applied sciences

MDPI

Article

AB-LaBSE: Uyghur Sentiment Analysis via the Pre-Training Model with BiLSTM

Yijie Pei [1,†], Siqi Chen [1,†], Zunwang Ke [2,*], Wushour Silamu [2] and Qinglang Guo [3]

[1] Xinjiang Multilingual Information Technology Laboratory, Xinjiang Multilingual Information Technology Research Center, School of Software, Xinjiang University, Urumqi 832001, China; poppy795@stu.xju.edu.cn (Y.P.); bubble777@stu.xju.edu.cn (S.C.)

[2] College of Information Science and Engineering, Xinjiang University, Urumqi 832001, China; wushour@xju.edu.cn

[3] National Engineering Research Center for Public Safety Risk Perception and Control by Big Data (RPP), China Academic of Electronics and Information Technology, Beijing 100041, China; gql1993@mail.ustc.cn

* Correspondence: kzwang@xju.edu.cn

† These authors contributed equally to this work.

Abstract: In recent years, more and more attention has been paid to text sentiment analysis, which has gradually become a research hotspot in information extraction, data mining, Natural Language Processing (NLP), and other fields. With the gradual popularization of the Internet, sentiment analysis of Uyghur texts has great research and application value in online public opinion. For low-resource languages, most state-of-the-art systems require tens of thousands of annotated sentences to get high performance. However, there is minimal annotated data available about Uyghur sentiment analysis tasks. There are also specificities in each task—differences in words and word order across languages make it a challenging problem. In this paper, we present an effective solution to providing a meaningful and easy-to-use feature extractor for sentiment analysis tasks: using the pre-trained language model with *BiLSTM* layer. Firstly, data augmentation is carried out by *AEDA* (An Easier Data Augmentation), and the augmented dataset is constructed to improve the performance of text classification tasks. Then, a pretraining model *LaBSE* is used to encode the input data. Then, *BiLSTM* is used to learn more context information. Finally, the validity of the model is verified via two categories datasets for sentiment analysis and five categories datasets for emotion analysis. We evaluated our approach on two datasets, which showed wonderful performance compared to some strong baselines. We close with an overview of the resources for sentiment analysis tasks and some of the open research questions. Therefore, we propose a combined deep learning and cross-language pretraining model for two low resource expectations.

Keywords: sentiment analysis; cross-lingual pre-trained language model; low-resource; *BiLSTM*; data augmentation

Citation: Pei, Y.; Chen, S.; Ke, Z.; Silamu, W.; Guo, Q. AB-LaBSE: Uyghur Sentiment Analysis via the Pre-Training Model with BiLSTM. *Appl. Sci.* **2022**, *12*, 1182. https://doi.org/10.3390/app12031182

Academic Editor: Valentino Santucci

Received: 23 November 2021
Accepted: 20 January 2022
Published: 24 January 2022

1. Introduction

With the rapid development of the Internet and the rise of communication platforms such as social media, online forums and e-commerce platforms, NLP technology plays a key role in the processing, understanding, and applications of text in the face of many unstructured text datasets generated on the Internet. Text sentiment analysis [1] refers to analyzing, processing, and extracting subjective texts with emotional color by using NLP and text mining techniques. Sentiment analysis is one of the important basic research tasks, which plays an important role in computer automatic processing and the understanding of natural languages.

In recent years, more and more research institutions and scholars have paid attention to sentiment analysis. In SIGIR, ACL, WWW, CIKM, WSDM, and other famous international conferences, research results from this issue emerge one after another. Traditional text sentiment analysis methods mainly include sentiment lexis-based and supervised machine learning based methods, which have good performance in processing small-data text sentiment analysis tasks and have higher interpretability. However, traditional methods mainly have a problem with the sparsity of text representation and weak generalization ability.

Felipe [2] proposed a method to train incremental word sentiment classifiers from time-varying distributed word vectors to automatically extract constantly updated sentiment words from a Twitter stream, resulting in a time-varying sentiment dictionary based on incremental word vectors. However, this method relies too much on the creation of an emotion dictionary. There is no context to this method, and the portability of the emotion dictionary is poor.

Ahmad [3] proposed an optimized sentiment analysis framework OSAF based on SVM, which uses SVM grid searches technology and 10K cross verification. Mathapati [4] proposed a sentiment analysis method based on emoticons and discussed the role of emoticons in sentiment analysis. Compared with the construction of emotion dictionary, the emotion classification method based on machine learning has some progress, but it still needs to mark the text features manually. Secondly, machine learning relies on a large amount of data, and such methods often cannot fully use contextual information in sentiment analysis, which affects the accuracy of the results.

