

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 PEML: PROTOTYPE-ENHANCED META-LEARNING FOR MULTI-LINGUAL TEXT CLASSIFICATION

Anonymous authors

Paper under double-blind review

## ABSTRACT

Multi-lingual text classification is a challenging task in natural language processing, which not only faces language differences between multiple languages but also faces the challenge of scarce annotated data. This paper proposes a prototype-enhanced meta-learning (PEML) method to address the challenges in the multi-lingual text classification task. The PEML method consists of two steps: firstly, to enhance the model’s ability to understand multi-lingual samples, we design a multi-lingual label-fusion technique to better map labels from different languages into a unified semantical space; secondly, in response to the problem that class prototypes for support sets are difficult to apply to query sets in meta-learning, we use a query-enhanced technique to associates the prototype vectors of the support set with samples in the query set. After training with our method, the classification model can quickly update the class prototypes to the data distribution of the query set, thereby expanding the model’s multi-lingual classification ability from the support set to the unseen query set. Extensive experiments demonstrate that the proposed method significantly outperforms state-of-the-art methods in multi-lingual text classification tasks. The code and data will be released on GitHub.

## 1 INTRODUCTION

Multi-lingual text classification (MLTC) tasks have wide applications in practical scenarios such as public opinion monitoring and industry intelligence analysis (Brauwers & Frasincar, 2023; Minaee et al., 2022). Due to language and culture differences, it is not easy to annotate text data in different languages using the same classification criteria since annotators need to understand both language differences and annotation standards (Snell et al., 2017; Bao et al., 2020). Therefore, Current researches mainly focus on improving classification performance with only a few annotated samples (few-shot scenario) (Shliazhko et al., 2024; Liu et al., 2023).

Prototypical Networks (PN) are frequently used for solving few-shot text classification tasks (Snell et al., 2017; Bao et al., 2020; Luo et al., 2021). A prototype vector is considered to be representative of a class, which is constructed with samples in the support set (Lei et al., 2023). Query samples are judged to belong to a class based on the distance to the prototype vector (Liu et al., 2024). However, in MLTC tasks, intra-class language differences increase the computational difficulty of class prototypes, which can easily lead to insufficient representativeness of class prototypes and incorrect classification of query samples (Han et al., 2023; Li et al., 2024a).

Recently, several methods were proposed to address intra-class sample differences in PN research. For instance, MILADA (Han et al., 2021) introduced adversarial domain adaptive networks to reduce intra-class differences; ContrastNet (Chen et al., 2023) amplified inter-class differences through contrastive learning; Meta-SN (Han et al., 2023) improved the discriminative ability of the model by introducing external knowledge of class labels and a task construction strategy driven by hard samples. However, these methods usually focus on monolingual tasks (Li et al., 2024a; Zhang et al., 2022) and address intra-class sample differences by updating the class prototype from a category perspective (Li et al., 2024b; Lv, 2024), they lack a design for updating the prototype vectors in a multi-lingual scenario, i.e., updating category information and language information simultaneously.

To better address the challenges in MLTC tasks, this paper proposes a Prototype Enhanced Meta Learning (PEML) method, which introduces a multi-lingual and multi-class fusion and matching

strategy to efficiently learn the class prototype vectors for multi-lingual samples and unseen categories in meta-learning. Specifically, PEML first uses a multi-lingual label-fusion module to better map labels from different languages into a unified semantic space, which enhances the model’s ability to understand multi-lingual samples. Then, PEML uses a query-enhanced technique to better associate the prototype vectors of the support set with samples in the query set. In this way, the classification model can quickly update the class prototypes to the data distribution of the query set, thereby expanding the model’s multi-lingual classification ability to the unseen query set. The advantages of the PEML method are: 1) it does not require a manually constructed prompt, making it suitable for multi-lingual scenarios; 2) it needs only a few annotated data and does not need external knowledge, making it easily be applied to low-resources languages; 3) extensive experiments show that PEML achieves a new state-of-the-art performance under a multi-lingual few-shot scenario. The main contributions of this paper are as follows:

- We propose a prototype-enhanced meta-learning (PEML) method for MLTC tasks.
- We design a label-fusion module and a query-enhancement module in PEML. The label-fusion module maps samples from different languages of the same category to a unified semantic space, enhancing the model’s ability to understand samples in multiple languages. The query-enhancement module associates the class prototype vector with samples in the query set, alleviating the problem of insufficient representativeness of prototype vectors.
- Extensive experiments show that PEML achieves new state-of-the-art performance. We give a detailed analysis of the results. The code and data of this paper will be released on GitHub.

## 2 RELATED WORK

Due to the high cost of manual annotation and the difficulty in unifying multi-lingual labels, few-shot learning has become the mainstream research paradigm in MLTC, which can be categorized into transfer learning and meta learning.

### 2.1 TRANSFER LEARNING

Transfer learning usually transfers knowledge from the source domain to the target domain with multi-lingual pre-trained language models (PLMs). In recent years, the prompt-based method has gradually become the mainstream transfer learning paradigm. This method constructs prompt words to guide the model to stimulate existing knowledge with only a few examples (Liu et al., 2023).

Schick & Schütze (2021) introduced a prompt learning paradigm that requires manual template design. The KPT series method (Hu et al., 2022) guides PLMs to predict the category of query samples by constructing an external knowledge graph. Dong et al. (2023) transformed few-shot text classification into a correlation estimation problem and constructed a universal prompt template. Ji (2024) enhanced text representation and improved text classification performance by introducing entity relationship information from knowledge graphs. Liu & Yang (2024) proposed a knowledge-enhanced prompt learning method (SKPT) that optimizes prompt templates by introducing external knowledge (such as open triplets). Meng et al. (2025) proposed a multi-granularity feature extraction method that integrates semantic relevance information of labels, which improves the performance of the model on label confusion problems. Dementieva et al. (2025) proposed an unsupervised cross-linguistic knowledge transfer method that avoids manual data annotation and utilizes large-scale multi-lingual encoders and translation systems for text classification. Although these methods have achieved improvements, they often require the introduction of external knowledge or manual construction of prompt templates, which increases labor costs and has limited generalization ability when dealing with different types of problems (Gao et al., 2025; Hatefi et al., 2025). In addition, although large language models (LLMs) are currently a research hotspot in natural language processing, there are few works on multi-lingual text classification based on LLMs and prompt learning. **Compared with existing transfer learning methods, the few-shot learning method we proposed does not require the introduction of external knowledge or manually constructed prompt templates and only utilizes the characteristics of the data itself, which has better generalization ability.**

108  
109

## 2.2 META LEARNING

110  
111  
112  
113  
114

Meta learning, also known as “learning how to learn”, trains a model across many few-shot tasks, each consisting of a support set (for learning task-specific information) and a query set (for evaluating generalization). It can help learning algorithms achieve better generalization for unknown tasks, even in situations where training data is extremely limited or insufficient. The current meta learning methods can be mainly divided into three categories:

115  
116  
117

1) **Optimization-based methods** focus on “how to optimize”. For example, MAML (Finn et al., 2017) and Reptile (Nichol et al., 2018) focus on optimizing the gradient descent process.

118  
119  
120  
121  
122  
123

2) **Model-based methods** learn an implicit feature space and predict the labels of query samples in an end-to-end manner, but often lack interpretability. Compared with optimization-based methods, model-based methods are usually easier to train, but have weaker generalization ability for out-of-distribution tasks (Hospedales et al., 2020). For example, MANN (Santoro et al., 2016) and SNAIL (Mishra et al., 2018), which aim to build adaptive internal states in fixed modules to achieve fast and efficient parameter updates.

