Multi-Task Classification Model for Muitilingual Patents

Yongxin Peng[1,†], Xinyu Tong[1,†], Yueyun Zhang[1] and Yonghe Lu[2,*]

[1]School of Information Management, Sun Yat-sen University, Waihuan east street No.132, Guangzhou, 510006, China
[2]School of Artificial Intelligence, Sun Yat-sen University, Tangjiawan, Zhuhai, 519082, China

Abstract

The existing patent text classification models perform well in many commonly-used languages but show suboptimal performance with corpora from the Association of Southeast Asian Nations (ASEAN). This paper proposes a multi-task modeling approach to train a multilingual patent classification model, which applies to patent texts of multiple languages, including commonly-used and low-resource languages. The model learns joint text embeddings by training a multilingual patent classifier and a cross-lingual text-pair similarity discriminator using monolingual classification labeled data and aligned bilingual corpora at the sentence or paragraph level. The method shows competitive performance, achieving up to 5% improvement in accuracy over mBERT on cross-lingual patent classification tasks in Thai and English.

Keywords

Patent classification, Multilingual classification, Multi-task learning,

1. Introduction

Automatic classification of patent documents is crucial for intellectual property protection, patent management, and information retrieval. Designing an accurate and automated patent classification model can provide significant support for patent inventors and examiners. While existing text classification models perform well with patents in commonly-used languages such as English and Chinese, they struggle with patents written in low-resource languages from the ASEAN region. This challenge is often due to the lack of labeled data, a common issue across various NLP tasks, including text classification, named entity recognition (NER), and sentiment classification.

Recently, multilingual pre-trained models have been widely used to fine-tune and train text classification models. Experiments have proven that fine-tuned multilingual pre-trained models perform well in situations with limited labeled data. One of these models is multilingual BERT, also called mBERT, proposed by Devlin et al.[1] and trained on 104 languages. With the powerful multilingual text representation capability of mBERT, Pelicon et al.[2] have created text representations for news datasets in Slovenian and Croatian and successfully transferred sentiment classification tasks across these two languages. XLM by Conneau and Lample[3], despite being pre-trained using masked language modeling (MLM) and causal language modeling (CLM) on mBERT, added a module called translation language modeling (TLM) to splice parallel translation sentence pairs. XLM has reported state-of-the-art performance on several NLP downstream benchmarks and has been widely used in many multilingual tasks for text representation. Choi et al.[4] validated the hypothetically strong cross-lingual transfer properties induced by XLM pre-training, and experiments with XLM-RoBERTa (XLM-R) indicated that cross-lingual transfer is most pronounced in semantic textual similarity. Considering the shortcomings of existing pre-trained models, new models are constantly being proposed, such as MultiFit by Eisenschlos et al.[5], which facilitates practitioners in effectively training and fine-tuning language models in their own language. However, while pre-trained models contribute significantly to text representation, there are still areas for improvement.

In addition to selecting the most advanced model at the pre-training stage, some scholars have also enhanced model performance by using bilingual dictionaries or multilingual parallel corpora [6, 7, 8]. The main purpose of using bilingual dictionaries or parallel corpora is to obtain the mapping relationship between different languages, which is valuable for models to learn cross-language text embedding. Moreover, some studies have shown that in cross-lingual classification tasks, models that introduce the mapping relationship between languages achieve better results than those relying only on monolingual data [9, 10, 11, 12]. Assuming we use a multilingual parallel corpus to assist training, one method is multi-task modeling [13], which involves training multiple related tasks together based on shared representations. The domain-related information possessed by related tasks serves as a derivation bias to improve the generalization effect of the main task. Some researchers have adopted multi-task learning at the stage of text representation to obtain better representations [14, 15, 16]. Others have improved the model's accuracy from the perspective of auxiliary tasks. For example, in the popular research field of sentiment classification, researchers have focused on the correlation between different emotions [17, 18, 19, 20] and the practice of adding auxiliary tasks. In text classification research, the basic goal is to use the potential correlation between related tasks to obtain common features, including language modeling [21] and domain information [22, 23], to enhance the model's performance.

In this work, we first adopt a multilingual pre-trained model to enhance text representation. Then, we jointly train a multilingual text classifier and a cross-lingual text similarity discriminator using a multi-task learning approach. This allows us to learn multilingual patent text embeddings effectively. During this process, the patent classification task and the text similarity discrimination task interact through shared embedding layers, continuously adjusting their respective parameters. The text similarity discrimination task, as a sub-task within the multi-task framework, influences the main patent classification task through its gradient adjustments.

