4-2022

Learning for amalgamation: A multi-source transfer learning framework for sentiment classification

Cuong V. Nguyen

Khiem H. Le

Hong Quang PHAM
Singapore Management University, hqpham.2017@phdis.smu.edu.sg

Quang H. Pham

Binh T. Nguyen

Learning for amalgamation: A multi-source transfer learning framework for sentiment classification

Cuong V. Nguyen [b,c,d], Khiem H. Le [b,c,d], Anh M. Tran [b,c,d], Quang H. Pham [a], Binh T. Nguyen [b,c,d,*],

a Singapore Management University, Singapore

b AISIA Research Lab, Ho Chi Minh City, Viet Nam

c Department of Computer Science, University of Science, Ho Chi Minh City, Viet Nam

d Vietnam National University, Ho Chi Minh City, Viet Nam

Corresponding author: ngtbinh@hcmus.edu.vn (B.T. Nguyen).

Published in Information Sciences, (2022 April), 590, 1-14.

DOI: 0.1016/j.ins.2021.12.059

Abstract:

Transfer learning plays an essential role in Deep Learning, which can remarkably improve the performance of the target domain, whose training data is not sufficient. Our work explores beyond the common practice of transfer learning with a single pre-trained model. We focus on the task of Vietnamese sentiment classification and propose LIFA, a framework to learn a unified embedding from several pre-trained models. We further propose two more LIFA variants that encourage the pre-trained models to either cooperate or compete with one another. Studying these variants sheds light on the success of LIFA by showing that sharing knowledge among the models is more beneficial for transfer learning. Moreover, we construct the AISIA-VN-Review-F dataset, the first large-scale Vietnamese sentiment classification database. We conduct extensive experiments on the AISIA-VN-Review-F and existing benchmarks to demonstrate the efficacy of LIFA compared to other techniques. To contribute to the Vietnamese NLP research, we publish our source code and datasets to the research community upon acceptance.

Keywords: LIFA, Low-resource NLP, Mixture of experts, Sentiment classification, Transfer learning

1. Introduction

Sentiment analysis has been extensively studied for the last two decades and has had a lot of practical applications in natural language processing (NLP), data mining (DM), information retrieval (IR), social networks, and e-commerce [1], [2]. With the rise of deep learning, there has been a tremendous success in sentiment classification for popular languages such as English and Chinese [3]. Besides the expressive power of deep neural networks, this success is also attributed to the transfer learning approach with high-quality pre-trained models such as Word2Vec [4], BERTs [5], [6], and GPTs [7]. Given the ubiquitous of pre-trained models, one might face the multi-source transfer learning problem where it is difficult to choose the appropriate pre-trained model for a task at hand. Moreover, it could be more beneficial to take advantage of several pre-trained models based on different architectures and trained on a diverse corpus.

In the multi-source transfer learning problem, most existing methods assume having access to a set of source datasets. They aim to transfer the source knowledge to perform well on a single target dataset. Various strategies have been developed under this setting and have shown promising results across a wide range of applications [8], [9]. However, having access to many source datasets might not be a realistic assumption and can even be prohibited in real-world scenarios. In practice,

it is more common to obtain and use a pre-trained model per source dataset, while the source data are kept away from access due to privacy concerns. Therefore, our work focuses on the multi-source transfer learning scenario where one only has access to the pre-trained models rather than the source data. This setting has recently gained much interest [10], but not yet widely explored in the NLP domain.

In this paper, we propose LIFA (**L**earn**I**ng **F**or **A**lmagamation), a learning framework to combine various pre-trained models from one source dataset into a unified embedding that can perform better than its components in the target task. Motivated from Mixture of Experts (MOE) [11], LIFA introduces an additional gating layer trained to combine existing embeddings and produce the final embedding for classification. As a result, without having access to the training data, our proposed LIFA takes advantage of rich-knowledge sources and allows our sentiment classification to leverage features dynamically and selectively from each source through a probabilistic mixture of expert mechanisms. Notably, it helps tackle the shortcomings of existing algorithms while only acquiring a small amount of data for training and boosting performance remarkably.

We conducted extensive experiments on four different datasets: two Vietnamese datasets, including one public dataset of AIVIVN and our large-scale dataset collected from Vietnamese e-commerce websites (namely AISIA-VN-Review-F Dataset), and two multi-domains English benchmark datasets, including Multi-Domain Dataset and Amazon Reviews Dataset. We consider three sources for each classification problem. For Vietnamese classification, we employ the sources of Fasttext [12], BERT [5] and PhoBERT [6]. For English classification, we use the sources of FastText [12], BERT [5], XLM [13]. Our LIFA variants of SIGMOID, WTA, COOP consider all three sources and learn how to combine them. The results show that our LIFA-SIGMOID consistently outperforms other approaches that transfer only from a single source and show a better performance than a traditional ensembling method of concatenation. In summary, our work makes the following contributions:

(a) We propose LIFA, a novel framework for transfer learning using multiple pre-trained sources with different embedding sizes. We also consider and compare different variants of LIFA, including LIFA-SIGMOID, LIFA-WTA, and LIFA-COOP.
(b) Through extensive experiments, we demonstrate the efficacy of our proposed LIFA compared to other existing techniques. Meanwhile, LIFA-SIGMOID shows the best performance.
(c) We construct the AISIA-VN-Review-F Dataset consisting of over 450 K reviewing comments that we manually labeled. We will publish the AISIA-VN-Review-F Dataset, both the raw and post-processed versions, and LIFA's implementation to facilitate the research community in Vietnamese sentiment analysis.

The rest of this paper is organized as follows. Sections 1 and 2 formulate the main problem and provide an overview of the literature. In Section 3, we present our proposed LIFA framework with different variants (LIFA-SIGMOID, LIFA-WTA, LIFA-COOP). In Section 4, we introduce the AISIA-VN-Review-F dataset and conduct extensive experiments to validate the efficacy of LIFA compared to existing techniques. Finally, we conclude this work in Section 5.

2. Related Work

2.1. Overview

There have many studies investigated the problem of sentiment classification. Traditionally, the methods usually employ the word embedding Word2Vec [4] or Fasttext [12] to transform sentences to vectors, design raw architectures, and train from scratch such as deep character-level CNNs [14], shallow word-level [15], recurrent networks [16], combination of convolutional and recurrent networks [17], or residual-based networks [18]. However, these methods' drawbacks come from the requirements to design and test with many raw architectures and train on a large-scale dataset to guarantee a good performance. With the rising of transfer learning, it has become increasingly common to utilize pre-trained models and finetune on a downstream task. The pre-trained models are usually trained on very large-scale datasets and are heavily designed with a complex architecture of millions or billions of parameters. This approach has been successfully applied and obtained state-of-the-art results in many of the most common NLP benchmarks but mainly limited to the English language such as BERT [5], XLNET [19], XLM [13], or UniLM [20]. For other languages, several variants of pre-trained models can be found such as AraBERT [21] for Arabic, ChineseBERT [22] for Chinese, DutchBERT [23] for Dutch, FrenchBERT [24] for French, PhoBERT [6] for Vietnamese. However, there is one rising question for exploring multi-source transfer learning to borrow knowledge from multi-trained models. There have existed several approaches proposed to explore transfer learning under multi-sources. Zhang et al. [25] assumed to have $\boldsymbol{m}$ word embeddings with corresponding dimensions and then join these at the final layer by simply concatenating to form the final feature vector. Yin and colleagues [26] introduced an ensemble approach of combining different public embedding sets with the aim of learning meta-embeddings, utilizing a simple neural network or Singular Value Decomposition (SVD) to define a projection from the meta-embedding space to the known embeddings.

