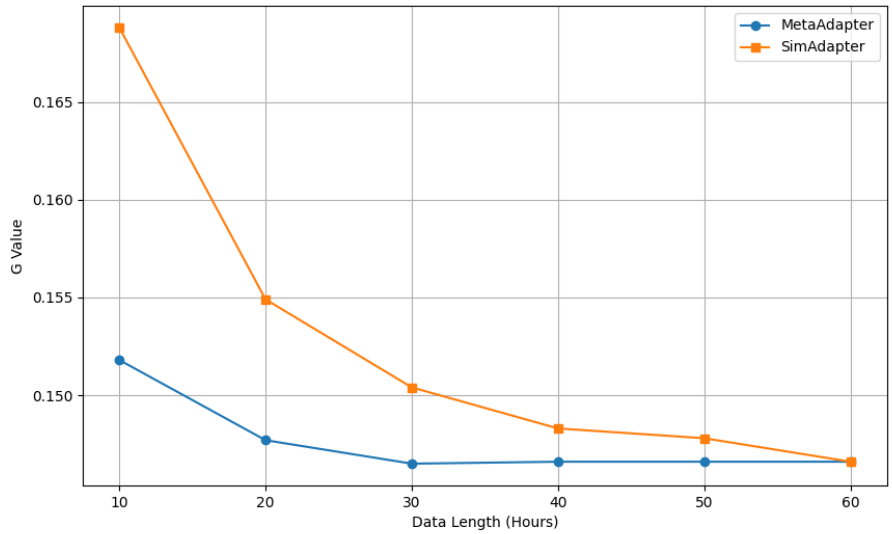

Figure 3: Performance of MetaAdapter and SimAdapter Across Different Data Lengths(Metric G)

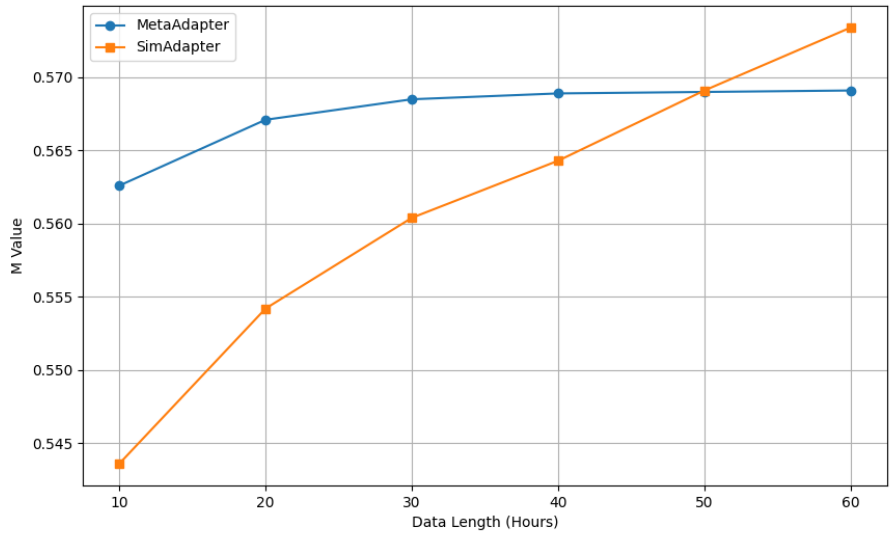

Figure 4: Performance of MetaAdapter and SimAdapter Across Different Data Lengths(Metric M)

To investigate the performance variation of the model under different adapter initialization methods across different dataset sizes, we conducted tests on the synthetic mms dataset. In total, we generated a dataset of 60.18 hours of speech data. Across these datasets, we conducted experiments using both MetaAdapter and SimAdapter. From Figures 3 and 4, it can be observed that when the dataset size is relatively small at 10 hours, the G value of SimAdapter, 0.1518, is lower than that of MetaAdapter, 0.1688, while the M value of SimAdapter, 0.5626, is higher than that of MetaAdapter, 0.5436. At the 50-hour mark, the M values of both methods are similar, with MetaAdapter at 0.5690 and SimAdapter at 0.5691. As the duration of the dataset increases to 60 hours, both G values continue to decrease, while both M values continue to increase. At the 60-hour mark, the G values of both methods are identical at 0.1466, with MetaAdapter achieving an M value of 0.5734, higher than SimAdapter's M value of 0.5691.

4.6. Ablation Study

Table 5: Ablation Results of Different Models on the Flores200 Dataset

Model	Accuracy ↑			G ↓	M ↑
	Chinese	Tibetan	Burmese		
OUR MODEL	84.67%	59.21%	48.71%	**0.1245**	**0.6420**
OUR MODEL w/o LCL	85.98%	58.61%	46.33%	0.1385	0.6364
OUR MODEL w/o MCL	86.44%	58.49%	46.70%	0.1383	0.6388
OUR MODEL w/o Adapter	69.26%	42.91%	33.17%	0.1655	0.4845

To validate the effectiveness of our proposed framework, we conducted a series of ablation experiments on the Flores200 dataset. As shown in Table 5, removing either the LCL, MCL, or adapter component from the framework results in performance degradation of the model. When either the MCL or LCL component is removed, the model's performance improves in the Chinese language but declines in Tibetan and Burmese languages, demonstrating a trade-off relationship. Particularly, when the adapter component is removed, the model's performance significantly decreases across all three languages, indicating the effectiveness of the adapter structure in MCML.