The motivation behind this method is to learn high-

PatentSemTech'24: 5th Workshop on Patent Text Mining and Semantic Technologies, July 14-18, 2024, Washington D.C., USA
*Corresponding author.
† Both authors contributed equally.
pengyx9@mail3.sysu.edu.cn (Y. Peng); tongxy7@mail2.sysu.edu.cn (X. Tong); zhangyy275@mail2.sysu.edu.cn (Y. Zhang); luyonghe@mail.sysu.edu.cn (Y. Lu)

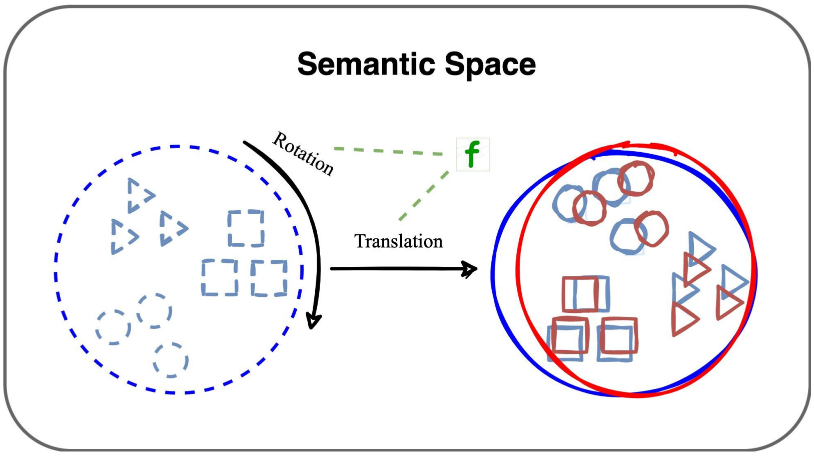

Figure 1: Mapping different languages into the same semantic space

quality multilingual text embeddings, including sentence-level and paragraph-level, in low-resource scenarios. The contributions of this research are twofold. Firstly, we construct a patent corpus for low-resource languages of ASEAN countries and a parallel patent corpus combining common and low-resource languages. Secondly, we combine state-of-the-art multilingual pre-trained models with multi-task frameworks for training. Specifically, the shared word embedding layer enables both the text classification task and the text similarity task to update parameters simultaneously during training. By jointly training these models, the framework can leverage both text classification data and parallel text pairs to learn multilingual text embeddings and achieve better patent classification results.

2. Methodology

In this work, multi-task learning is used to jointly train patent text representations. The main task is patent classification, while the secondary task is cross-lingual text similarity, which transfers additional linguistic semantic information to the main task.

2.1. Patent Classification

As the main task, patent classification involves dividing patents into multiple categories. During training, the objective is to minimize classification error. We use features extracted from a multilingual pre-trained model as the text representation, and a two-layer feed-forward network with a softmax layer as the classifier.

Given a multi-class classification scenario, the softmax layer normalizes output values to ensure that each class's probability is non-negative and their sum is one. The calculation formula for one label's probability is shown in Eq.(1). Let P denote the probability that the patent text belongs to the i-th class, and K denote the number of labels. Here, e is a natural constant, and z represents the input vector's power value.

$$P_i = \frac{e^{z_i}}{\sum_{j=0}^{K-1} e^{z_j}},\ i \in \{0, 1, ..., K-1\} \tag{1}$$

The following equation Eq.(2) is about the loss function, in which J denotes the loss of patent text classification and y represents the real label of the example. If the model predicts correctly, y is 1; otherwise, it is 0.

$$J_{clsf} = -\sum_{i=0}^{K-1} y_i \cdot log(P_i) \tag{2}$$

2.2. Cross-lingual Text Similarity Task

The cross-lingual text similarity task actually learns more language-related semantic information by increasing the correspondence between different languages as auxiliary signals. The aim of this task is to map different languages into the same semantic space, which means finding the mapping function f for two languages. Assuming that one language is Language A and the other is Language B, our purpose is to determine the parameters that satisfy Eq.(3).

$$f(L_A) = L_B \tag{3}$$

We regard function f as a series of operations, such as transforming and rotating the semantic space, as shown in Fig.1. However, f consists of a complex neural network, which is difficult to express as a specific mathematical formula.