2.2. Multi-source Transfer Learning

Multi-source transfer learning in NLP is an important approach in facilitating the development of tasks and languages that only have a limited amount of labeled data. Existing multi-source transfer learning approaches can be broadly categorized into two groups, based on the availability of the source data when learning on the target task.

With Access to Source Data. In the first category, existing methods assume having access to both the source domains' data and their corresponding pre-trained models. Chen et al. [27] presented a mixture-of-experts (MoE) model [11] to combine a set of language expert networks, one per source language, each responsible for learning language-specific features for that source language during training. Jian and colleagues [28] presented two weakly supervised directions for the cross-lingual named entity recognition (NER) with the assumption that there is no human annotation in a target language. By automatically creating labeled NER data for the target language using an annotation projection on selected corpora and projecting word embeddings from the target language to a source language, the proposed techniques bypassed three other weakly supervised approaches on the CoNLL data. Xingjian et al. [29] proposed a new deep transfer earning algorithm, namely XMixup, that could efficiently utilize the knowledge transfer from the source to the target domains for different classification tasks. The experimental results on six vision datasets showed the better performance and efficiency of the XMixup in comparison with several baseline algorithms. Han and co-workers [30] investigated the multi-source domain adaptation for text classification using a new DistanceNet-Bandit model. The proposed method can utilize a multi-armed bandit controller that dynamically takes source domain data (labeled) among different source domains and combines the target domain data (unlabeled) to extract the feature representations during the training process and learn an optimal transfer from sources domains to the target domain. Zheng et al. [31] studied the cross-domain sentiment classification using two parameter-shared adversarial memory networks that could utilize a set of labeled data and unlabeled data in a source domain to predict the polarity of unlabeled samples from the target domains. The proposed memory networks can automatically capture the associated important sentiment words using the attention mechanism without manual selection and share them in both source and target domains to minimize the classification error. The experiments showed that the proposed technique could outperform other state-of-the-art approaches on the Amazon reviews benchmark dataset. Stephen and colleagues [32] proposed an efficient approach for cross-lingual named entity recognition that could use a lexicon to translate annotated data available from different high-resource languages to a low-resource language. Then, using the newly translated data, it could learn the corresponding NER model in the target language. The method also outperformed other state-of-the-art NER results in seven languages. Phillip et al. [33] proposed a new adversarial learning scheme with multilingual BERTs for zero-resource cross-lingual text classification and named entity recognition. The method used English text (labeled) and unlabeled non-English text (unlabeled) during training and selected hyperparameters using English evaluation sets. The experimental results demonstrated the improved performance on the multilingual ML- Doc text classification and CoNLL 2002/2003 named entity recognition tasks.

Without Access to Source Data. In the second category, existing methods only need access to the pre-trained models on the source domains while the source data are not required, which could be more suitable for real-world applications because of the data privacy. [10] investigated the multi-source transfer learning problem and focused on the scenario when the target task only had minimal (few-shot) training samples. The proposed method, MCW [10] learned to weigh the features in the source domains based on the Maximal Correlation Analysis principle [34]. It is important to note that MCW focused on computer vision applications and required having access to the features extracted from the source domains, which is impossible in our application with pre-trained language models (although this constraint is weaker than the first category). Therefore, we cannot empirically compare with MCW in our experiments. For NLP applications, [35] proposed to linearly combine the features of the pre-trained source models via a set of learnable coefficients, each of which is associated with one pre-trained source model. Our LIFA framework can be considered as a generalized version of this strategy because the coefficients are generated from the gating network (as depicted in Fig. 1). Moreover, LIFA requires less effort in analyzing the strategy to combine the source domains compared to [10].

2.3. Transfer Learning for the Vietnamese Language

For the Vietnamese Language, there was little work about utilizing transfer learning for the Vietnamese sentiment analysis. Nguyen et al. [36] finetuned BERT models on Vietnamese datasets and showed experimental results; those using BERT could slightly outperform other models using Glove and FastText. PhoBERT [6] (regarded as the first public large-scale monolingual language and the state-of-the-art model for the Vietnamese language) is trained on a very large corpus of Vietnamese and improves the state-of-the-art in multiple Vietnamese-specific NLP tasks. However, to the best of our knowledge, the problem of multi-source transfer learning for the Vietnamese sentiment classification has not been explored. Moreover, existing methods presented in Section 2.2 assume access to the source domains' data, which may not be suitable in real-world scenarios due to privacy issues. Therefore, our work focuses on a more general setting where only pre-trained models on the source domains are available. Such a scenario has only been explored in vision applications [10] and has not yet been studied in the NLP domain.

3. Methodology

In this section, we first formulate the sentiment classification problem and then describe the proposed **LIFA**, a simple yet effective method for integrating several pre-trained models to improve the performance on the sentiment classification task. LIFA makes a prediction of an input based on a novel gating mechanism to combine the embedding features from several experts, each of which may have **different embedding sizes**. Moreover, LIFA allows for an easy mechanism to enforce cer-