124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140

3) **Metric-based methods** learn an appropriate distance or similarity function to distinguish samples from different tasks. When faced with new tasks, this method calculates the distance between input samples and known samples, and then classifies them into the most similar or nearest category, such as Matching Network (Vinyals et al., 2016), Prototypical Network (Snell et al., 2017), and the Relation Network (Sung et al., 2018). Among them, the Prototype Network aggregates all annotated samples in each category into a category prototype, and then classifies them by measuring the distance between the query sample and each category prototype. Lei et al. (2023) proposed the TART method, which converts class prototypes into fixed reference points in a task-adaptive metric space to enhance the model’s generalization ability. The LAQDA proposed by Liu et al. (2024) utilizes label information and query samples to optimize class prototypes, to alleviate the problem of large intra-class differences and small inter-class differences between support set samples. Li et al. (2024a) proposed a prototype network optimization method that combines label propagation and attention mechanisms, aiming to improve the quality of prototype representation and metric flexibility. **The PEML method proposed in this paper belongs to the metric-based meta-learning method. Unlike existing methods, the PEML method utilizes a label-fusion and a query-enhancement technique to obtain and update the multi-lingual class prototype vectors, thereby better transferring the classification ability of the model to unseen multi-lingual categories.**

141  
142

## 3 THE PROPOSED PEML METHOD

143  
144

## 3.1 PROBLEM FORMULATION

145  
146  
147  
148  
149

Meta-learning provides an effective paradigm for few-shot text classification, typically adopting the  $N$ -way  $K$ -shot task setting. For each task, there are  $N$  classes, and each class has only  $K$  labeled samples for training. The data in meta-learning is divided into two parts: the source classes  $Y_{train}$  and the target classes  $Y_{test}$ , where  $Y_{train} \cap Y_{test} = \emptyset$ . In general, meta-learning contains two phases: meta-training and meta-testing.

150  
151  
152  
153  
154  
155  
156

**Meta-training:** During the meta-training phase, the model is trained on a set of meta-tasks constructed from the training set  $D_{train}$ , whose class labels are sampled from  $Y_{train}$ . Each meta-task consists of a support set and a query set. Specifically, for each task,  $N$  classes are randomly sampled from  $Y_{train}$ ,  $K$  labeled examples are sampled as the support set  $S_{train}$  and another  $M$  examples as the query set  $Q_{train}$  per class, denoted as  $S_{train} = \{(x_i, y_i)\}_{i=1}^{N \times K}$  and  $Q_{train} = \{(x_j, y_j)\}_{j=1}^{N \times M}$ . The model makes predictions about the query set  $Q_{train}$  based on the given support set  $S_{train}$ . Then the model updates the parameters by minimizing the loss with  $Q_{train}$ .

157  
158  
159  
160  
161

**Meta-testing:** During the meta-testing phase, the model is used to predict the labels of query samples in  $D_{test}$ . For each task,  $N$  novel classes will be sampled from  $Y_{test}$ , which is disjoint from  $Y_{train}$ . Then the support set  $S_{test}$  and the query set  $Q_{test}$  will be sampled from the  $N$  classes, like in meta-training. The query set  $Q_{test}$  is denoted as  $Q_{test} = \{x_j\}_{j=1}^{N \times M}$ , where the label of each example is unknown to the model. The performance of the model will be evaluated through the average classification accuracy on the query set  $Q_{test}$  across all the testing tasks.

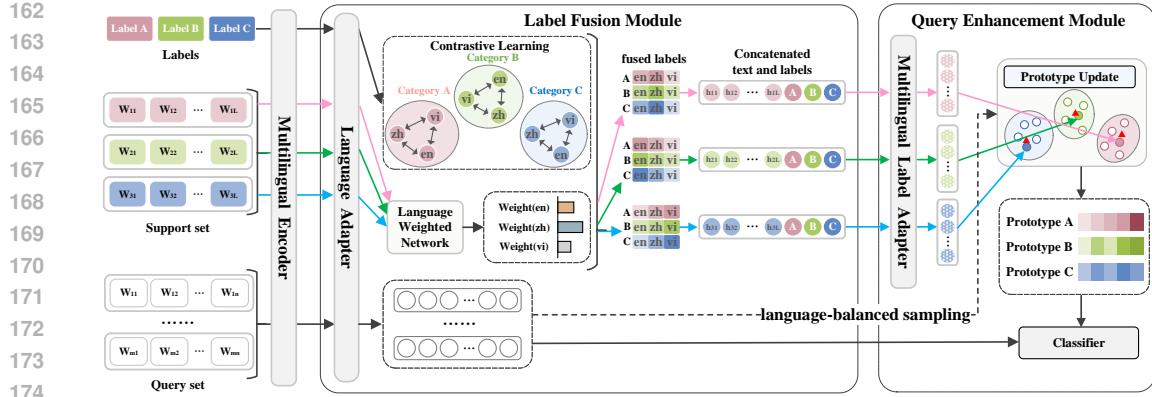


Figure 1: PEML on a 3-way 1-shot classification task. The dashed arrow is the query process.

### 3.2 FRAMEWORK

Figure 1 gives an overview of PEML. First, a *multi-lingual Encoder* gets the representations of the input sentences and label names. **In the label-fusion module**, a *Language Adapter* is applied to enhance language-specific features. Then, contrastive learning is employed to align the semantic spaces of the multiple languages under each label. Meanwhile, a Weighting Network computes language weights for each query sample across multiple languages and adaptively fuses the multi-lingual representations of candidate labels. **In the query-enhancement module**, a *multi-lingual Label Adapter* is used to map the fused representations (label and examples) to a class prototype space. During meta-training, the prototype of the support set is used to calculate losses. During meta-testing, the prototype is used to query samples from the query set  $Q_{test}$  as data augmentation to update itself, mitigating the problem of intra-class differences. Finally, query samples are inferred by a *Classifier* which calculates the distance between query samples and the updated prototype vector. Next, we will introduce the PEML method in detail.

### 3.3 LABEL FUSION MODULE

#### 3.3.1 INPUT TEXT ENCODING

Firstly, an input text is tokenized and encoded with a multi-lingual encoder (e.g., mBERT) to obtain its representation, i.e., the final-layer hidden state sequence  $\{h_1, h_2, \dots, h_L\} \in \mathbb{R}^{L \times d}$ , where  $L$  is the sequence length and  $d$  is the hidden size. After initial encoding, we design a Language Adapter (LA) module to further optimize the representations of different languages. Different languages will be automatically recognized and input into the corresponding sub-adapter. Each sub-adapter consists of a down-projection of the hidden vector  $h_i$  to a lower dimension  $r$  (where  $r = d/4$ ), followed by a non-linear transformation, and an up-projection back to the original dimension  $d$ . The adapted vector is then added to the original vector via a residual connection, followed by layer normalization (LayerNorm). This process is:  $\tilde{h}_i = \text{LA}_\Phi(h_i)$ ,  $i = \{1, 2, \dots, L\}$ . Here,  $\Phi$  denotes the trainable parameters of the adapter, and we maintain a separate parameter set for each language, i.e.,  $\Phi_{en}$ ,  $\Phi_{zh}$ , and  $\Phi_{vi}$  for English, Chinese, and Vietnamese (we use these three languages for a multi-lingual demonstration in this paper), respectively. This process transforms the original encoding results into a more language-specific representation, enabling the model to capture subtle semantic differences and language-specific features in the multi-lingual embedding space. After applying the adapter transformation to each hidden vector in the sequence, we perform Mean Pooling over the sequence length to obtain a sentence-level vector  $\mathbf{v}_{text} = \frac{1}{L} \sum_{i=1}^L \tilde{h}_i$ ,  $\tilde{h}_i \in \mathbb{R}^d$ , where  $\tilde{h}_i$  denotes the adapter-transformed hidden state at position  $i$ . This sentence vector preserves the contextual information learned by the multi-lingual encoder through multi-lingual pre-training, while also integrating language-specific features extracted by the adapter.