Text similarity discrimination is essentially a binary classification task. The final prediction results of the model are either similar or dissimilar. The calculation method of the loss is shown in Eq.(4), where J represents the loss of the discriminator and y represents the sample label. If two sentences or paragraphs are similar, the label is 1; otherwise, the label is 0. Similarly, P refers to the probability that y belongs to a positive sample (labeled as 1).

$$J_{\text{sim}} = -\left[y_i \cdot \log(P_i) + (1 - y_i) \cdot \log(1 - P_i)\right] \tag{4}$$

Finally, the total loss of multi-task cross-lingual patent classification is the sum of two loss functions, as shown in Eq.(5).

$$J_{\text{total}} = J_{\text{clsf}} + J_{\text{sim}} \tag{5}$$

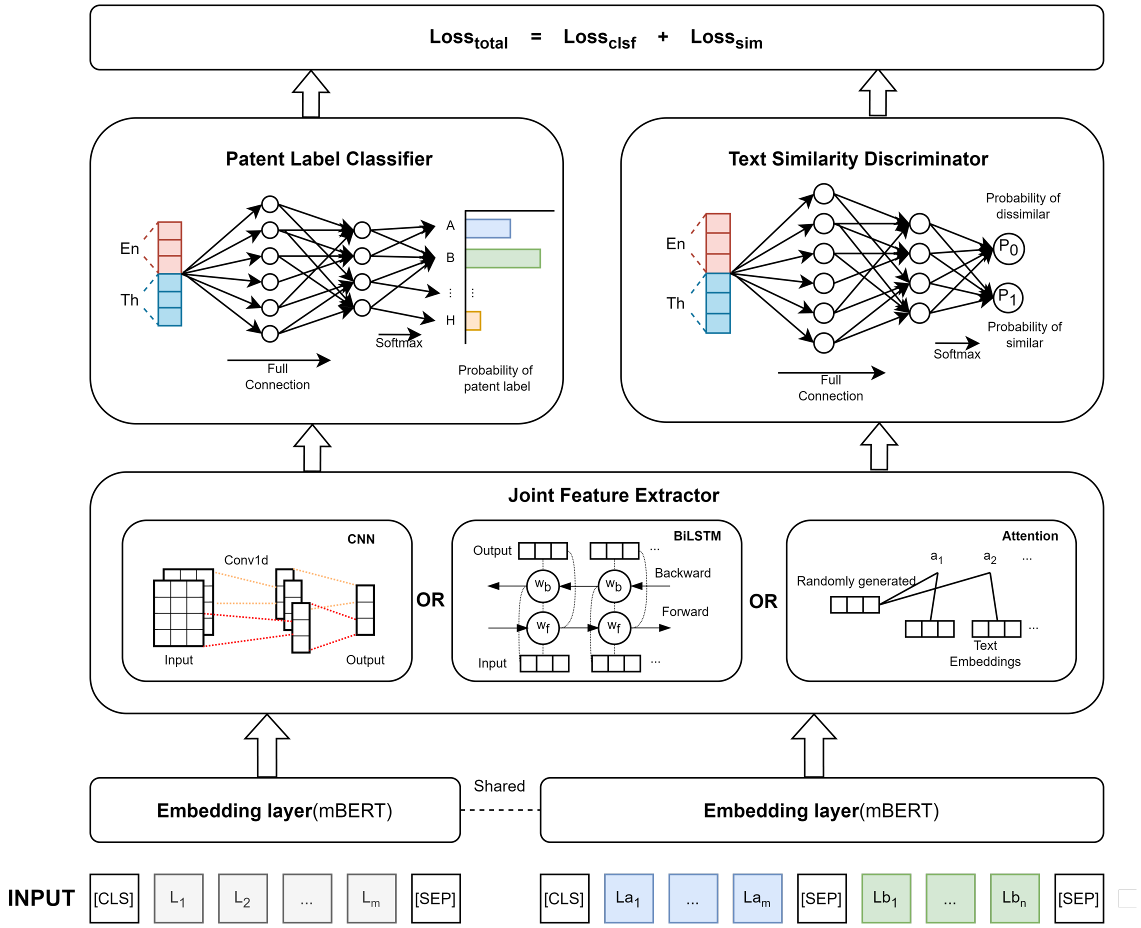

Figure 2: The structure of the proposed model incorporates: (1) an Embedding layer using mBERT for generating multilingual text representations, (2) a Joint Feature Extractor for extracting valuable features, (3) a Patent Label Classifier for prediction, and (4) a Text Similarity Discriminator for calculating the similarity of text pairs from different languages.