Deep learning is the application of a multi-layer neural network of learning, which solves a lot of problems that are difficult to be solved by machine learning in the past. At present, deep learning models include CNN, *RNN*, *LSTM* [5], Transformer, *BiLSTM* [6], GRU [7], and attention mechanism. Rehman [8] proposed a hybrid model using *LSTM* and deep CNN models, which also used dropout technology, normalization, and correction linear elements to improve accuracy.

Xu [9] proposed the *DomBERT* model by combining ELMo [10] and *BERT* [11], which showed excellent performance in aspect based sentiment analysis. As an application of transfer learning, a pretraining model can transfer knowledge learned from the open domain to downstream task to improve a low-resource task, which is also very beneficial for low-resource language processing.

Wu [12] proposes two variants of context-guided BERT (CG-Bert) that learn to allocate Attention in different contexts. This modified Quasi-attention CG-Bert model can learn combinatorial Attention that supports subtractive Attention. Mao [13] proposed a complete solution for ABSA and constructed two machine reading comprehension (MRC) problems and solved all subtasks by jointly training two bert-MRC models with shared parameters. Li [14] proposed a new direction-based sentiment analysis method, GBCN, which uses a gating mechanism with an up-down file aspect embedding to enhance and control the BERT representation of aspect-oriented sentiment analysis.

For the sentiment analysis model based on a deep neural network, excellent results have been achieved in many sentiment analysis tasks, but its success depends heavily on large-scale training data with tags. For some widely used languages, the acquisition of manually tagged data may be relatively easy, so deep learning-based models succeed in sentiment analysis in these languages.

In the research community, using high-quality manual tagging, we have proposed sentiment analysis task training data in some languages with a large numbers of speakers, such as English. Using these high-quality annotation data, based on the cross-language transfer method, we can build sentiment analysis models for the Uyghur language with little or no annotation data. Therefore, although deep learning has gained preliminary application in NLP, it is still rarely used in the Uyghur language. This paper uses a pre-training model to automatically learn Uyghur language features and explores its feasibility in Uyghur sentiment analysis.

What causes the unique characteristics of Uyghur texts? There are a lot of parallel words in the Uyghur text, so there are ambiguities [15]. The reason for this phenomenon is that it belongs to the Turkic language family of the Altai language family, which is an agglutinative language with complex morphological changes. The modern Uyghur language is based on the Arabic alphabet and has similar pronunciation and borrows many Arabic words, which are formed by combining affixes and stems. For example, Uyghur names lack a unified writing style, with some names being written in multiple ways. Therefore, its adhesion leads to data bias and results in data sparsity. This creates many unregistered people, institutions, and place names, leading to sparse data. Therefore, new ideas and methods are needed to further improve the accuracy of Uyghur sentiment analysis.

In this paper, the pre-training model was used to share the vocabulary layer, and they implemented the data augmentation strategy on the dataset. Thus, the generalization ability of the model was improved effectively. As depicted in Figure 1, we propose a low-resource language model with the *BiLSTM* layer that can address these issues: *AB-LaBSE*. First, we propose data augmentation by *AEDA* [16] and using the shared vocabulary characteristics of the pre-training model *LaBSE* [17] to fine-tune the cross-lingual pre-training model, which is combined with deep learning *BiLSTM* model.

We used *LaBSE* to pre-train a language model on a large cross-lingual corpus.We introduce an *LSTM* (Long- and short-term memory neural network) that selects relevant semantic and syntactic information from the pre-trained low-resource language model. To evaluate our model, we collected and annotated two datasets for Uyghur sentiment analysis. Respectively, these are hotel review dichotomies and five categories of emotion datasets. The experimental results show that *AB-LaBSE* can significantly improve the performance with a few labeled examples.