#### 3.3.2 INPUT LABEL ENCODING AND FUSION

In terms of multi-lingual label modeling, assume the current task involves  $N$  candidate category labels. For the  $j$ -th label, its multi-lingual representation includes Chinese, English, and Vietnamese

216 versions, denoted as  $l_j^{(\text{zh})}$ ,  $l_j^{(\text{en})}$ , and  $l_j^{(\text{vi})}$ , respectively. Each version is a sequence of words or  
 217 subword tokens, represented as  $l_j^{(\text{lang})} = [l_1, l_2, \dots]$ , where  $\text{lang} \in \{\text{zh}, \text{en}, \text{vi}\}$ . We first encode  
 218 the label names in each language independently using a multi-lingual encoder. The resulting hidden  
 219 states for each language are mean-pooled and then passed through a language-specific adapter,  
 220 yielding three language-specific label embeddings:  $\hat{l}_j^{(\text{lang})} = \text{LA}_{\Phi}^{(\text{lang})} \left( \frac{1}{m} \sum_{k=1}^m \text{Enc}(l_j^{(\text{lang})})_k \right)$ ,  
 221 where  $\text{Enc}(\cdot)$  denotes the multi-lingual encoder, and  $\hat{l}_j^{(\text{lang})} \in \mathbb{R}^d$  is the representation of the  $j$ -th  
 222 label in language  $\text{lang} \in \{\text{zh}, \text{en}, \text{vi}\}$  after processing with the corresponding Language Adapter.  
 223

224 To enable dynamic semantic fusion of multi-lingual label representations, we introduce a language-  
 225 weighted Network, which takes the pooled text representation  $\mathbf{v}_{\text{text}}$  as input and predicts an attention  
 226 distribution over the three language-specific label embeddings. The Query Network is implemented  
 227 as a two-layer multilayer perceptron (MLP), whose output is a 3-dimensional vector normalized  
 228 by a Softmax function to produce the attention weights  $(w_{\text{zh}}, w_{\text{en}}, w_{\text{vi}})$ . These weights are used  
 229 to compute a weighted combination of the multi-lingual label embeddings. The final fused multi-  
 230 lingual label representation for the  $j$ -th label is given by:  $\tilde{l}_j = w_{\text{zh}} \cdot \hat{l}_j^{(\text{zh})} + w_{\text{en}} \cdot \hat{l}_j^{(\text{en})} + w_{\text{vi}} \cdot \hat{l}_j^{(\text{vi})}$ .  
 231

232 To further enhance the semantic consistency among multi-lingual label representations, we incorpo-  
 233 rate a cross-lingual contrastive loss based on InfoNCE (Information Noise Contrastive Estimation)  
 234 (Gao et al., 2021). For each category label  $j$ , we obtain its embeddings under three different lan-  
 235 guages:  $\{\hat{l}_j^{(\text{zh})}, \hat{l}_j^{(\text{en})}, \hat{l}_j^{(\text{vi})}\}$ . These embeddings are first  $\ell_2$ -normalized and treated as positive pairs.  
 236 In contrast, embeddings of different category labels across languages are considered as negative  
 237 samples. Assume a mini-batch contains  $B$  labels in total. The contrastive loss  $\mathcal{L}_{\text{contra}}$  is defined as:  
 238

$$239 \quad - \frac{1}{3B} \sum_{j=1}^B \sum_{(a,b) \in \Omega} \log \frac{\exp \left( \text{sim} \left( \tilde{l}_j^{(a)}, \tilde{l}_j^{(b)} \right) / \tau \right)}{\sum_{k=1}^B \exp \left( \text{sim} \left( \tilde{l}_j^{(a)}, \tilde{l}_k^{(b)} \right) / \tau \right)}, \quad (1)$$

243 where  $\Omega = \{(\text{zh}, \text{en}), (\text{zh}, \text{vi}), (\text{en}, \text{vi})\}$ ,  $\text{sim}(\cdot)$  refers to cosine similarity, and  $\tau$  is a temperature  
 244 hyperparameter. This loss encourages the embeddings of the same label across different languages to  
 245 be semantically aligned, while pushing apart embeddings of different labels. As a result, it improves  
 246 the cross-lingual alignment of label representations in the shared semantic space.  
 247

248 We concatenate the word vector sequence of the input sentence  $[\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_L]$ , and the multi-  
 249 lingual label representations of all  $N$  candidate categories in the current task  $[\tilde{l}_1, \tilde{l}_2, \dots, \tilde{l}_N]$  to form  
 250 the overall input sequence to the multi-lingual label adapter:  $[\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_L, \tilde{l}_1, \tilde{l}_2, \dots, \tilde{l}_N]$ .  
 251

### 252 3.4 QUERY ENHANCEMENT MODULE

#### 253 3.4.1 MULTI-LINGUAL LABEL ADAPTER

254 To construct a task-adaptive metric space in which intra-class samples are more tightly clustered and  
 255 inter-class samples are better separated, we introduce a multi-lingual Label Adapter (MLA), which  
 256 enhances the conventional self-attention framework by explicitly incorporating multi-lingual label  
 257 semantics into sentence representation learning. The goal of this module is to enhance the model’s  
 258 discriminative capacity across languages by embedding task-specific label information directly into  
 259 the sentence encoding process.  
 260

261 To enable deep integration between the input sentence and multi-lingual label information, we in-  
 262 troduce a set of multi-head attention blocks. This module is parameterized by a set of trainable  
 263 meta-parameters  $\theta$ , and is defined as:  $\text{MLA}_{\theta}(Q, K, V) = \sigma(QK^T) \cdot V$ , where the pairwise dot-  
 264 product  $QK^T$  measures the similarity amongst features and is used for feature weighting com-  
 265 puted through an activation function  $\sigma$ . Intuitively, each feature of  $V$  will get more weight if the  
 266 dot-product between  $Q$  and  $K$  is larger. In  $\text{MLA}_{\theta}(Q, K, V)$ , following the self-attention mecha-  
 267 nism, we have  $Q = K = V$ . We input  $[\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_L, \tilde{l}_1, \tilde{l}_2, \dots, \tilde{l}_N]$  to the MLA. The output  
 268 of  $\text{<CLS>}$  position (denoted as  $\tilde{h}_0^*$ ) serves as the final representation vector  $v$  of the sentence:  
 269  $v = h_0^* = \text{MLA}_{\theta} \left( [\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_L, \tilde{l}_1, \tilde{l}_2, \dots, \tilde{l}_N] \right)$ .  
 270

270 By explicitly incorporating multi-lingual label embeddings into the sentence representation construction process, the model not only preserves the original linguistic context of the text but also captures the cross-lingual semantic features of all candidate categories. This mechanism enables the sentence representation to align more closely in the embedding space with the multi-lingual centroid representations of its corresponding labels, thereby significantly enhancing the model’s accuracy and robustness in category discrimination. This mechanism demonstrates superior generalization and discriminative capabilities, especially in typical few-shot scenarios where multi-lingual samples are imbalanced or label semantics are ambiguous.

278 **3.4.2 PROTOTYPE UPDATE**

280 In multi-lingual few-shot text classification, due to the limited number of samples in the support set 281  $S$  and the semantic variance within classes, the class prototypes constructed solely from the support 282 set often suffer from bias and fail to accurately represent the true class centers. In contrast, the 283 query set  $Q$  typically contains more unlabeled samples with a broader distribution. Therefore, we 284 incorporate information from the query set to update and enhance the representational capacity of 285 the class prototypes. For each support sample, we select the top  $R$  most similar samples from the 286 query set to assist in constructing more accurate class prototypes. This is based on the intuition 287 that richer sample information can generally lead to better prototype estimation. Meanwhile, to 288 mitigate semantic bias caused by language imbalance, we apply language-balanced sampling to the 289 top  $R$  candidate samples. Specifically, we allocate  $\lfloor R/3 \rfloor$  samples to each language. Any remaining 290 slots are filled by randomly selecting from the remaining candidate pool, ensuring that the selected 291 samples are as balanced across the three languages as possible.

292 For an  $N$ -way  $K$ -shot classification task, let the  $K$  support samples of novel class  $c$  be denoted as 293  $\{x_{c1}^S, \dots, x_{cK}^S\}$ , with their corresponding representations  $\{v_{c1}^S, \dots, v_{cK}^S\}$ . The query set contains  $M$  294 unlabeled samples  $\{x_1^Q, \dots, x_M^Q\}$ , whose representations are  $\{v_1^Q, \dots, v_M^Q\}$ . We treat each sample as 295 a random variable following a Gaussian distribution and use an Optimal Transport (OT) technique 296 that can help align data distributions between query samples and class prototypes from the support 297 set. Specifically, for each sample in the  $c$ -th class support set  $S_c$ , we first retrieve its  $R$  most similar 298 samples in the query set  $Q$  based on the OT distance:

$$299 \quad M_c^Q = \operatorname{argmin}_{c \in N} \mathcal{W}(Q, S_c) = \operatorname{argmin}_{c \in N} \min_{\mathbf{T} \in \Sigma(Q, x_{ci}^S)} \langle \mathbf{C}, \mathbf{T} \rangle = \{a_1, a_2, \dots, a_R\}, \quad (2)$$

300 where  $\mathbf{C}$  denotes the cost matrix and each element is computed by the Euclidean distance between a 301 query sample and a support sample:  $c(x_i^Q, x_j^S) = \|v_i^Q - v_j^S\|_2^2$ .  $\mathbf{T} \in \mathbf{R}_+^{K \times m}$  represents the optimal 302 transport plan matrix that satisfies the marginal constraints  $\{\mathbf{T} \cdot \mathbf{1}_m = Q, \mathbf{T} \cdot \mathbf{1}_K = S_c\}$ , indicating 303 the optimal pairing between support and query samples. The optimal transport matrix  $\mathbf{T}_c$  can be 304 efficiently solved using the Sinkhorn Algorithm (Cuturi, 2013).