3. Model

The model we proposed consists of an embedding layer, a joint feature extractor, a patent label classifier, and a text similarity discriminator, as shown in Fig.2.

3.1. Embedding layer

The purpose of the embedding layer is to convert text into vectors, which are numerical representations that can be processed by machine learning models. In this work, we chose English, a widely-used language, and Thai, one of the ASEAN languages, to demonstrate the model's capability to handle multilingual data.Despite the different inputs, the text representation model mBERT shares a series of parameters.

To optimize for two distinct tasks—patent label classification and cross-lingual text similarity discrimination—we modify the form of inputs according to their specific requirements. For patent label classification, the input consists of the entire sentence or paragraph in either English or Thai. This task aims to classify the patent text into predefined categories based on its content.

In contrast, the input for the similarity discrimination task involves paired sentences or paragraphs in English and Thai, separated by the delimiter [SEP]. This setup enables the model to learn the relationship and semantic similarity between texts in different languages. Despite the different input structures, the text representation model mBERT shares a series of parameters, ensuring that the multilingual embeddings are consistent and comparable across both tasks.

Furthermore, by leveraging mBERT's pre-trained capabilities, we significantly reduce the amount of data required for effective training. This is particularly advantageous in scenarios involving low-resource languages, where labeled data is scarce. The embedding layer thus plays a critical role in bridging linguistic gaps and enhancing the model's applicability to diverse multilingual tasks.

3.2. Joint Feature Extractor

In our conception, we aim to learn relationships between different languages and incorporate them into text classification as complementary knowledge. If the feature extractor can extract features that satisfy both text classification and text similarity tasks, the model may effectively map text in different languages to the same semantic space. This unified semantic space allows for improved cross-lingual understanding and classification accuracy.

To achieve this, we designed a joint feature extractor with the following combinations: mBERT+FC, mBERT+CNN, mBERT+BiLSTM, and mBERT+Attention.

(1) mBERT+FC: This combination uses a fully connected

layer to reduce the dimensionality of the embeddings from mBERT. It provides a straightforward approach to integrating and processing the extracted features.

(2) mBERT+CNN: TextCNN captures local patterns and n-grams within the text. By applying convolutional filters, it detects key phrases and local dependencies, which are crucial for understanding the contextual meaning in sentences and paragraphs.

(3) mBERT+BiLSTM: BiLSTM captures long-range dependencies and sequential patterns by processing the text in both forward and backward directions. This approach is essential for understanding the nuances and intricate relationships within complex patent texts.

(4) mBERT+Attention: Attention mechanisms allow the model to focus on the most relevant parts of the text, dynamically weighing the importance of different words and phrases. This selective focus enhances the model's ability to highlight critical information, improving the accuracy of both classification and similarity tasks.

This joint training process ensures that the embeddings are optimized for multiple tasks simultaneously, improving the model's overall performance. The joint feature extractor enables the model to understand and classify multilingual patent texts more effectively, particularly in low-resource scenarios where annotated data is limited.

3.3. Cross-lingual Text Similarity Discriminator

Generally, the similarity between two sentences or paragraphs is a score representing the relevance and likelihood between them. Despite using the same words with identical writing, text similarity should encompass both syntactic and semantic information. Syntactic similarity is based on string matching, while semantic similarity refers to the similar meaning between words, sentences, paragraphs, and even documents. Since the calculation of text similarity in this paper involves different languages, more attention should be given to the degree of semantic similarity across languages. When calculating semantic similarity, it is natural to consider a common semantic vector space [24, 25], which can be used to calculate the similarity between texts within the same dimension.

By leveraging the shared semantic space provided by the multilingual pre-trained model, the cross-lingual text similarity discriminator effectively measures the semantic relevance between texts in different languages. This capability is essential for enhancing the performance of multilingual applications, ensuring that the model can accurately identify and utilize semantic similarities across languages. The output of the discriminator is the probability of text similarity, represented by a binary classification: 1 for similar and 0 for dissimilar.