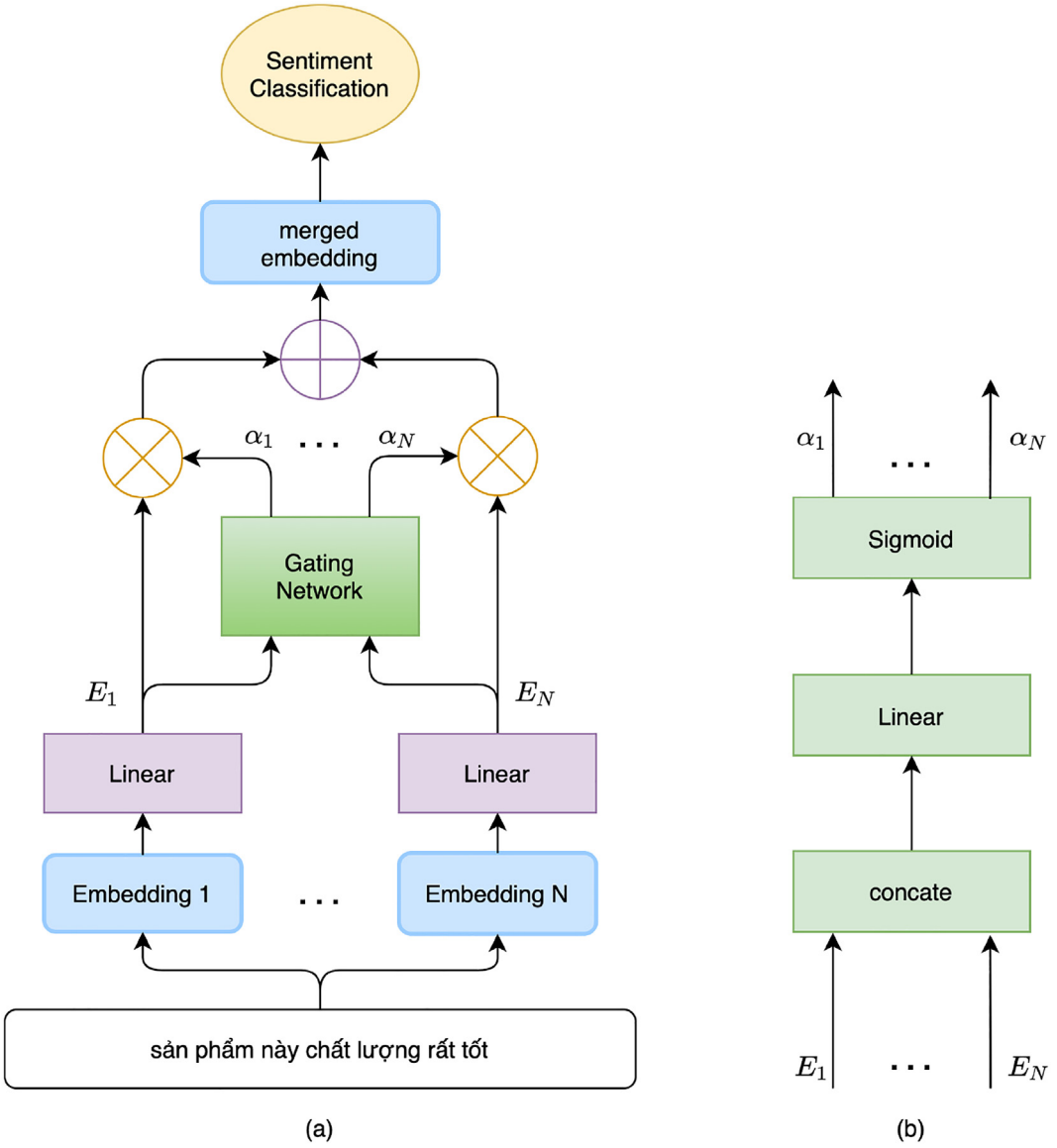

Fig. 1. Our proposed LIFA framework using a gating network. (a) The overall framework demonstrated on N embeddings. (b) Our Gating Network architecture. It is worth noting that from given input data, LIFA can select N different embedding models (which can have distinct embedding dimensions) for extracting feature vectors. These feature vectors then go through linear transformation layers to project these feature vectors into the same feature space. Finally, a gating network can be employed for combining these newly computed features to learn an optimal classifier for the sentiment classification problem.

tain structures to the experts, which results in three variants: (i) LIFA-COOP: experts cooperate with one another, (ii) LIFA-WTA: experts compete with one another, (iii) LIFA-SIGMOID: no specific structure.

3.1. Preliminary

We first introduce the sentiment classification problem studied in this work. Let $\mathscr{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$ be a training set consisting of $\{\boldsymbol{x}, y\}$, where $\boldsymbol{x} = \{x_1, \ldots, x_T\}$ is a training sequence $\boldsymbol{x}$ of T tokens x_j and its corresponding label $y \in \{0, 1\}$, which represents *positive* (1) or *negative* sentiment (0).

A classification model is composed of an embedding model $g(\cdot; \varphi)$ parameterized by φ, and a classifier $f(\cdot; \theta)$ parameterized by θ. For simplicity, we omit the dependency of the embedding and classification models on their parameters φ and θ in

the rest of this paper. Given an input sequence $\boldsymbol{x}$, a sequence of token-embeddings is first generated as: $\boldsymbol{e}(\boldsymbol{x}) = \{g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_T\}$. Then, the classifier takes the embedding tokens as input and makes a prediction as:

$$\hat{y} = f \circ g(\boldsymbol{x}) = f(\{g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_T\}) \tag{1}$$

Both the embedding parameters φ and the classifier's parameters θ are jointly optimized by minimizing the empirical loss $\mathscr{L}(\hat{y}, y)$, which is usually implemented as the cross-entropy loss for classification problems. In practice, the classifier f is implemented as a deep neural network such as Recurrent Neural Networks and its variants [37]. Meanwhile, the embedding model g can be a pre-trained word embedding such as Word2vec [4], GLOVE [38]), or even complex pre-trained models (e.g., BERT [5] or GPT [7]).

3.2. LIFA: Multisource Transfer Learning for Vietnamese Sentiment Classification

One particular challenge often faced in practice is that it is usually costly to label and obtain adequate data to achieve satisfactory results. As a result, one can leverage the language structure through transfer learning schemes from pre-trained models to the embedding model φ. Such approaches are ubiquitous in practice and have been shown to improve performance significantly, especially when training data are limited. However, it gives rise to a multi-source transfer learning problem: given a set of pre-trained models (sources), how can one choose the appropriate model for transfer learning given a task at hand?.

This paper proposes LIFA (**L**earn**I**ng **F**or **A**lmagamation): a framework for tackling this multi-source transfer learning with pre-trained models problem. LIFA employs a Mixture-of-Experts (MoE) layer that learns to combine various sources by taking advantage of the knowledge from all of them. Therefore, LIFA can assign appropriate importance to each expert (pre-trained source model) such that the performance on the target problem is maximized.

Moreover, we propose three variants of LIFA that enforce specific structures on the source knowledge:

(i) LIFA-SIGMOID (Learning for Amalgamation by Cooperation without any constraint) leads the experts working altogether without any constraint by nominating high weights to the experts having good performance and low weights to the experts having a poor performance.
(ii) LIFA-COOP (Learning for Amalgamation by Cooperation) forces the experts to cooperate with each other by smoothing out the weights among experts.
(iii) LIFA-WTA (Learning for Amalgamation by Winner-Take-All) encourages the competition among experts by assigning most of the weight to the expert having the best performance and almost zero weight to other experts.

3.2.1. LIFA-SIGMOID

We assume having access to n embedding models $\boldsymbol{E}_1, \ldots, \boldsymbol{E}_n$ with different embeddings' dimensions. Our LIFA first applies linear transformation layers on the embedding models $\{\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_n\}$ to obtain new embedding models $\{\boldsymbol{E}'_1, \boldsymbol{E}'_2, \ldots, \boldsymbol{E}'_n\}$ that have same dimensionality of K, which can be formulated as follows:

$$\boldsymbol{E}'_i(\boldsymbol{x}) = \boldsymbol{W}_i \boldsymbol{E}_i(\boldsymbol{x}), \tag{2}$$

where $\boldsymbol{W}_i$ is the parameter matrix of the i-th linear layer. We then compute the final embedding as a weighted combination of the experts' embeddings:

$$\boldsymbol{E}(\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{E}'_i(\boldsymbol{x}). \tag{3}$$

Here, $G(\boldsymbol{x})$ is the gating network that can learn to combine the embeddings to predict importance coefficients $[\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$ of each expert. In our experiments, we have up to three experts for each classification problem of Vietnamese or English. In this work, the gating network receives the embeddings $\{\boldsymbol{E}'_1, \boldsymbol{E}'_2, \ldots, \boldsymbol{E}'_n\}$ as the input data and generates an output as a sparse n-dimensional vector. We implement the gating network by multiplying the concatenation of these embeddings by a trainable weight matrix W and then apply the *Sigmoid* function, which can be formulated as follows:

$$G(\boldsymbol{x}) = Sigmoid([\boldsymbol{E}'_1(\boldsymbol{x}), \boldsymbol{E}'_2(\boldsymbol{x}), \ldots, \boldsymbol{E}'_n(\boldsymbol{x})]\boldsymbol{W}). \tag{4}$$

Since there is no constraint on the experts' weights, each expert's weight will be updated proportionally to its contribution in the final prediction. Therefore, this variance of LIFA is named LIFA-SIGMOID. Once acquiring the consolidated embedding vector $\boldsymbol{E}(\boldsymbol{x})$, we can make a prediction $\hat{y}$ in the same manner as Eq. (1). Regarding the sentiment classification module $\boldsymbol{\theta}$, we use a linear layer with the output features of "2", indicating the number of sentiment polarities. Fig. 1 demonstrates the workflow of our LIFA and the architecture of our gating layer.