305 We next adapt the augmented information  $M_c^Q$  from the query set  $Q$ , mapping to the task as follows:

$$309 \quad \hat{a}_i = \operatorname{argmin}_{a_i \in M_c^Q} \sum_{j=1}^K \mathbf{T}_c(i, j) \cdot c(a_i, v_{cj}^S), \quad i = \{1, \dots, R\}, \quad (3)$$

310 where  $\hat{a}_i$  denotes the projected representation of the  $i$ -th sample representation in  $M_c^Q$ , and  $\mathbf{T}_c(i, j)$  311 is the element at position  $(i, j)$  of the transport matrix  $\mathbf{T}_c$ . This projection maps the query samples to 312 the distribution space of the support set through the optimal transport matrix  $\mathbf{T}_c$  to reduce language 313 bias. Previous studies (Liu et al., 2024) have shown that when the cost function is the squared 314 Euclidean norm, this projection process is equivalent to a weighted average of  $S_c$  (Courty et al., 2017) 315 as follows:  $\hat{S}_c = \operatorname{diag}(\mathbf{T}_c \mathbf{1}_{n_c})^{-1} \mathbf{T}_c S_c$ , where  $\hat{S}_c$  denotes the enhanced support sample 316 representations for class  $c$ , and  $\operatorname{diag}$  is a diagonal matrix. Following previous work (Liu et al., 2024), we 317 obtain the adapted augment information  $\hat{S}_c$ , which is then combined with the original support sample 318 representations to compute the final prototype of the  $c$ -th class:  $P_c = \operatorname{mean}(\operatorname{union}(S_c, \hat{S}_c))$ . By 319 integrating auxiliary sample information from the query set, this enhanced prototype becomes more 320 robust and better aligned with the true class center, thereby improving classification performance in 321 multi-lingual few-shot scenarios.

324

325

Table 1: Experimental results under different few-shot settings.

Methods	2-shot			4-shot			8-shot		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
mBERT	49.31	39.15	41.10	72.83	70.46	70.60	77.21	75.31	75.37
PET	28.03	23.01	25.27	31.84	26.63	27.65	41.46	40.48	39.57
MetricPrompt	67.49	64.55	64.45	75.09	73.29	73.13	80.90	79.79	79.80
KPT	75.48	74.16	72.18	80.22	78.84	78.67	82.50	80.78	80.81
PBML	86.77	86.41	85.87	87.57	87.39	86.95	88.89	88.65	88.76
DML	11.84	14.40	10.17	29.62	23.14	22.58	52.70	43.69	44.27
LAQDA	88.35	86.12	87.75	89.62	89.35	89.29	91.51	90.39	90.21
EMPT	88.49	88.50	88.44	89.34	89.16	89.21	90.10	89.91	90.09
Sailor2-8B	76.88	75.34	76.10	78.73	76.95	77.83	79.35	77.86	78.60
PEML(ours)	<b>91.26</b>	<b>90.65</b>	<b>90.57</b>	<b>92.61</b>	<b>92.15</b>	<b>92.11</b>	<b>93.54</b>	<b>93.29</b>	<b>93.27</b>

## 339 3.5 TRAINING AND TESTING PHASES

340  
341 **Training Phase:** For each query sample  $x_i^Q$  with representation  $v_i^Q$ , the model calculates the Eu-  
342 clidean distance between  $v_i^Q$  and each class prototype. The resulting distances are passed through a  
343 softmax function to estimate the probability that the sample belongs to  $c$ -th class, as follows:  
344

345 
$$P(y_c|x_i^Q, \mathcal{P}) = \frac{\exp(-\|v_i^Q - P_c\|_2^2)}{\sum_{j=1}^N \exp(-\|v_i^Q - P_j\|_2^2)}, \quad (4)$$
346  
347

348 where  $P_c$  denotes the prototype of class  $c$ , and  $N$  is the number of classes in the task. To optimize  
349 the model parameters, we use the cross-entropy loss function to compute the prediction error across  
350 all query samples and all classes:  $\mathcal{L}_{ce} = \sum_{q=1}^n \sum_{c=1}^N y_{qc} \log P(y_c|x_i^Q, \mathcal{P})$ , where  $y_{qc} = 1$  if the  
351 query sample  $x_i^Q$  belongs to the  $c$ -th class; otherwise,  $y_{qc} = 0$ .  $n$  denotes the number of samples in  
352 the query set. We define the optimization objective of PEML as:  $\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{contra}$ , where  $\lambda$  is  
353 a hyperparameter that balances the classification loss and contrastive loss. By minimizing  $\mathcal{L}$ , all the  
354 trainable model parameters can be learned.  
355356 **Testing Phase:** For a given  $N$ -way  $K$ -shot task, we first generate the adapted representations for the  
357 query samples and combine them with the original support set representations to construct the final  
358 support set. The Prototypical network is then used to predict the class label for each query sample  
359  $x_i^Q$ , as given by:  $\tilde{y} = \arg \max_c P(y_c | x_i^Q, \mathcal{P})$ .  
360

## 361 4 EXPERIMENTAL RESULTS AND ANALYSIS

362 The experimental settings are introduced in Appendix A. We aim to answer the following questions  
363 with the experiments: 1) whether PEML achieves state-of-the-art performance in few-shot MLTC  
364 tasks (section 4.1); 2) how does each component contribute to the overall performance (Section  
365 4.2); 3) whether prototype enhancement leads to more compact intra-class distributions and better-  
366 separated inter-class representations (Section 4.3); 4) what is the difference when select different  
367 number of query samples (Section 4.4); 5) what is the difference with different  $\lambda$  (Appendix B).  
368

## 369 4.1 MAIN RESULTS

370 We conducted experiments on multi-lingual text classification tasks under different few-shot conditions.  
371 Table 1 shows the experimental results compared with the baseline models.  
372373 We can see that the proposed method PEML (Ours) outperforms the baseline models in different  
374 few-shot settings and evaluation metrics, demonstrating strong performance advantages. Specifically,  
375 under the 2-shot setting, PEML achieved an F1 score of 90.57%, an improvement of 2.82%  
376 and 2.13% compared to the strong baselines LAQDA (87.75%) and EMPT (88.44%), respectively.  
377

378  
 379 Table 2: The ablation study results. “PEML w/o LF” removes the Label Fusion module, “PEML w/o  
 380 LA” removes the Language Adapter, “PEML w/o CL” removes the Contrastive Learning process,  
 381 “PEML w/o QE” removes the query enhancement module, “PEML w/o MLA” removes the multi-  
 382 lingual Label Adapter, “PEML w/o PU” removes the Prototype Update process.