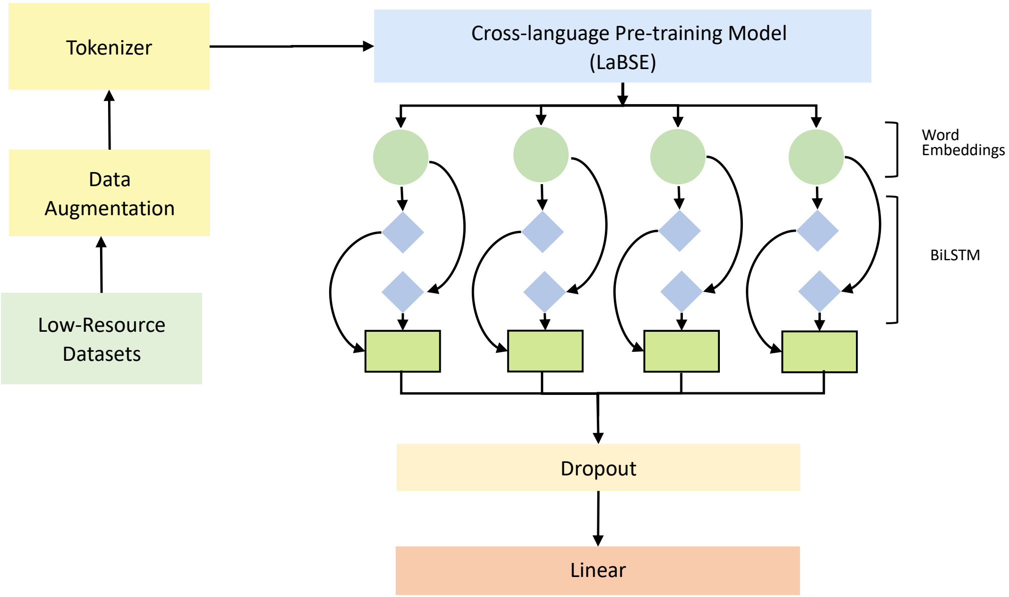


Figure 1. Low resource cross-language pretraining Model with *BiLSTM-AB-LaBSE*.

As a combined method, our contributions are as follows:

1. We constructed two sentiment analysis datasets, one of which was about two categories of hotel review sentiment analysis, and the other was a dataset containing five categories of emotion analysis, including happiness, surprise, sadness, anger, and neutral. We also completed the annotation of Uyghur datasets with the help of Uyghur language experts. They divided each of the language datasets for 80% training, 10% validation, and 10% testing.
2. We propose a fine-tuning strategy for Uyghur agglutinative languages—a data augmentation method, based on the feature that a cross-language pre-training model shares a vocabulary layer. They base it on the pre-training model for the Uyghur language, and at present, there is little research on the Uyghur language using a pre-training model.

3. We propose a method to add *BiLSTM* layers, in which our datasets from outputs that have been pre-trained across languages are associated with *BiLSTM* layers for better learning context features. In this task, the method can better select relevant semantic and feature information from the pre-trained language model. This method can improve the performance of downstream tasks effectively by taking advantage of the characteristics of context association of cohesive languages.

2. Related Works

Sentiment analysis has a wide range of applications, which are an important task in NLP. We summarized the related works on three topics: (1) data augmentation; (2) cross-lingual pre-trained language models; and (3) *BiLSTM*.

Data augmentation. Data augmentation is a class of methods used for synthesizing new data from existing data. It has played an important role in the experiment. We find that the larger the scale and the higher the quality, the better the generalization ability of the model. Data augmentation solves the problem of data accuracy by using original documents to generate similar examples, solving the problem of insufficient data effectively. It is an effective method used to expand the sizes of the data samples. In recent years, more and more people are creating data augmentation techniques in research. According to the diversity of generated samples, methods of data augmentation are divided into the following three categories: (1) Paraphrasing; (2) Noising; (3) Sampling.

Paraphrasing is used to make some changes to the words, phrases, and sentence structure in the sentence, keeping the original meaning. This method can make use of dictionaries, knowledge maps, semantic vectors, *BERT* [11] models, rules, Machine Translation, etc. It can randomly replace non-stop words with synonyms. For example, back translation [18] can change the syntactic structure and retain semantic information, and it can often increase the diversity of textual data. However, the data generated by the back-translation method depends on the quality of the translation, and most of the translation results are not very accurate.

Noising is used to increase some discrete or continuous noises while keeping the label unchanged, which has little impact on semantics. People are immune to noise when reading text, such as words out of order and typos [19]. Based on this idea, some noise can be added to the data to improve the robustness of the model. Since there are few synonyms in Uyghur language, replacing synonyms does not produce positive meaning. The method of data augmentation with noise is simple to use, but it will affect the sentence structure and semantics, with limited diversity, and mainly improve the robustness [20].

Sampling aims to select new samples based on the current data distribution, which will generate more diverse data. Sampling aims to select new samples based on the current data distribution, which will generate more diverse data [21]. Sampling refers to sampling new samples from the data distribution. Different from general paraphrasing, sampling is more task-dependent and needs to increase more diversity while ensuring data reliability. Rules, Seq2Seq Models, Language Models, and Self-training are usually used in experiments to increase data diversity.

Above all, the method of data augmentation can be used as a powerful tool to solve the problems of data imbalance and missing data quickly when we train the NLP model.

Research on sentiment analysis in machine learning. Firstly, we will analyze the research of sentiment analysis function in machine learning. Kim [22] proposed the use of a convolutional neural network, CNN, to complete sentiment analysis and problem classification, which achieved good results. Li [23] proposed to use undersampling and dynamically generating random subspace strategies to solve the problem of unbalanced datasets in sentiment analysis.