3.4. Patent Label Classifier

The patent label classifier corresponds to the patent classification task, which is the main goal of our work. This classifier is designed to categorize patent texts into predefined classes based on their content, helping in the efficient organization and retrieval of patent information.

The classifier operates by first receiving the vector representations from the joint feature extractor. These vectors encapsulate the semantic information derived from the original patent texts. To make these high-dimensional vectors manageable and suitable for classification, the classifier reduces their dimensionality through a fully connected layer. This reduction helps in simplifying the complexity of the data, making the subsequent classification process more efficient and effective.

An important feature of our patent label classifier is its applicability to multilingual data. Since both English and Thai texts are converted into vectors using the same embedding process, the classifier can seamlessly handle patents in both languages.

4. Experiment

4.1. Datasets

In this study, we constructed two main datasets: a parallel corpus for training the similarity computation model and a classification corpus for the classification task.

4.1.1. Cross-lingual Parallel Corpus

The cross-lingual parallel corpus is a multilingual dataset composed of original texts and their corresponding target texts by translation. Given the scarcity of available corpora in the field of patents, we independently built a patent parallel corpus for ASEAN languages. Specifically, we selected 10,000 English patents from the Google BigQuery dataset and translated their titles and abstracts of claims into Thai using the Google Translate API. This translation process created a bilingual parallel corpus, enabling the model to learn mapping relationships between texts from different languages. The parallel corpus includes both sentence-level and paragraph-level data, with labels indicating whether the text pairs have translation correspondence (1) or not (0), with the latter obtained through shuffling.

4.1.2. Patent Classification Corpus

For the classification task, we accessed patent data in both English and Thai from the Google BigQuery dataset by writing SQL queries to randomly download patents. Considering that Lee and Hsiang [26] have shown that only patent claims are sufficient to complete the classification task, we constructed a patent classification corpus that includes the patent title, the abstract of the claim, and the international patent number, which corresponds to the Cooperative Patent Classification (CPC) system. The CPC is a result of a partnership between the EPO and the USPTO to develop a common, internationally compatible classification system for technical documents, including sections, classes, subclasses, groups, and complete classification symbols.

To ensure the dataset was balanced, we used random sampling to adjust the number of patents in each language. After processing, the total amount of data for each class is very close, with approximately 8,000 samples for each section, although slight variations occur due to the random sampling process. The dataset is then divided into a training set, validation set, and test set in the ratio of 8:1:1, resulting in 64,000 training samples, 8,000 validation samples, and 8,000 test samples. See details in Table 1.

By constructing these two datasets, we aim to facilitate both the cross-lingual text similarity task and the patent classification task, ensuring that our multi-task learning framework can leverage the rich semantic information from multilingual texts.

Table 1
Details of Dataset

Section	Totals	Training Set	Validating Set	Testing Set
A	7508	6006	751	751
B	9555	7643	956	956
C	8413	6731	841	841
D	8170	6536	817	817
E	7908	6326	791	791
F	8774	7020	877	877
G	7417	5933	742	742
H	8113	6491	811	811

Table 2
Results of Experiments

Model	Acc			Macro-F1		
	Base	Multi-Task (Par-level)	Multi-Task (Sen-level)	Base	Multi-Task (Par-level)	Multi-Task (Sen-level)
mBERT+FC	0.6914	0.7372	**0.7453**	0.6883	0.7360	**0.7433**
mBERT+CNN	0.7475	0.7515	**0.7575**	0.7410	0.7473	**0.7529**
mBERT+BiLSTM	0.7184	**0.7388**	0.7307	0.7134	**0.7348**	0.7255
mBERT+Attention	0.7610	0.7618	**0.7654**	0.7565	0.7542	**0.7603**

*Multi-Task(Par-level): Cross-lingual text similarity task with paragraph-level parallel corpus
*Multi-Task(Sen-level): Cross-lingual text similarity task with sentence-level parallel corpus

4.2. Setting

In the experiments, we set the following parameters. The number of iterations is 20 rounds, and the learning rate is uniformly set to 1e-4. Since the two tasks need to use shared text embeddings, the parameters of the two pre-trained models are shared, which is equivalent to using the same mBERT for the vectorized representation of text. Additionally, since BERT is a 12-layer transformer structure, we freeze the parameters of the first 9 layers of mBERT and only allow the parameters in the last three layers to be updated, with the aim of saving time.