Methods	2-shot			4-shot			8-shot		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
PEML	<b>91.26</b>	<b>90.65</b>	<b>90.57</b>	<b>92.61</b>	<b>92.15</b>	<b>92.11</b>	<b>93.54</b>	<b>93.29</b>	<b>93.27</b>
PEML w/o LF	82.75	81.43	82.08	84.42	83.21	83.59	86.96	85.91	86.61
PEML w/o LA	86.67	85.26	85.87	88.03	87.74	87.62	90.85	89.43	90.23
PEML w/o CL	84.22	83.43	83.73	85.98	84.85	85.41	88.08	87.84	87.96
PEML w/o QE	85.54	83.24	84.36	87.26	87.16	87.21	90.06	89.82	89.96
PEML w/o MLA	<u>91.24</u>	<u>90.59</u>	<u>90.48</u>	<u>91.03</u>	<u>90.74</u>	<u>90.62</u>	<u>92.08</u>	<u>91.73</u>	<u>91.85</u>
PEML w/o PU	86.44	85.58	85.65	87.86	87.07	87.12	88.23	88.01	88.02

393 It also far surpasses traditional fine-tuning and prompt-based methods such as mBERT, PET, KPT,  
 394 etc. PEML also outperforms the state-of-the-art multi-lingual LLM, such as Sailor2. The perfor-  
 395 mance of Sailor2 is lower than methods based on smaller PLMs (e.g., PBML, LAQDA EMPT). This  
 396 result indicates that even the state-of-the-art multi-lingual LLM faces a challenge in multi-lingual  
 397 few-shot text classification tasks. In summary, the 2-shot results verify PEML’s ability to construct  
 398 more accurate category prototypes under multi-lingual and extremely few-shot conditions.

399 In the 4-shot and 8-shot settings, most baseline models showed steady performance improvements.  
 400 The DML method has the most significant improvement when given more labeled samples. Taking  
 401 the F1 metric for instance, DML’s 4-shot has increased by more than 2 times compared to 2-shot,  
 402 and 8-shot has increased by nearly 2 times compared to 4-shot. This indicates that the DML method  
 403 can learn multi-lingual classification patterns from an increasing number of examples. However,  
 404 there is still a significant gap between it and the best-performing models. Among all models, PEML  
 405 still performed the best. The results again verify PEML’s good capability for few-shot MLTC tasks.

## 4.2 ABLATION STUDY

408 To evaluate the impact of the two modules proposed in this paper on model performance, we con-  
 409 ducted ablation experiments under different few-shot settings, and the results are shown in Table  
 410 2. We can see that **when removing the Label Fusion (LF) modules**, the performance of PEML  
 411 drops significantly (the F1 metric drops 8.49/8.52/6.66 on the 2/4/8-shot setting, respectively). The  
 412 results show that the LF module plays an important role in helping the model understand multi-  
 413 lingual text. When the Language Adapter (LA) was removed, the F1 score decreased by 4.70,  
 414 4.49, and 3.04 percentage points under the three settings, respectively. The results indicate that LA  
 415 helps capture language-specific features and enhances cross-lingual representation. The Contrastive  
 416 Learning (CL) process also helps to map different languages into a unified semantic space, remov-  
 417 ing CL causes performance declines. **When removing the Query Enhancement (QE) modules**,  
 418 the F1 values dropped to 84.36/87.21/89.96 on the 2/4/8-shot setting, respectively. Among them,  
 419 the performance degradation was most significant under the 2-shot setting, reaching 6.21 percentage  
 420 points. This indicates that in extremely few-shot scenarios, the PE module can alleviate the problem  
 421 of sparse or insufficiently representative support set samples, improve the robustness and discrimi-  
 422 nation of class prototypes. When the multi-lingual Label Adapter (MLA) module was removed, the  
 423 F1 value decreased by 0.09/1.49/1.42 under the 2/4/8-shot setting, respectively. This shows that the  
 424 MLA module introduces semantic information of multi-lingual labels, which helps to build a more  
 425 task-consistent and semantically distinguishable representation space. When the Prototype-Update  
 426 (PU) module was removed, the F1 value decreased by 4.92/4.99/5.25 under the 2/4/8-shot setting,  
 427 respectively. This shows that the PU module helps to build more robust prototype vectors, which  
 428 increases the classification performance.

## 4.3 VISUALIZATION

430 To evaluate the ability of different models in building better class prototypes, we used t-SNE (Van der  
 431 Maaten & Hinton, 2008) to visualize the sample representations and class prototypes generated by  
 PBML, LAQDA, and our method. We randomly selected 3 categories with 150 query samples per

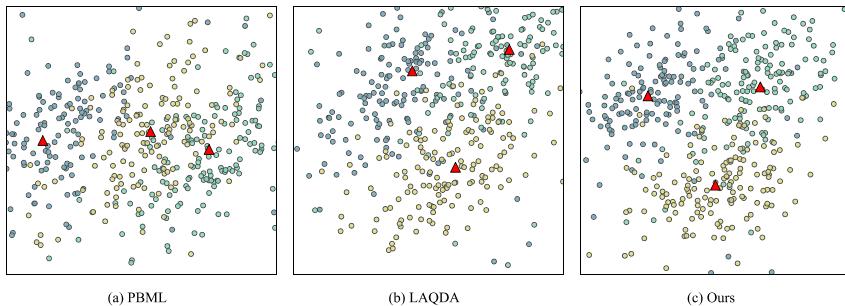


Figure 2: Visualization of sample text representations extracted from three new categories on the multi-lingual industry dataset. The triangle represents the calculated category prototype.

Table 3: Experimental results under different  $R$ .

Number of $R$	2-shot			4-shot			8-shot		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
R=3	89.99	89.13	88.99	92.19	91.59	91.52	92.86	91.79	91.73
R=6	<u>91.12</u>	90.43	90.34	92.36	92.01	91.93	<u>93.22</u>	<u>92.57</u>	<u>92.53</u>
R=12	<b>91.26</b>	<b>90.65</b>	<b>90.57</b>	<b>92.61</b>	<b>92.15</b>	<b>92.11</b>	<b>93.54</b>	<b>93.29</b>	<b>93.27</b>
R=24	<b>91.26</b>	<u>90.61</u>	90.54	<u>92.59</u>	<u>92.12</u>	<b>92.11</b>	93.19	92.46	92.40

category from the collected data. The results are shown in Figure 2 (a), (b), and (c). We can see that our PEML method makes samples have stronger compactness within the class, so as to alleviate the problem of insufficient representativeness of support sets in a multi-lingual scenario.

#### 4.4 EXPERIMENTS WITH DIFFERENT QUERY SAMPLES

To verify the impact of the number of query samples  $R$  in the prototype update module on the model performance, we conducted parameter sensitivity experiments in different few-shot settings (2/4/8-shot), setting  $R=\{3,6,12,24\}$ , respectively. It means that we introduce 1, 2, 4, and 8 auxiliary samples in each language. The experimental results as shown in Table 3. We can see that in 2/4/8-shot settings, the overall F1 value increases when  $R$  increases, indicating that introducing more auxiliary samples appropriately can help improve the representation quality of the prototype. Among them, under the 2-shot setting, the F1 value increased by about 1.55 percentage points (from 88.99% to 90.54%) from  $R=3$  to  $R=24$ ; Under the conditions of 4-shot, F1 increased by 0.59% from  $R=3$  to  $R=24$ . This indicates that under few-shot conditions, the gain brought by auxiliary samples is more significant, and additional language-balanced samples can alleviate the problem of class center shift, thereby improving classification performance. However, under the conditions of 8-shot, the performance gain tends to drop when  $R$  is larger than 12, indicating that too many auxiliary samples may bring redundant information or noise, affecting the query-enhancement effect. Hence, during the experiments, we select  $R=12$ .

**For more experiments.** Appendix B shows the experiments for deciding  $\lambda$  in the final loss  $\mathcal{L}$ , Appendix C shows the confusion matrix results of the test set. Appendix D shows experiments on other public MLTC datasets with 6 languages.

## 5 CONCLUSION

We propose a prototype-enhanced meta-learning (PEML) framework for few-shot MLTC tasks. The method introduces two key components: a label-fusion module that jointly encodes label names and text features to construct a task-adaptive metric space, and a query-enhancement module that refines class prototypes via optimal transport using query samples, addressing support set randomness and intra-class variation. Extensive experiments show that PEML achieves state-of-the-art performance. Further experiments are now conducting and will be released in the next version of the paper (including experiments on more languages and different kinds of encoders).

486 REFERENCES  
487

488 Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. Few-shot text classification with  
489 distributional signatures. In *8th International Conference on Learning Representations, ICLR*  
490 *2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=H1emfT4twB>.

491

492 Gianni Brauwers and Flavius Frasincar. A survey on aspect-based sentiment classification. *ACM*  
493 *Comput. Surv.*, 55(4):65:1–65:37, 2023. doi: 10.1145/3503044. URL <https://doi.org/10.1145/3503044>.