3.2.2. LIFA-COOP

We now describe a simple and effective technique to enforce a prior structure to experts. Mainly, we are interested in *two* specific structures in which the experts cooperate or compete with one another. Thanks to the gating network's design, we can quickly achieve this goal by changing the gating activation function from *Sigmoid* to *Softmax* in Eq. 4 as follows:

$$G(x) = \text{Softmax}_\tau([\boldsymbol{E}'_1(\boldsymbol{x}), \boldsymbol{E}'_2(\boldsymbol{x}), \ldots, \boldsymbol{E}'_n(\boldsymbol{x})]\boldsymbol{W}), \tag{5}$$

where $\text{Softmax}_\tau : \mathbb{R}^K \to \mathbb{R}^K$ is the standard *Softmax* function with temperature τ defined over K inputs $z_1, z_2, \ldots, z_K$ as:

$$\text{Softmax}_\tau(z_i) = \frac{\exp^{z_i/\tau}}{\sum_{j=1}^{K} \exp^{z_j/\tau}}. \tag{6}$$

It is important to note that as $\tau \to \infty$, we have $\text{Softmax}_\tau(z_i) \approx \text{Softmax}_\tau(z_j), \forall i, j$, i.e., the *Softmax* function with high temperatures will produce a uniform distribution over its input. Conversely, when $\tau \to 0$, the *Softmax* function will converge to a Dirac delta distribution peaking at the largest value input, i.e. $\arg\max_i z_i$. Therefore, we propose implementing the LIFA variants by first replacing the *Sigmoid* activation in LIFA's gating network with the *Softmax* function. For LIFA-COOP, we raise the temperature τ to a high value so that experts will receive similar weight signals regardless of their performance.

3.2.3. LIFA-WTA

LIFA-WTA has the same architecture as LIFA-COOP. However, we instead lower the temperature τ so that only the experts making the correct prediction receive most of the rewards, reflecting the winner-take-all principle. In what follows, we also compare the performance of different variants of LIFA in chosen datasets.

4. Experiments

We design our experiments to investigate the following hypotheses: (i) it is more beneficial to take transfer knowledge from several pre-trained models compared to using just one (Section 4.4 and 4.5); (ii) our LIFA framework can efficiently transfer knowledge compared to the naive strategy of concatenating all the embeddings (Sections 4.4 and 4.5); and (iii) LIFA allows for a flexible mechanism to control to which degree pre-trained knowledge should be utilized (Sections 4.6.1 and a mechanism to enforce prior structure to the experts (Section 4.6.2).

4.1. Datasets

We consider four different datasets in our experiments. Two Vietnamese datasets are used for the Vietnamese sentiment classification, including the **AIVIVN dataset** and the **AISIA-VN-Review-S** (a subset of the AISIA-Review-F). We also extend to English sentiment analysis on two multi-domains datasets: the **Multi-Domain Dataset** [39] and the **Amazon Review Dataset** [40]. The statistics of each dataset are shown in Tables 1 and 2. In the following, we provide the details of the aforementioned datasets.

AIVIVN Dataset is a Vietnamese review dataset that consists of around 16 K reviews in the training set and around 11 K reviews in the testing set. This dataset was used for the Vietnam Sentiment Analysis Challenge 2019[1]. All labels in the testing set are not available and kept private from the competition organizers. We carefully labeled all the testing reviews by ourselves and cross-checked multiple times between our team members and experts to guarantee the labeling quality. Along with this work, we will publish this dataset with our label for further research.

In **AISIA-VN-Review-S** and **AISIA-VN-Review-F** datasets, we first collect 450 K customer reviewing comments from various e-commerce websites. Then, we manually label each review to be either positive or negative, resulting in 358,743 positive reviews and 100,699 negative reviews. We named this dataset the sentiment classification from reviews collected by AISIA, the full version (AISIA-VN-Review-F). However, in this work, we are interested in improving the model's performance when the training data are limited; thus, we only consider a subset of up to 25 K training reviews and evaluate the model on another 170 K reviews. We refer to this subset from the full dataset as AISIA-VN-Review-S. It is important to emphasize that our team spends a lot of time and effort manually classifying each review into positive or negative sentiment. To the best of our knowledge, AISIA-VN-Review-F is the most extensive large-scale dataset for Vietnamese sentiment analysis until now, which can be considered as an additional contribution to the research community in Vietnamese Natural Language Processing. Due to our text data collected from social networks and e-commerce websites, we observe that many informal texts and words do not conform to the usual standard of the Vietnamese language. Thus, we apply various pre-processing steps for the text data, as described in Table 3. Lastly, although some of our pre-processing steps, such as Step #2, may potentially remove additional information about emotions, we choose to standardize all the writing styles in this work for consistent comparison across all datasets. We will also publish the unprocessed dataset to facilitate future research exploring such properties.

Multi-Domain Dataset [39] consists of a short English dataset from four different domains of books, DVD, electronics, kitchen, and housewares, taken from Amazon.com. Also, each domain contains 1 K positive reviews and 1 K negative reviews.

Amazon Review Dataset [40] is a very large-scale English review dataset with 19 different categories and millions of reviews. In this paper, we follow [41] and select a subset of four domains and reviews: Cell Phones and Accessories, Clothing

[1] https://www.aivivn.com/contests/1

Table 1
Two Vietnamese datasets used in our experiments.

AIVIVN Dataset	Positive	Negative	Total
Train	8690	7383	16073
Test	5767	5214	10981
AISIA-VN-Review-S Dataset	Positive	Negative	Total
Train 5 K	3912	1088	5000
Train 15 K	11736	3264	15000
Train 25 K	19559	5441	25000

Table 2
Two English datasets used in our experiments.

Multi-Domain Dataset	Positive	Negative	Total
Books	1000	1000	2000
DVD	1000	1000	2000
Electronics	1000	1000	2000
Kitchen and Housewaves	1000	1000	2000
Amazon Review Dataset	Positive	Negative	Total
Cell Phones and Accessories	10000	10000	20000
Clothing Shoes and Jewelry	10000	10000	20000
Home and Kitchen	10000	10000	20000
Tools and Home Improvement	10000	10000	20000

Shoes and Jewelry, Home and Kitchen, Tools and Home Improvement. We randomly select a subset of 20 K reviews divided equally among the positive and negative sentiment in each domain.

4.2. Baselines

We compare our LIFA with various baselines, from individual models to classical ensemble methods as described below.

Recurrent CNN [17] proposes a combination of Recurrent and Convolutional Neural Networks for text classification and shows its remarkable improvement compared to the individual RNNs or CNNs. The model takes advantage of RNN to capture long-term dependencies and contextual information and CNN to extract local and position-invariant features very well. We implement this model from scratch with the same settings as presented in the original paper.

BERT, PhoBERT, XLM The second set of baselines we consider is transformer-based models. Particularly we consider the pre-trained models of $BERT_{base-multilingual}$ [5], $PhoBERT_{base}$ [6], and XLM [13] for our task. $BERT_{base-multilingual}$ is pre-trained on the cased text on the top 104 languages with the largest Wikipedias, while $PhoBERT_{base}$ is the first public large-scale monolingual language model pre-trained for Vietnamese. These pre-trained models are also regarded as experts for our proposed LIFA.

Concatenation A traditional method of ensembling multiple pre-trained expert models [25] in which the experts' embeddings are concatenated before feeding into the fully connected layers.

Supervised Contextual Embeddings Additionally, we compare LIFA with a multi-source transfer learning method, Supervised Contextual Embeddings (SCEs), which was recently proposed by [35]. SCEs use a convex combination to aggregate the projected output of all different source models into one vector. In contrast, such coefficients are generated from the gating network in our LIFA.