For training, we alternated between text classification and text similarity tasks in each batch. Specifically, one batch would be used to update the model based on the text classification task, and the next batch would be used to update the model based on the text similarity task. This alternating training approach helps in jointly optimizing the model for both tasks.

The input dimension for mBERT is 768, which corresponds to the size of the embeddings produced by the model. After passing through the feature extractor (which can be a fully connected layer, CNN, BiLSTM, or Attention mechanism), the representations are reduced to 128 dimensions before being fed into the classifier or similarity discriminator.

4.3. Evaluation Methods

We adopt the commonly used classification accuracy (Acc) and macro-F1 value as evaluation metrics to measure our proposed model. To comprehensively analyze the performance of our multi-task model, we choose fully connected (FC) layers, CNN, BiLSTM, and Attention as feature extraction models for comparison. In addition, we set corresponding single-task learning models as baseline methods. The evaluation was carried out on both the patent classification task and the text similarity task to ensure a holistic assessment of the model's performance.

4.4. Results Analysis

With the settings above, the results of the models are shown in Table 2. Compared to the baseline experiments, our proposed multi-task learning model shows better performance in patent classification overall. The results indicate that multi-task learning can achieve performance improvements through related auxiliary tasks during the training process. After multi-task learning, the performance gap among models is reduced to less than 3.47% in accuracy and less than 3.48% in macro-F1. This suggests that learning sentence pair similarity classification is the most important factor affecting the performance of multi-task learning models. The mBERT+Attention combination achieves the best results in both single-task and multi-task learning. Attention uses an auto-focusing mechanism to capture the most important semantic information in sentences, enhancing text representation and classification performance. Results on different datasets in multi-task learning show that sentence-level learning performs better than paragraph-level learning, except in the mBERT+BiLSTM model. We suppose that sentence-level learning can focus more on local information, which is representative of the text topic. Meanwhile, CNN and Attention have advantages in capturing local information to an extent, resulting in good outcomes in our multi-task learning. BiLSTM is more effective at obtaining semantic information from long texts and performs better in paragraph-level learning, but the introduction of BiLSTM has not brought significant improvement in multi-task learning overall.

5. Conclusion

The results show that joint multi-task learning of patent text classification and cross-lingual text similarity based on a multilingual pre-trained model is effective, improving accuracy by 5.39% and macro-F1 by 5.50% over the baseline, with the highest result reaching 76.54% in accuracy and 76.03% in macro-F1. This demonstrates that joint multitasking can fully utilize the mapping relationship between two languages in parallel text pairs and incorporate this information into the main task of patent classification through the sub-task of discriminating text similarity. Theoretically, the overall scheme designed in this study can be transferred to any low-resource language, and its application can be extended to multiple tasks such as sentiment classification and natural language reasoning, not only text classification. In addition to the portability of tasks, the domain is also transferable; that is, in addition to patents, it can also be applied to fields such as medicine and social media.

6. Discussion and Future Work

In this paper, we demonstrate that cross-lingual text discrimination can supplement semantic information for multilingual text classification tasks. However, this approach still requires the use of parallel corpora or translation text pairs. Considering the convenience of data acquisition, future research needs to focus on reducing dependence on large data volumes and enhancing the model's ability to learn from limited data. Currently, technologies such as adversarial transfer and domain adaptation can address this expectation. These technologies eliminate the reliance on parallel data, enabling the training of models suitable for multiple languages in an unsupervised manner. However, these technologies are still in development, particularly in addressing the issue of distributional shift across languages. Our future work will also explore more advanced methods for intelligent multilingual information processing.

Acknowledgement

The authors warmly thank reviewers for their valuable suggestions. This research was partly supported by Key-Area Research and Development Program of Guangdong Province (NO.2021B0101420004).