494

495 Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. Contrastnet: A contrastive learning frame-  
496 work for few-shot text classification. *CoRR*, abs/2305.09269, 2023. doi: 10.48550/ARXIV.2305.  
497 09269. URL <https://doi.org/10.48550/arXiv.2305.09269>.

498

499 Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for do-  
500 main adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1853–1865, 2017. doi: 10.1109/  
501 TPAMI.2016.2615921. URL <https://doi.org/10.1109/TPAMI.2016.2615921>.

502

503 Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C.  
504 Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural*  
505 *Information Processing Systems 26: 27th Annual Conference on Neural Information Processing*  
506 *Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United*  
507 *States, pp. 2292–2300, 2013*. URL <https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html>.

508

509 Daryna Dementieva, Valeria Khylenko, and Georg Groh. Cross-lingual text classification trans-  
510 fer: The case of ukrainian. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend  
511 Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st In-  
512 ternational Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, Jan-  
513 uary 19-24, 2025*, pp. 1451–1464. Association for Computational Linguistics, 2025. URL  
<https://aclanthology.org/2025.coling-main.97/>.

514

515 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep  
516 bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and  
517 Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of*  
518 *the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*  
519 *2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–  
520 4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL  
<https://doi.org/10.18653/v1/n19-1423>.

521

522 Hongyuan Dong, Weinan Zhang, and Wanxiang Che. Metricprompt: Prompting model as a  
523 relevance metric for few-shot text classification. In Ambuj K. Singh, Yizhou Sun, Leman  
524 Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (eds.),  
525 *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Min-  
526 ing, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pp. 426–436. ACM, 2023. doi:  
527 10.1145/3580305.3599430. URL <https://doi.org/10.1145/3580305.3599430>.

528

529 Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation  
530 of deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th Interna-  
531 tional Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*,  
532 volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 2017. URL  
<http://proceedings.mlr.press/v70/finn17a.html>.

533

534 Jia Gao, Shuangquan Lyu, Guiran Liu, Binrong Zhu, Hongye Zheng, and Xiaoxuan Liao. A hybrid  
535 model for few-shot text classification using transfer and meta-learning. *CoRR*, abs/2502.09086,  
536 2025. doi: 10.48550/ARXIV.2502.09086. URL <https://doi.org/10.48550/arXiv.2502.09086>.

537

538 Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sen-  
539 tence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-  
540 tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Lan-  
541 guage Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11*

540 November, 2021, pp. 6894–6910. Association for Computational Linguistics, 2021. doi:  
 541 10.18653/V1/2021.EMNLP-MAIN.552. URL <https://doi.org/10.18653/v1/2021.emnlp-main.552>.

542

543 Chengcheng Han, Zeqiu Fan, Dongxiang Zhang, Minghui Qiu, Ming Gao, and Aoying Zhou. Meta-  
 544 learning adversarial domain adaptation network for few-shot text classification. In Chengqing  
 545 Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Compu-  
 546 tational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP  
 547 2021 of *Findings of ACL*, pp. 1664–1673. Association for Computational Linguistics, 2021.  
 548 doi: 10.18653/V1/2021.FINDINGS-ACL.145. URL <https://doi.org/10.18653/v1/2021.findings-acl.145>.

549

550 Chengcheng Han, Yuhe Wang, Yingnan Fu, Xiang Li, Minghui Qiu, Ming Gao, and Aoying Zhou.  
 551 Meta-learning siamese network for few-shot text classification. In Xin Wang, Maria Luisa Sapino,  
 552 Wook-Shin Han, Amr El Abbadi, Gill Dobbie, Zhiyong Feng, Yingxiao Shao, and Hongzhi Yin  
 553 (eds.), *Database Systems for Advanced Applications - 28th International Conference, DASFAA  
 554 2023, Tianjin, China, April 17-20, 2023, Proceedings, Part III*, volume 13945 of *Lecture Notes  
 555 in Computer Science*, pp. 737–752. Springer, 2023. doi: 10.1007/978-3-031-30675-4\\_54. URL  
 556 [https://doi.org/10.1007/978-3-031-30675-4\\_54](https://doi.org/10.1007/978-3-031-30675-4_54).

557

558 Arezoo Hatefi, Xuan-Son Vu, Monowar H. Bhuyan, and Frank Drewes. The efficiency of pre-  
 559 training with objective masking in pseudo labeling for semi-supervised text classification. *CoRR*,  
 560 abs/2505.06624, 2025. doi: 10.48550/ARXIV.2505.06624. URL <https://doi.org/10.48550/arXiv.2505.06624>.

561

562 Timothy M. Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. Meta-learning in  
 563 neural networks: A survey. *CoRR*, abs/2004.05439, 2020. URL <https://arxiv.org/abs/2004.05439>.

564

565 Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and  
 566 Maosong Sun. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer  
 567 for text classification. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.),  
 568 *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume  
 569 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 2225–2240. Association  
 570 for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.158. URL <https://doi.org/10.18653/v1/2022.acl-long.158>.

571

572 Xue Ji. A cultural industry text classification method based on knowledge graph information con-  
 573 straints and knowledge fusion. *Int. J. Web Eng. Technol.*, 19(2):127–147, 2024. doi: 10.1504/IJWET.2024.139844. URL <https://doi.org/10.1504/IJWET.2024.139844>.

574

575 Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, and Chang-Tien Lu. TART: improved  
 576 few-shot text classification using task-adaptive reference transformation. In Anna Rogers,  
 577 Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meet-  
 578 ing of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023,  
 579 Toronto, Canada, July 9-14, 2023*, pp. 11014–11026. Association for Computational Linguis-  
 580 tics, 2023. doi: 10.18653/V1/2023.ACL-LONG.617. URL <https://doi.org/10.18653/v1/2023.acl-long.617>.

581

582 Haorui Li, Jie Shao, Xiangqiang Zeng, and Hui Xu. Improving meta-learning for few-shot text  
 583 classification via label propagation. In Albert Bifet, Jesse Davis, Tomas Krilavicius, Meelis  
 584 Kull, Eirini Ntoutsi, and Indre Zliobaite (eds.), *Machine Learning and Knowledge Discovery  
 585 in Databases. Research Track - European Conference, ECML PKDD 2024, Vilnius, Lithuania,  
 586 September 9-13, 2024, Proceedings, Part V*, volume 14945 of *Lecture Notes in Com-  
 587 puter Science*, pp. 389–405. Springer, 2024a. doi: 10.1007/978-3-031-70362-1\\_23. URL  
 588 [https://doi.org/10.1007/978-3-031-70362-1\\_23](https://doi.org/10.1007/978-3-031-70362-1_23).

589

590 Shujie Li, Guanghu Yuan, Min Yang, Ying Shen, Chengming Li, Ruifeng Xu, and Xiaoyan Zhao.  
 591 Improving semi-supervised text classification with dual meta-learning. *ACM Trans. Inf. Syst.*,  
 592 42(4):109:1–109:28, 2024b. doi: 10.1145/3648612. URL <https://doi.org/10.1145/3648612>.

594 Jinshuo Liu and Lu Yang. Knowledge-enhanced prompt learning for few-shot text classification.  
 595 *Big Data Cogn. Comput.*, 8(4):43, 2024. doi: 10.3390/BDCC8040043. URL <https://doi.org/10.3390/bdcc8040043>.

596

597 Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig.  
 598 Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language  
 599 processing. *ACM Comput. Surv.*, 55(9):195:1–195:35, 2023. doi: 10.1145/3560815. URL  
 600 <https://doi.org/10.1145/3560815>.

601

602 Xinyue Liu, Yunlong Gao, Linlin Zong, and Bo Xu. Improve meta-learning for few-shot text classifi-  
 603 cation with all you can acquire from the tasks. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung  
 604 Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami,*  
 605 *Florida, USA, November 12-16, 2024*, pp. 223–235. Association for Computational Linguistics,  
 606 2024. URL <https://aclanthology.org/2024.findings-emnlp.12>.

607

608 Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International*  
 609 *Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.  
 610 OpenReview.net, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.