4.3. Implementation

All experiments are performed on a deep learning workstation with Intel Core i9-7900X CPU, 128 GB RAM, and two GPUs RTX-2080Ti with Pytorch framework [42]. For Recurrent CNN [17], we train the model from scratch with the feature embedding of dimension 256 and the Fasttext word embedding vectors[2] with a dimension of 300. For $PhoBERT_{base}$, we first apply the Vietnamese word segmenter RDRsegmenter [43] to process raw data and generate segmented words. We employ a pre-trained model[3] and fine-tune on all datasets. The dimension of feature embedding from the PhoBERT model is 768. For $BERT_{base-multilingual}$ and XLM, we utilizes the pre-trained models[4] and fine-tune on all datasets. The dimension of feature embedding from BERT and XLM are 768 and 1024, respectively. For LIFA, we consider three-component experts: Recurrent CNN, BERT, and PhoBERT for the

[2] https://fasttext.cc/docs/en/crawl-vectors.html
[3] https://github.com/VinAIResearch/PhoBERT
[4] https://huggingface.co/transformers/

Table 3
An illustration for the preprocessing step in **AISIA-VN-Review-F Dataset**.

Step	Description	Customer Review (Raw)	Post-processed Text
1	Lowercase all characters	mua xong về bỏ sọt rác luôn. màn hình kính thì mờ, mặt kính thì trầy xước, nhìn không khác gì đồ cũ phế liệu đem bán cho khách hàng. **NẾU MÀ TIẾP TỤC XEM CÁI NÀY CHẮC LÀ HƯ LUÔN CON MẮT... BỰC MÌNH**... *we throw it into the trash after buying due to the glass screen is blur and scratched, so it looks like a garbage which sold to customer. If continuously watching this screen, it will harm our eyes ... so angry*	mua xong về bỏ sọt rác luôn. màn hình kính thì mờ, mặt kính thì trầy xước, nhìn không khác gì đồ cũ phế liệu đem bán cho khách hàng. **Nếu mà tiếp tục xem cái này chắc là hư luôn con mắt... bực mình**...
2	Correct elongated words	Giao hàng nhanh hơn dự kiến, vải **đẹpppppppppppppppppp**! *The delivery is faster than expected, the fabric is so beautifullllllllll!*	giao hàng nhanh hơn dự kiến vải **đẹp**
3	Remove URLs	Các bác tham khảo ở đây, rẻ hơn hẳn 100-150k **https://noithatluongson.vn/ban-chan-sat** *you guys can refer here, cheaper than 100-150k*	Các bác tham khảo ở đây, rẻ hơn hẳn 100-150k
4	Translate	Mình đặt chiều hôm qua đến sáng nay thì có hàng rồi. Nhanh hú hồn. **Thanks shop** nhé *I ordered yesterday but we received the product today. So fast.*	mình đặt chiều hôm qua đến sáng nay thì có hàng rồi nhanh hú hồn **cảm ơn cửa hàng** nhé
5	Remove punctuation marks and special characters	Hàng đúng chuẩn**,** đóng gói cẩn thận**,** dùng tốt **,** ủng hộ**!** *The product is nice, the packaging is careful, the usage is good.*	Hàng đúng chuẩn đóng gói cẩn thận dùng tốt ủng hộ
6	Exclude other language reviews (Korean, Chinese, English, etc.)	The quality is good and suitable for using at the library, but the click is not good.	-
7	Correct free-style letters and acronyms	dày , êm , rất tốt so với giá tiền . hình in trên miếng lót rất chi tiết và rõ nét . xài **tgian** thì sẽ đánh giá thêm *the product is thick, smooth, deserved with its price. It is very detailed and clear. I will feedback more after usage*	dày , êm , rất tốt so với giá tiền. hình in trên miếng lót rất chi tiết và rõ nét. xài **thời gian** thì sẽ đánh giá thêm

Vietnamese datasets; Recurrent CNN, BERT, and XLM for the English datasets. We train the component experts individually and store the best checkpoint, which is then used as the experts to train LIFA. All methods are optimized to minimize the cross-entropy loss using the Adam optimizer with a batch size of eight over 30 epochs with the early stopping of five based on the validation accuracy. Lastly, we compared the methods over three evaluation metrics: AUC, Accuracy, and F1-score.

4.4. Experimental results on Vietnamese review datasets

Standard evaluation. Tables 4 and 5 report the experimental results on the AIVIVN and AISIA-VN-Review-S Datasets. It is worth noting that all methods considered in this work perform much better than the AIVIN 2019's champion concerning our label test set. For our baselines, we observe that BERT and PhoBERT perform slightly better than Recurrent CNN, thanks to

Table 4
The results of our proposed LIFA and other methods (including the AIVIVN 2019 Sentiment Champion's solution) on **AIVIVN Dataset**. Here, we consider three different performance metrics: Accuracy (ACC), AUC, and F1-score (F1).

Methods	AUC	ACC	F1
AIVIVN 2019 Sentiment Champion	-	-	90.01
Recurrent CNN	98.33	93.42	92.98
BERT	98.82	94.05	93.94
PhoBERT	98.67	94.04	93.79
Concatenation	98.12	94.37	94.09
LIFA-WTA	98.04	93.41	93.02
LIFA-COOP	99.02	95.11	94.87
LIFA-SIGMOID	**99.12**	**95.46**	**95.20**
SCEs	98.75	94.42	94.14

Table 5
The results of our proposed LIFA and other methods on **AISIA-VN-Review-S Dataset** with 10 K, 15 K and 20 K training reviews consecutively. Here, we consider three different performance metrics: Accuracy (ACC), AUC, and F1-score (F1).

Methods	**5 K reviews**			**15 K reviews**			**25 K reviews**		
	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
Recurrent CNN	90.10	86.57	66.53	93.35	88.92	74.35	93.83	89.57	75.00
BERT	90.01	87.41	67.43	94.09	89.45	76.03	94.62	90.00	75.80
PhoBert	91.95	88.04	68.98	94.07	89.66	75.25	94.62	90.57	77.65
Concatenation	92.57	88.14	69.26	94.49	90.24	77.89	95.17	90.6	77.57
LIFA-WTA	92.05	88.06	69.00	94.42	90.04	76.82	94.57	90.81	78.30
LIFA-COOP	92.70	88.11	69.36	95.39	91.18	79.18	95.16	91.07	79.21
LIFA-SIGMOID	**92.95**	**88.58**	**69.55**	**95.71**	**91.42**	**79.83**	**95.79**	**91.76**	**80.18**
SCEs	92.70	88.25	69.39	95.12	91.04	78.53	95.36	91.02	79.25

their stronger transformer backbone. Moreover, the Concatenation baseline performs slightly better than the single model, suggesting that taking advantage of multiple pre-trained models can potentially improve the performance. However, the Concatenation strategy is quite simple and may not efficiently utilize the rich knowledge of all experts, making it performs worse than our LIFA strategies. We also observe that SCEs perform worse than both LIFA-COOP and LIFA-SIGMOID, suggesting that it lacks the ability to combine the source knowledge compared to our proposed LIFA efficiently. Overall, we observe that two LIFA variants (LIFA-COOP and LIFA-SIGMOID) consistently outperform the remaining methods on both datasets across all metrics. For LIFA, we observe that LIFA-SIGMOID achieves the best performance, while LIFA-WTA performs the worst. This result suggests that LIFA-SIGMOID achieves great flexibility in knowledge sharing experts while enforcing prior structure to the experts does not perform well in all scenarios.