References

[1] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2018).
[2] A. Pelicon, M. Pranjić, D. Miljković, B. Škrlj, S. Pollak, Zero-shot learning for cross-lingual news sentiment classification, Applied Sciences 10 (2020) 5993.
[3] A. Conneau, G. Lample, Cross-lingual language model pretraining, Advances in neural information processing systems 32 (2019).
[4] H. Choi, J. Kim, S. Joe, S. Min, Y. Gwon, Analyzing zero-shot cross-lingual transfer in supervised nlp tasks, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 9608–9613.
[5] J. Eisenschlos, S. Ruder, P. Czapla, M. Kadras, S. Gugger, J. Howard, MultiFiT: Efficient multi-lingual language model fine-tuning, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5702–5707. URL: https://aclanthology.org/D19-1572. doi:10.18653/v1/D19-1572.
[6] J. Zhang, L. Zhu, J. Liu, Unsupervised cross-language model for patent recommendation based on representation, Data Analysis and Knowledge Discovery 4 (2020) 93–103.
[7] M. Artetxe, H. Schwenk, Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, Transactions of the Association for Computational Linguistics 7 (2019) 597–610.
[8] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, K. Weinberger, Adversarial deep averaging networks for cross-lingual sentiment classification, Transactions of the Association for Computational Linguistics 6 (2018) 557–570.
[9] E. Schumacher, J. Mayfield, M. Dredze, Cross-lingual transfer in zero-shot cross-language entity linking, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 583–595.
[10] X. Meng, R. Cui, Y. Zhao, M. Fang, Multilingual text classification method based on bi-directional long short-term memory and convolutional neural network, Application Research of Computers (2020) 037.
[11] R. Zhang, C. Westerfield, S. Shim, G. Bingham, A. Fabbri, W. Hu, N. Verma, D. Radev, Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations, in: 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, Association for Computational Linguistics (ACL), 2020, pp. 3173–3179.
[12] W. Ma, H. Yu, K. Zhao, D. Zhao, J. Yang, Tibetan-chinese cross-lingual word embeddings based on muse, in: Journal of Physics: Conference Series, volume 1453, IOP Publishing, 2020, p. 012043.
[13] Z. Zhang, Y. Su, X. Niu, F. Gao, Y. Zhao, D. R, Q, Domain information sharing method in mongolian-chinese machine translation application, Computer Engineering and Applications 56 (2020) 9.
[14] G. Lu, J. Gan, J. Yin, Z. Luo, B. Li, X. Zhao, Multi-task learning using a hybrid representation for text classification, Neural Computing and Applications 32 (2020) 6467–6480.
[15] B. Tian, Y. Zhang, J. Wang, C. Xing, Hierarchical inter-attention network for document classification with multi-task learning, in: Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19, 2019.
[16] K. Singla, D. Can, S. Narayanan, A multi-task approach to learning multilingual representations, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, pp. 214–220.
[17] N. Lin, S. Fu, X. Lin, L. Wang, Multi-label emotion classification based on adversarial multi-task learning, Information Processing & Management 59 (2022) 103097.
[18] K. B. Nelatoori, H. B. Kommanti, Multi-task learning for toxic comment classification and rationale extraction, Journal of Intelligent Information Systems (2022) 1–25.
[19] J. Barnes, E. Velldal, L. Øvrelid, Improving sentiment analysis with multi-task learning of negation, Natural Language Engineering 27 (2021) 249–269.

[20] X. Gu, K. Xia, Y. Jiang, A. Jolfaei, Multi-task fuzzy clustering–based multi-task tsk fuzzy system for text sentiment classification, Transactions on Asian and Low-Resource Language Information Processing 21 (2021) 1–24.

[21] Y. Qi, S. Lin, Multi-task learning with bidirectional language models for text classification, in: 2019 International Joint Conference on Neural Networks (IJCNN), 2019.

[22] J. Xie, J. Li, S. Kang, Q. Wang, Y. Wang, A multi-domain text classification method based on recurrent convolution multi-task learning, Journal of Electronics & Information Technology 43 (2021) 2395–2403.

[23] W. Zhao, H. Gao, S. Chen, N. Wang, Generative multi-task learning for text classification, IEEE Access 8 (2020) 86380–86387.

[24] J. Eronen, M. Ptaszynski, F. Masui, M. Arata, G. Leliwa, M. Wroczynski, Transfer language selection for zero-shot cross-lingual abusive language detection, Information Processing & Management 59 (2022) 102981.

[25] G. Glavaš, M. Franco-Salvador, S. P. Ponzetto, P. Rosso, A resource-light method for cross-lingual semantic textual similarity, Knowledge-based systems 143 (2018) 1–9.

[26] J.-S. Lee, J. Hsiang, Patent classification by fine-tuning bert language model, World Patent Information 61 (2020) 101965.