611

612 Qiaoyang Luo, Lingqiao Liu, Yuhao Lin, and Wei Zhang. Don't miss the labels: Label-semantic  
 613 augmented meta-learner for few-shot text classification. In Chengqing Zong, Fei Xia, Wen-  
 614 jie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguis-  
 615 tics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Find-  
 616 ings of ACL*, pp. 2773–2782. Association for Computational Linguistics, 2021. doi: 10.  
 617 18653/V1/2021.FINDINGS-ACL.245. URL [https://doi.org/10.18653/v1/2021.  
 618 findings-acl.245](https://doi.org/10.18653/v1/2021.findings-acl.245).

619

620 Xiaobao Lv. Few-shot text classification with an efficient prompt tuning method in meta-learning  
 621 framework. *Int. J. Pattern Recognit. Artif. Intell.*, 38(3):2451006:1–2451006:18, 2024. doi: 10.  
 622 1142/S0218001424510066. URL <https://doi.org/10.1142/S0218001424510066>.

623

624 Chunyun Meng, Yuki Todo, Cheng Tang, Li Luan, and Zheng Tang. MFLSCI: multi-granularity  
 625 fusion and label semantic correlation information for multi-label legal text classification. *Eng.  
 626 Appl. Artif. Intell.*, 139:109604, 2025. doi: 10.1016/J.ENGAPPAL.2024.109604. URL <https://doi.org/10.1016/j.engappai.2024.109604>.

627

628 Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jian-  
 629 feng Gao. Deep learning-based text classification: A comprehensive review. *ACM Comput.  
 630 Surv.*, 54(3):62:1–62:40, 2022. doi: 10.1145/3439726. URL <https://doi.org/10.1145/3439726>.

631

632 Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-  
 633 learner. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver,  
 634 BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.  
 635 URL <https://openreview.net/forum?id=B1DmUzWAW>.

636

637 Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR*,  
 638 abs/1803.02999, 2018. URL <http://arxiv.org/abs/1803.02999>.

639

640 Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan,  
 641 Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas  
 642 Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy,  
 643 Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style,  
 644 high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina  
 645 Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in  
 646 Neural Information Processing Systems 32: Annual Conference on Neural Information Pro-  
 647 cessing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp.  
 648 8024–8035, 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html>.

649

650 Adam Santoro, Sergey Bartunov, Matthew M. Botvinick, Daan Wierstra, and Timothy P. Lillicrap.  
 651 Meta-learning with memory-augmented neural networks. In Maria-Florina Balcan and Kilian Q.

648 Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1842–1850. JMLR.org, 2016. URL <http://proceedings.mlr.press/v48/santoro16.html>.  
649  
650  
651  
652

653 Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and  
654 natural language inference. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Pro-  
655 ceedings of the 16th Conference of the European Chapter of the Association for Computational  
656 Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 255–269. Association  
657 for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EACL-MAIN.20. URL  
658 <https://doi.org/10.18653/v1/2021.eacl-main.20>.  
659

660 Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and  
661 Tatiana Shavrina. mgpt: Few-shot learners go multilingual. *Trans. Assoc. Comput. Linguistics*, 12:  
662 58–79, 2024. doi: 10.1162/TACL\_A\_00633. URL [https://doi.org/10.1162/tacl\\_a\\_00633](https://doi.org/10.1162/tacl_a_00633).  
663

664 Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Is-  
665 abelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vish-  
666 wanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30:  
667 Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long  
668 Beach, CA, USA*, pp. 4077–4087, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html>.  
669

670 Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales.  
671 Learning to compare: Relation network for few-shot learning. In *2018 IEEE Conference  
672 on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June  
673 18-22, 2018*, pp. 1199–1208. Computer Vision Foundation / IEEE Computer Society, 2018.  
674 doi: 10.1109/CVPR.2018.00131. URL [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Sung\\_Learning\\_to\\_Compare\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Sung_Learning_to_Compare_CVPR_2018_paper.html).  
675

676 Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine  
677 learning research*, 9(11), 2008.  
678

679

680 Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Match-  
681 ing networks for one shot learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg,  
682 Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems  
683 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016,  
684 Barcelona, Spain*, pp. 3630–3638, 2016. URL <https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html>.  
685

686

687 Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi,  
688 Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick  
689 von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger,  
690 Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural  
691 language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference  
692 on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020  
693 - Demos, Online, November 16-20, 2020*, pp. 38–45. Association for Computational Linguistics,  
694 2020. doi: 10.18653/V1/2020.EMNLP-DEMONS.6. URL <https://doi.org/10.18653/v1/2020.emnlp-demos.6>.  
695

696

697 Haoxing Zhang, Xiaofeng Zhang, Haibo Huang, and Lei Yu. Prompt-based meta-learning for  
698 few-shot text classification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.),  
699 *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,  
700 EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 1342–1357. Association  
701 for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.87. URL  
<https://doi.org/10.18653/v1/2022.emnlp-main.87>.

702

703

Table 4: Statistics of the collected dataset.

704

705

Category	A	B	C	D	E	F	G	H	I
<b>Chinese</b>	20,243	3,880	8,252	18,854	7,425	9,958	9,995	15,590	19,584
<b>English</b>	14,076	1,574	2,773	8,061	3,381	4,716	5,163	9,160	7,751
<b>Vietnamese</b>	3,320	253	210	579	2,915	737	2,930	2,966	2,848
<b>Total</b>	37,639	5,707	11,235	27,494	13,721	15,411	18,088	27,716	30,183

706

707

## A EXPERIMENTAL SETTINGS

708

### A.1 DATASET

709

Due to the lack of multi-lingual text classification datasets under a unified classification standard in real-world scenarios, we collected a total of 187,194 news articles from publicly available news websites, including 113,781 in Chinese, 56,655 in English, and 16,758 in Vietnamese. The data statistics are shown in Table 4. The average word length of Chinese/English/Vietnamese news is around 245/326/513, respectively. The collected data covers multiple categories. We use the classification tags provided by the website and classify the data into 9 categories based on the same classification criteria. The Category index A is “Agriculture, forestry, animal husbandry and fishery”, B is “Mining”, C is “Manufacturing”, D is “Transportation, storage and postal services”, E is “Finance”, F is “Real estate”, G is “Education”, H is “Health and social work”, I is “Culture, sports and entertainment”. To verify the ability of few sample learning, data from 9 categories were divided into training category, validation category, and testing category in a 4:2:3 ratio in the experiment. We also tested with a 3:3:3 division, and the results are similar to those with a 4:2:3 division.

710

711

### A.2 EVALUATION METRICS

712

713

Following previous work (Liu et al., 2024; Lv, 2024), we use Precision, Recall, and F1 score to measure the performance of text classification. The above indicators are calculated based on weighted averages in subsequent sections to ensure that the results accurately reflect the model’s overall performance in each category.

714

### A.3 TRAINING DETAILS

715

The experimental implementation is based on the PyTorch (Paszke et al., 2019) and Huggingface (Wolf et al., 2020) frameworks. To ensure fairness in comparison, all methods used BERT-base-multi-lingual-uncased (Devlin et al., 2019) as the backbone model, and all experiments were run on an NVIDIA RTX 4090 GPU. The model parameters were optimized using the AdamW optimizer (Loshchilov & Hutter, 2019), with a learning rate set to 1e-6. Early stopping strategy is executed when the performance of the validation set does not improve for 20 consecutive epochs. We conducted few-shot classification experiments under the 2-way 2-shot, 2-way 4-shot, and 2-way 8-shot settings. Following previous work, 100, 100, and 1000 tasks are randomly sampled for training, validation, and testing in each round for each method, respectively. For each category, 30 query samples in each task are randomly selected. We also test with different numbers of query samples, and the results still verify the effectiveness of PEML. Due to the page limit, we only show the 30 query sample experiments.

746

747

### A.4 BASELINE MODELS

748

This paper chooses the following method as the baseline model (including both transfer learning methods and meta learning methods):

749

**mBERT** (Devlin et al., 2019) uses the output of the <CLS> position as text representations, and maps them to specific categories for classification through linear layers.

750

751

**PET** (Schick & Schütze, 2021) uses manual templates to construct prompts by adding prefixes or suffixes to the input text and masking certain markers, converting the given tasks into fill-in-the-blank phrases.