Performance with an increasing number of training data. When the number of training samples in the target domain is limited, it is essential to rely on the knowledge stored in the pre-trained models. Therefore, we also report the results of our AISIA-VN-Review-S Dataset in Table 5 with three different amounts of training data: 5 K, 15 K, and 25 K, while keeping the same testing dataset of 170 K reviews. Generally, the performance improves with more training data across all methods, which is easy to understand as deep models require a large amount of training data to achieve good performance. Interestingly, the performance gap between our LIFA-SIGMOID and other methods is most significant with 5 K reviews and becomes stable when moving to 15 k and 25 k reviews. Moreover, we also observe that LIFA-SIGMOID consistently outperforms SCEs across all scenarios, showing that LIFA is a generalized framework of SCEs. This result shows that our LIFA-SIGMOID can efficiently utilize pre-trained knowledge under limited training samples while still maintains its ability to adapt with more training data.

4.5. Experimental results on English review datasets

Similarly, we conduct the experiments of the two English multi-domains datasets (“Multi-Domain Dataset” and “Amazon Reviews Dataset”) [41]. We employ LIFA-SIGMOID, the best one in LIFA variants, and compare with other baselines, as mentioned in [41]. It is worth noting that GLU, GTU, and GTRU are three state-of-the-art models for those two datasets [41]. The experiments on these multi-domain datasets require training the models on one domain and testing them on other remaining domains. Tables 6 and 7 present the experimental results of this experiment.

Consistent with the previous experiments on the Vietnamese datasets. Here we also observe that for the single expert methods such as GLU, GTU, BERT, etc., the performance increases with better backbones: BERT and XLM outperform the other baselines. Second, Concatenation and LIFA-SIGMOID consistently perform better than the remaining baselines, thanks

Table 6
The accuracy between our proposed LIFA and the baselines on the **Multi-Domain Dataset**.

Source → Target	Recurrent CNN	BERT	XLM	Concatenation	LIFA-SIGMOID	GLU [41]	GTU [41]	GTRU [41]
Books → DVD	77.95	79.80	79.40	82.35	**82.75**	79.50	79.25	79.25
Books → Electronics	74.45	75.15	77.35	78.20	**79.40**	71.75	71.75	71.75
Books → Kitchen	76.35	79.70	79.85	82.80	**83.55**	73.00	72.50	72.50
DVD → Books	75.05	77.60	79.15	82.65	**83.65**	78.00	80.25	77.25
DVD → Electronics	73.60	76.65	78.80	79.70	**81.60**	73.00	74.50	69.25
DVD → Kitchen	74.20	78.90	79.75	82.00	**83.85**	77.00	76.00	74.75
Electronics → Books	70.15	77.05	70.15	81.35	**81.35**	71.75	68.75	67.25
Electronics → DVD	70.60	75.10	76.40	79.80	**80.40**	71.75	69.00	68.25
Electronics → Kitchen	80.90	83.90	85.50	88.60	**89.10**	82.25	82.25	79.00
Kitchen → Books	71.50	75.40	78.45	80.00	**81.25**	70.00	67.75	63.25
Kitchen → DVD	72.30	73.40	75.30	76.55	**78.25**	73.75	73.50	69.25
Kitchen → Electronics	78.10	80.85	82.90	84.85	**85.55**	82.00	82.00	81.25

Table 7
The accuracy between our proposed LIFA and the baselines on **Amazon Reviews Dataset**.

Source → Target	Recurrent CNN	BERT	XLM	Concatenation	LIFA-SIGMOID	GLU [41]	GTU [41]	GTRU [41]
Cell Phone → Clothing	83.88	87.61	88.85	89.87	**90.06**	85.13	84.95	84.80
Cell Phone → Home	88.03	89.17	92.17	92.33	**93.03**	84.85	84.20	84.55
Cell Phone → Tools	88.09	90.14	91.75	92.62	**93.13**	79.50	79.28	80.23
Clothing → Cell Phone	84.69	87.06	87.94	88.93	**89.04**	80.93	80.25	83.10
Clothing → Home	89.70	90.45	91.61	92.16	**92.25**	83.95	83.40	84.03
Clothing → Tools	88.65	88.50	90.47	90.78	**90.96**	79.48	77.85	79.38
Home → Cell Phone	86.89	86.89	90.79	90.72	**90.92**	83.18	81.85	82.10
Home → Clothing	85.87	89.43	91.64	91.94	**92.14**	82.75	84.10	85.43
Home → Tools	89.94	91.73	93.36	94.08	**94.19**	82.55	81.78	81.83
Tools → Cell Phone	84.84	88.24	89.36	89.97	**90.28**	82.13	80.81	81.83
Tools → Clothing	87.96	89.86	90.96	91.48	**91.79**	82.63	83.98	84.78
Tools → Home	87.54	91.33	91.4	92.52	**92.94**	84.70	83.95	85.28

to the knowledge from multiple experts. Lastly, our LIFA-SIGMOID achieves the best results, outperforms the Concatenation by efficiently sharing knowledge across experts.

4.6. Ablation Study

In this section, we conduct various ablation studies to demonstrate the robustness of LIFA under different embedding sizes and have a better understanding of the LIFA-COOP and LIFA-WTA behaviors.

4.6.1. Different Gating embedding sizes

Individual experts play a vital role in our LIFA framework. However, different pre-trained models can provide different embedding dimensions, i.e., 256, 768, and 768 in our Vietnamese datasets. Therefore, LIFA employs a linear layer to map such embeddings to the same dimensions before combining them. In this section, we investigate the effect of the standard mapping dimension on the final performance. We consider the "AIVIVN" and "AISIA-VN-Review-S" datasets with the LIFA-SIGMOID model and vary the standard mapping sizes, from 256, 512, to 768. The experimental results in Table 8 show that the increase of this dimension **does not consistently improve** the performance of LIFA-SIGMOID. On both datasets, the performance increases when the typical mapping size increases from 256 to 512 but decreases when we increase the mapping size to 768. One possible explanation is that the standard mapping size is the bottleneck to transfer the knowledge from the pre-trained models to the current task. Therefore, a small mapping size (256) limits the knowledge transfer to learn the current task. On the other hand, a larger mapping size allows for more knowledge, but not all are useful, especially with limited training data. As a result, controlling the typical mapping size in LIFA enables a flexible knowledge transfer mechanism to facilitate training across different datasets with different amounts of training data.

4.6.2. Cooperative and Competitive LIFA

In this section, we explore the effect of prior knowledge-sharing structure between experts and how it affects the final performance of LIFA. Our LIFA presents a simple yet effective way for users to enforce certain behaviors among experts via a single temperate hyper-parameter in the softmax gating layer: experts could cooperate together (LIFA-COOP, high temperature) or compete against one another (LIFA-WTA, low temperature). By increasing the temperature to infinity, LIFA-

Table 8
The experimental results of the proposed LIFA with different dimensions on the **AIVIVN Dataset** and **AISIA-VN-Review-S Dataset**.