5. Conclusion

In this paper, we focus on the task of multi-language text classification and propose a novel framework to address the issue of language imbalance in multi-language pre-trained models. Our core idea is to leverage the correlations between speech among Sino-Tibetan languages to enhance the model's performance on low-resource languages, thereby mitigating language imbalance. However, our current work is limited by the lack of real-world data and does not encompass other languages within the Sino-Tibetan language family. In future work, we aim to explore the integration of more modalities, including text, speech, and images, and experiment with larger language models to verify the effectiveness of our method.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62376062), the Ministry of Education of Humanities and Social Science Project (No. 23YJAZH220, No. 24YJAZH244), the Philosophy and Social Sciences 14th Five-Year Plan Project of Guangdong Province (No. GD23CTS03), and the Guangdong Basic and Applied Basic Research Foundation of China (No. 2023A1515012718).

References

Lukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. Language-family adapters for low-resource multilingual neural machine translation. In Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jade Abbott, Jonathan Washington, Nathaniel Oco, Valentin Malykh, Varvara Logacheva, and Xiaobing Zhao, editors, *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.loresmt-1.5. URL https://aclanthology.org/2023.loresmt-1.5.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Koel Dutta Chowdhury, Mohammed Hasanuzzaman, and Qun Liu. Multimodal neural machine translation for low-resource language pairs using synthetic data. In Reza Haffari, Colin Cherry, George Foster, Shahram Khadivi, and Bahar Salehi, editors, *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 33–42, Melbourne, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-3405. URL https://aclanthology.org/W18-3405.

Maha Elbayad, Anna Sun, and Shruti Bhosale. Fixing MoE over-fitting on low-resource languages in multilingual machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14237–14253, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.897. URL https://aclanthology.org/2023.findings-acl.897.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095889.

Fredric C Gey. How similar are chinese and japanese for cross-language information retrieval? In *NTCIR*, 2005.

Corrado Gini. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]*. Tipogr. di P. Cuppini, 1912.

Qi Gou, Zehua Xia, and Wenzhe Du. Cross-lingual data augmentation for document-grounded dialog systems in low resource languages. In Smaranda Muresan, Vivian Chen, Kennington Casey, Vandyke David, Dethlefs Nina, Inoue Koji, Ekstedt Erik, and Ultes Stefan, editors, *Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 1–7, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.dialdoc-1.1. URL https://aclanthology.org/2023.dialdoc-1.1.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. Survey of low-resource machine translation. *Computational Linguistics*, 48(3): 673–732, September 2022. doi: 10.1162/coli_a_00446. URL https://aclanthology.org/2022.cl-3.6.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. Improving low-resource languages in pre-trained multilingual language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.822. URL https://aclanthology.org/2022.emnlp-main.822.

Wenxin Hou, Yidong Wang, Shengzhou Gao, and Takahiro Shinozaki. Meta-adapter: Efficient cross-lingual adaptation with meta-learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7028–7032. IEEE, 2021a.

Wenxin Hou, Han Zhu, Yidong Wang, Jindong Wang, Tao Qin, Renjun Xu, and Takahiro Shinozaki. Exploiting adapters for cross-lingual low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:317–329, 2021b.

Jaeseong Lee, Dohyeon Lee, and Seung-won Hwang. Script, language, and labels: overcoming three discrepancies for low-resource language specialization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13004–13013, 2023.

Nankai Lin, Junheng He, Zhenghang Tang, Dong Zhou, and Aimin Yang. Model and evaluation: Towards fairness in multilingual text classification. *arXiv preprint arXiv:2303.15697*, 2023.

Ronghao Lin and Haifeng Hu. Multimodal contrastive learning via uni-modal coding and cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 511–523, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.36.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=r1X3g2_xl.

Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. Improved text classification via contrastive adversarial training. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11130–11138. AAAI Press, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/21362.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. Contrastive learning for many-to-many multilingual neural machine translation. 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 244–258, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.21. URL https://aclanthology.org/2021.acl-long.21.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Xi'ao Su, Ran Wang, and Xinyu Dai. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–679, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.75. URL https://aclanthology.org/2022.acl-short.75.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.16. URL https://aclanthology.org/2020.repl4nlp-1.16.

Kaicheng Yang, Hua Xu, and Kai Gao. CM-BERT: cross-modal BERT for text-audio sentiment analysis. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 521–528, 2020. doi: 10.1145/3394171.3413690. URL https://doi.org/10.1145/3394171.3413690.

Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325, 2023.