756 Please classify the given sentence into one of the 9 categories (A. "Agriculture, forestry, animal husbandry and fishery",  
 757 B. "Mining", C. "Manufacturing", D. "Transportation, storage and postal services", E. "Finance", F. "Real estate", G.  
 758 "Education", H. "Health and social work", I. "Culture, sports and entertainment").  
 759 **Examples:** Sentence: "国际教育招生规模进一步扩大。", Category: G. Sentence: "Tin ore production increased by  
 760 10%.", Category: B. **Question:** Sentence: "Số lượng du học sinh hai nước qua biên giới tăng hàng năm", Category: ?

761

762

Figure 3: The 2-shot prompt example for Sailor2 (Including three languages).

763

764

765

766 Table 5: Experimental results with different  $\lambda$ . Best results are highlighted in bold.

Value of $\lambda$	2-shot			4-shot			8-shot		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
$\lambda=0.05$	91.05	<u>90.58</u>	90.42	92.32	91.87	91.83	92.88	<u>92.96</u>	<u>92.92</u>
$\lambda=0.10$	<b>91.26</b>	<b>90.65</b>	<b>90.57</b>	<b>92.61</b>	<b>92.15</b>	<b>92.11</b>	<b>93.54</b>	<b>93.29</b>	<b>93.27</b>
$\lambda=0.15$	<u>91.19</u>	90.52	<u>90.51</u>	<u>92.42</u>	<u>92.02</u>	<u>92.05</u>	<u>92.98</u>	92.76	92.81
$\lambda=0.20$	91.16	90.51	90.50	92.21	91.93	92.01	92.44	92.33	92.32
$\lambda=0.25$	91.11	90.43	90.45	92.12	91.89	91.97	92.41	92.30	92.29

775

776

777

778 **MetricPrompt** (Dong et al., 2023) transforms few-shot text classification into a correlation esti-  
 779 mation task, using a prompt model as a correlation measure and supervised training of the model  
 780 through cross-entropy loss.

781

782 **KPT** (Hu et al., 2022) expands the search space of label words by introducing external knowledge  
 783 and uses a PLM to refine the expanded label word space before prediction, aiming to improve and  
 784 stabilize the performance of prompt tuning.

785

786 **PBML** (Zhang et al., 2022) combines a prompt mechanism and meta learning framework for few-  
 787 shot text classification. with few samples by assigning label words and template learning.

788

789 **DML** (Li et al., 2024b) introduces a dual meta learning mechanism to optimize the teacher and stu-  
 790 dent models through pseudo-label correction and feedback supervision, improving the performance  
 791 of semi-supervised text classification.

792

793 **LAQDA** (Liu et al., 2024) uses label information and query samples to optimize class prototypes,  
 794 to alleviate the problem of large intra-class differences and small inter-class differences between  
 795 support set samples.

796

797 **EMPT** (Lv, 2024) introduces an efficient prompt optimization method in the meta learning frame-  
 798 work, which normalizes label and sample information and uses regression to solve closed-form  
 799 solutions, improving inference speed and classification accuracy, and enhancing stability under a  
 800 small amount of meta-training data.

801

802

803

## B EXPERIMENTS WITH DIFFERENT $\lambda$

804

805

806 We set  $\lambda=\{0.05, 0.1, 0.15, 0.2\}$  in final loss  $\mathcal{L}$  to conduct parameter sensitivity experiments with  
 807 2/4/8-shot settings. The results are shown in Table 5. We can see that the F1 value is best when  
 808  $\lambda=0.1$  in all settings. When  $\lambda$  became larger or smaller, the model's performance all dropped.  
 809 Hence, during the experiments, we select  $\lambda=0.1$ . However, the overall trend of change is relatively  
 810 small, indicating that contrastive learning can steadily improve the performance of the model.

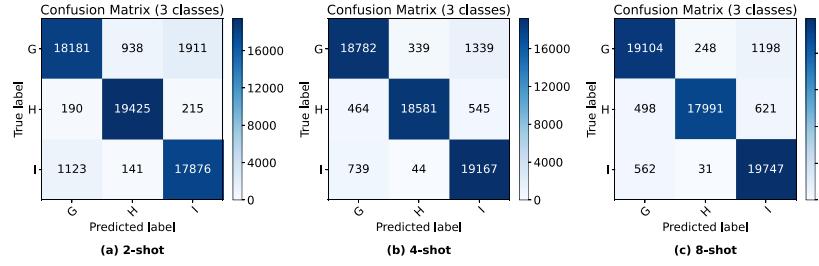


Figure 4: Confusion matrix results of test set.

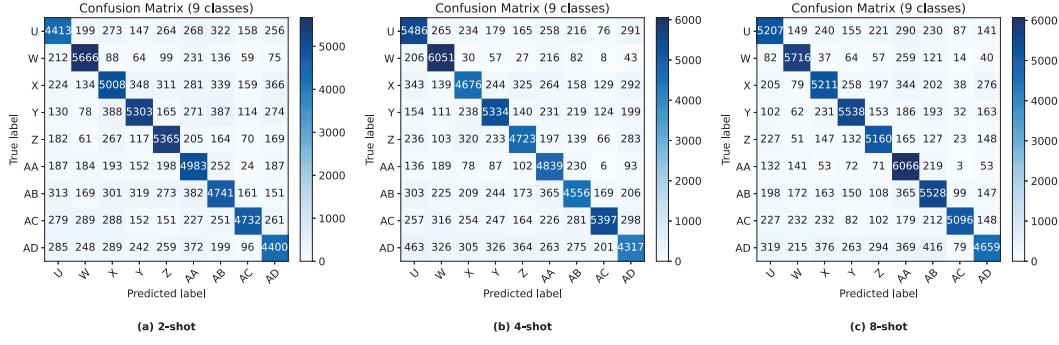


Figure 5: Confusion matrix results on the test test of the AMR data.

### C CONFUSION MATRIX RESULTS OF TEST SET

From the confusion matrix shown in Figure 4, it can be observed that as the number of shots increases from 2 to 4 and 8, the overall accuracy gradually improves, multilingual cross-class misclassification continuously decreases, class boundaries become clearer, and intra-class consistency increases. This trend validates the effectiveness of the proposed PEML method: by leveraging multilingual label fusion, labels from different languages are aligned into a unified semantic space, significantly alleviating cross-lingual semantic bias and inter-class confusion. Furthermore, through query enhancement, the model can rapidly adapt class prototypes to the query distribution during inference, thereby overcoming the representation bottleneck of multilingual models in low-resource scenarios.

Table 6: Experimental results with public multilingual datasets.

Methods	2-shot			4-shot			8-shot		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
LAQDA	74.59	73.45	73.07	78.02	77.17	76.95	81.21	80.40	80.22
PEML	<b>76.53</b>	<b>75.34</b>	<b>75.03</b>	<b>79.18</b>	<b>78.29</b>	<b>77.24</b>	<b>83.62</b>	<b>82.28</b>	<b>82.06</b>

### D EXPERIMENTAL RESULTS ON PUBLIC MULTILINGUAL DATASETS

To further validate the effectiveness of our proposed method, we conducted experimental evaluations on a public multilingual dataset. Specifically, we used the *multilingual-amazon-reviews-6-languages* dataset (AMR: Amazon Multilingual Reviews Dataset, available at: <https://huggingface.co/datasets/srvmishra832/multilingual-amazon-reviews-6-languages>), which covers 6 languages and over 30 categories. We split this dataset into 15 categories for training, 5 for validation, and 9 for testing. Meanwhile, we compared our method with state-of-the-art baseline model LAQDA, and the experimental results along with confusion matrices are presented below.

It can be seen from Table 6 that compared with the baseline model LAQDA, PEML has significant advantages in terms of precision, recall rate and F1 value under 2-shot, 4-shot and 8-shot Settings.

864 Among them, the F1 value at 2/8-shot outcome LAQDA around 2%. This result demonstrates that  
865 PEML shows stronger generalization ability in multilingual classification tasks. From the confu-  
866 sion matrix in Figure 5, it can be seen that the category boundaries of PEML are clear, indicating  
867 that it can still maintain good category discriminability in more languages and few-shot scenarios,  
868 fully verifying the core design effectiveness of the method in addressing the challenges of language  
869 differences and few-shot in MLTC.

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917