Methods	AIVIVN Dataset			AISIA-VN-Review-S Dataset		
	AUC	ACC	F1	AUC	ACC	F1
LIFA 256	98.95	95.12	94.83	94.99	90.54	78.12
LIFA 512	**99.12**	**95.46**	**95.20**	**95.71**	**91.42**	**79.83**
LIFA 768	98.85	94.98	94.68	94.9	90.47	77.71

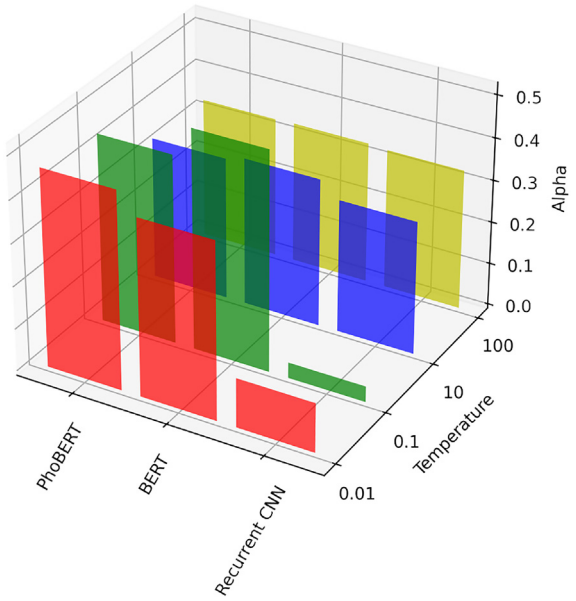

Fig. 2. The weight distribution learnt by our LIFA with different Softmax temperatures of 0.01, 0.1, 10 and 100 on **AIVIVN** Dataset.

COOP forces the experts to cooperate, and their predictions contribute equally to the final predictions. In contrast, LIFA-WTA lowers the temperature towards zero, which allocates all the weights to one expert who has the best performance. Since each expert's gradient is multiplied by its weight, experts making the wrong decision will have their weights lowered towards zero, which vanishes the gradient signals. As a result, LIFA-WTA will aggressively select a few experts who contribute the most to the correct predictions.

To verify such behaviours, we train several LIFA-COOP and LIFA-WTA models with different temperature values and report the weight distribution $[\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$ in Figs. 2 and 3. We observe that the weight distributions become smoother with higher temperature values on both datasets, which supports our hypothesis. The results show that it is more beneficial for the experts to cooperate rather than compete. From the weight distributions, one can see that LIFA-WTA essentially performs model selection to choose the best performing experts. As a result, LIFA-WTA does not encourage knowledge sharing among experts and thus, performs poorly compared to LIFA-COOP. This experiment's results shed light on the success of LIFA-SIGMOID and LIFA-COOP by showing that encourage cooperation is more beneficial than model selection for transfer learning.

5. Conclusion

In this work, we studied sentiment classification with a focus on the Vietnamese language. We explored the potentials and limitations of the existing approaches and showed that it is beneficial to take advantage of multiple pre-trained models for transfer learning. This observation motivated us to propose LIFA, an efficient framework to learn a unified embedding from several pre-trained models (experts) and perform better than its components. We further proposed a simple technique to enforce specific prior structures to such experts, resulting in two more LIFA variants that encouraged the experts to either cooperate or compete. Moreover, we also constructed the AISIA-VN-Review-F dataset, which is the first large-scale sentiment classification database for the Vietnamese language. Through extensive experiments on several benchmarks, we demonstrated the efficacy of LIFA compared to existing techniques and comprehensively studied the benefits and drawbacks

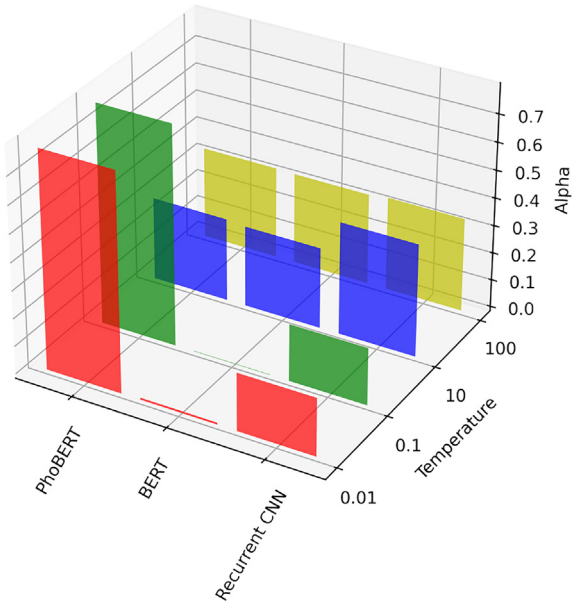

Fig. 3. The weight distribution learnt by our LIFA with different Softmax temperatures of 0.01, 0.1, 10 and 100 on **AISIA-VN-Review-S** Dataset.

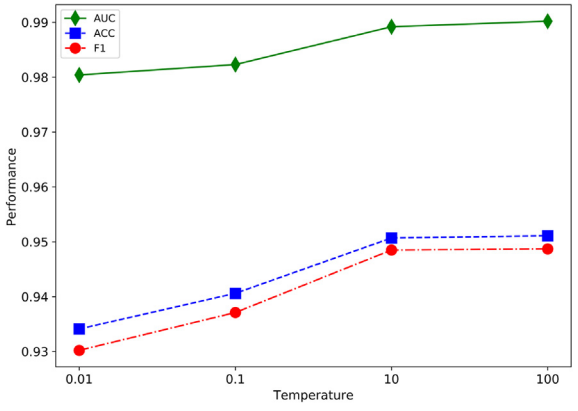

Fig. 4. The performance (AUC, ACC, F1) of our LIFA with different Softmax temperatures of 0.01, 0.1, 10, and 100 on **AIVIVN** dataset.

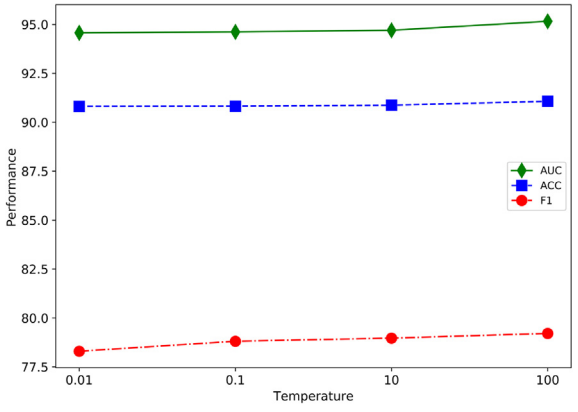

Fig. 5. The performance (AUC, ACC, F1) of our LIFA with different Softmax temperatures of 0.01, 0.1, 10, and 100 on **AISIA-VN-Review-S** dataset.

of its variants. We firmly believe our work will significantly contribute to the Vietnamese NLP research community. Finally, we will publish our codes and datasets used in this work upon acceptance. see Fig. 4,5.

CRediT authorship contribution statement

Cuong V. Nguyen: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Visualization. **Khiem H. Le:** Methodology, Investigation, Writing - original draft, Visualization. **Anh M. Tran:** Formal analysis, Investigation, Writing - original draft. **Quang H. Pham:** Methodology, Investigation, Writing - review & editing, Writing - original draft. **Binh T. Nguyen:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under Grant No. NCM2019-18-01. We want to thank the University of Science, Vietnam National University in Ho Chi Minh City, and AISIA Research Lab in Vietnam for supporting us throughout this paper.

References

[1] G. Wang, J. Sun, J. Ma, K. Xu, J. Gu, Sentiment classification: The contribution of ensemble learning, Decision Support Systems 57 (2014) 77–93, URL: http://www.sciencedirect.com/science/article/pii/S0167923613001978.
[2] F.G. Contratres, S.N. Alves-Souza, L.V.L. Filgueiras, L.S. DeSouza, Sentiment analysis of social network data for cold-start relief in recommender systems, in: Á. Rocha, H. Adeli, L.P. Reis, S. Costanzo (Eds.), Trends and Advances in Information Systems and Technologies, Springer International Publishing, Cham, 2018, pp. 122–132.
[3] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: A survey, Wiley Interdisciplinary Reviews, Data Mining and Knowledge Discovery 8 (2018) e1253.
[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, Curran Associates Inc., Red Hook, NY, USA, 2013, p. 3111–3119..
[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL:https://www.aclweb.org/anthology/N19-1423. doi: 10.18653/v1/N19-1423..
[6] D.Q. Nguyen, A.T. Nguyen, Phobert: Pre-trained language models for vietnamese, 2020. URL:https://arxiv.org/abs/2003.00744. arXiv:2003.00744..
[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019. URL:https://openai.com/blog/better-language-models/..
[8] M. Long, Y. Cao, J. Wang, M.I. Jordan, Learning transferable features with deep adaptation networks, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, p. 97–105..
[9] R. Gupta, L. Ratinov, Text categorization with knowledge transfer from heterogeneous data sources, in: Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2, AAAI'08, AAAI Press, 2008, p. 842–847..
[10] J. Lee, P. Sattigeri, G. Wornell, Learning new tricks from old dogs: Multi-source transfer learning from pre-trained networks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates Inc, 2019. URL:https://proceedings.neurips.cc/paper/2019/file/6048ff4e8cb07aa60b6777b6f7384d52-Paper.pdf..
[11] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton, Adaptive mixtures of local experts, Neural Computation 3 (1991) 79–87, https://doi.org/10.1162/neco.1991.3.1.79.
[12] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651 (2016)..
[13] A. Conneau, G. Lample, Cross-lingual language model pretraining, in: Advances in Neural Information Processing Systems, volume 32, Curran Associates Inc, 2019. URL:https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf..
[14] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, in: Advances in neural information processing systems, 2015, pp. 649–657..
[15] T.Z. Rie Johnson, Convolutional neural networks for text categorization: Shallow word-level vs. deep character-level, 2016. URL:https://arxiv.org/abs/1609.00718. arXiv:1609.00718..
[16] H.S., S.J., Long short-term memory, Neural Computing 9 (1997) 1735–1780. doi: https://doi.org/10.1162/neco.1997.9.8.1735..
[17] S. Lai, L.X. andKang Liu andJun Zhao, Recurrent convolutional neural networks for text classification, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015) 2267–2273..
[18] R. Johnson, T. Zhang, Deep pyramid convolutional neural networks for text categorization, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics 1, 2017, pp. 562–570, https://doi.org/10.18653/v1/P17-1052.
[19] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates Inc, 2019. URL:https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf..
[20] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, H.-W. Hon, Unified language model pre-training for natural language understanding and generation, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates Inc, 2019. URL:https://proceedings.neurips.cc/paper/2019/file/c20bb2d9a50d5ac1f713f8b34d9aac5a-Paper.pdf..

[21] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based model for Arabic language understanding, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, European Language Resource Association, Marseille, France, 2020, pp. 9–15. URL:https://www.aclweb.org/anthology/2020.osact-1.2..
[22] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, G. Hu, Pre-training with whole word masking for chinese bert, 2019. arXiv:1906.08101..
[23] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, M. Nissim, Bertje: A dutch bert model, 2019. arXiv:1912.09582..
[24] L. Martin, B. Muller, P.J. Ortiz Suarez, Y. Dupont, L. Romary, E. de la Clergerie, D. Seddah, B. Sagot, CamemBERT: a tasty French language model, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7203–7219, URL:https://www.aclweb.org/anthology/2020.acl-main.645.
[25] Y. Zhang, S. Roller, B.C. Wallace, MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1522–1527, URL:https://www.aclweb.org/anthology/N16-1178.
[26] W. Yin, H. Schütze, Learning word meta-embeddings, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1351–1360. URL:https://www.aclweb.org/anthology/P16-1128..
[27] X. Chen, A.H. Awadallah, H. Hassan, W. Wang, C. Cardie, Multi-source cross-lingual model transfer: Learning what to share, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3098–3112, https://doi.org/10.18653/v1/P19-1299, URL:https://www.aclweb.org/anthology/P19-1299.
[28] J. Ni, G. Dinu, R. Florian, Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection, CoRR abs/1707.02483 (2017). URL:http://arxiv.org/abs/1707.02483. arXiv:1707.02483..
[29] X. Li, H. Xiong, H. An, C. Xu, D. Dou, Xmixup: Efficient transfer learning with auxiliary samples by cross-domain mixup, CoRR abs/2007.10252 (2020). URL:https://arxiv.org/abs/2007.10252. arXiv:2007.10252..
[30] H. Guo, R. Pasunuru, M. Bansal, Multi - source domain adaptation for text classification via distancenet - bandits, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 7830–7838..
[31] Z. Li, Y. Zhang, Y. Wei, Y. Wu, Q. Yang, End-to-end adversarial memory network for cross-domain sentiment classification, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence IJCAI-17, 2017, pp. 2237–2243, https://doi.org/10.24963/ijcai.2017/311, URL: https://doi.org/10.24963/ijcai.2017/311.
[32] S. Mayhew, C.-T. Tsai, D. Roth, Cheap translation for cross-lingual named entity recognition, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2536–2545, https://doi.org/10.18653/v1/D17-1269, URL:https://www.aclweb.org/anthology/D17-1269.
[33] P. Keung, Y. Lu, V. Bhardwaj, Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER, 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. 1355–1360, https://doi.org/10.18653/v1/D19-1138, URL: https://www.aclweb.org/anthology/D19-1138.
[34] H.O. Hirschfeld, A connection between correlation and contingency, in: Mathematical Proceedings of the Cambridge Philosophical Society, volume 31, Cambridge University Press, 1935, pp. 520–524..
[35] M. Kale, A. Siddhant, S. Nag, R. Parik, M. Grabmair, A. Tomasic, Supervised contextual embeddings for transfer learning in natural language processing tasks, 2019. arXiv:1906.12039..
[36] Q.T. Nguyen, T.L. Nguyen, N.H. Luong, Q.H. Ngo, Fine-tuning bert for sentiment analysis of vietnamese reviews, in: 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), 2020, pp. 302–307, https://doi.org/10.1109/NICS51282.2020.9335899.
[37] B. Wang, Disconnected recurrent neural networks for text categorization, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 2311–2320.
[38] E. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)s, 2014, pp. 1532–1543, https://doi.org/10.3115/v1/D14-1162.
[39] J. Blitzer, M. Dredze, F. Pereira, Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 440–447. URL:https://www.aclweb.org/anthology/P07-1056..
[40] R. He, J. McAuley, Ups and downs, Proceedings of the 25th International Conference on World Wide Web - WWW '16 (2016). URL:https://doi.org/10.1145/2872427.2883037. doi: 10.1145/2872427.2883037..
[41] A. Madasu, V.A. Rao, Gated convolutional neural networks for domain adaptation, in: E. Métais, F. Meziane, S. Vadera, V. Sugumaran, M. Saraee (Eds.), Natural Language Processing and Information Systems, Springer International Publishing, Cham, 2019, pp. 118–130.
[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al, Pytorch: An imperative style, high-performance deep learning library, Advances in Neural Information Processing Systems (2019) 8024–8035.
[43] N. Dat Quoc, N. Dai Quoc, V. Thanh, D. Mark, J. Mark, A Fast and Accurate Vietnamese Word Segmenter, in: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), 2018, pp. 2582–2587.