After the above analysis, we can conclude that it proved these methods to be simple and effective in sentiment analysis task, but there are some limitations: Without a large amount of manual annotation data, traditional machine learning methods will not perform well. The method, based on a shallow neural network, relies on grammatical features and word order features and has difficulties grasping the deep semantic information. We have paid little attention to the low-resource Uyghur language.

Convolutional neural network in deep learning has achieved good results, but it did not consider the potential theme of the text. Dwivedi [24] proposed a rule model based on the Restricted RBM (Restred Bolzmann Machine) to analyze the sentiment analysis of sentences. Can [25] proposes a limited data model based on the *RNN* framework and a language with the largest dataset, and applies it to languages with limited resources, which has a better effect on the sentiment analysis of small languages. Based on *LSTM*, Wang [26] proposed a memory network classification of long-term and short-term aspect-oriented emotion based on attention. Chen [27] proposed a new sentiment analysis scheme based on the data of Twitter and Weibo, embedding the attention-based short and long-term memory network to analyze the emotional factors of facial expressions, and training a sentiment classifier. Sangeetha [28] proposed *LSTM* with multi-layer fusion to process sentiment analysis. This method uses a multi-layer attention mechanism to process sentence input sequence in parallel and uses different pruning ratios to improve accuracy. Then, multi-layer information is fused, and it fed the results to the *LSTM* layer as the input.

The Cross-lingual Pre-trained Language Model. In recent years, studies have shown that the generative training of natural language understanding proposed in monolingual languages such as English is very effective. Therefore, more and more people focus on the cross-language field with low resources. The researchers start from the *ELMo* [10] model, which uses BiLM (bidirectional language model) to pre-train the vector representation of words, which can dynamically generate the vector representation of words according to the training set. Sun [29] proposed an aspect based sentiment analysis, fine-tuning *BERT*'s pre-training model and getting new and good results on the SentiHood and SemEVAL-2014 Task 4 datasets. Yin [30] proposed Senci-Bert (Senti-bert), a variant of *BERT*, in which SentiBERT's method has more advantages than the baseline in capturing negative and contrastive relationships and constructing combinatorial models.

3. Methodology

In this section, we will explain our approach. We divide our training into four stages. We first performed data enhancement on the unbalanced characteristics of our sentiment analysis dataset. Secondly, the language model is pre-trained on a large-scale cross-language text corpus. In addition, the *BiLSTM* layer is added to the output results of the pre-trained model. Finally, we use the dropout layer and a full connection to get the final sentiment analysis prediction.

3.1. Data Augmentation

In data augmentation, because the application of machine learning for textual research is still a highly active section, especially with the small amount of initial annotation data in research experiments, we need data augmentation to improve the data diversity of experimental datasets. After we use data enhancement in our model research, the number of datasets is improved, noise is added to the experimental data, the generalization ability of the model is improved, and the robustness of the model is improved. We used the latest *AEDA* [13] technology, which is a data enhancement technique that is easier to implement than the *EDA* [31] approach, which we compared with our results. In addition, only the insertion of punctuation marks for the original data sequence information modification is not obvious. It also preserves the order of the words while maintaining their position in the sentence lead, creating performance. In addition, deletion in *EDA* can lead to information loss and thus mislead the network, whereas *AEDA* preserves all input information.

In this paper, based on the cross-language characteristics of the *LaBSE* [14] model, we use *AEDA* to expand the training data and to improve the ability of the model, aiming at the characteristics of Uyghur language with low resource stickiness. The word stem, through which words are made and connected by multiple suffixes, represented the content of this language. We first select the position where we want to insert the symbol. In order to ensure the correct insertion mark position, there are inserts, but not too many inserts to cause noise. Too much noise will have a negative effect on the model. Insertion positions in the sequence are determined by randomly specifying numbers between 1 and 1/3 of the sequence. To summarize, we use the shared lexical layer of the pre-trained model for each selected location, from “.”, “;”, “?”, “:”, “!”, “,“, and one punctuation mark is chosen at random. We show the specific augmentation in Section 4.1.

3.2. Cross-Language Model Pre-Training

Researchers at Google have developed a multilingual *BERT* Embedding model called language-agnostic *BERT* Sentence Embedding(*LaBSE*), which generates language-agnostic cross-language Sentence Embedding for 109 languages on a single model. In simple terms, *LaBSE* combines MLM and TLM pre-training on a 12-layer Transformer that contains 500,000 tokens with translation sorting tasks performed using bidirectional biencoders. In this paper, we utilized *LaBSE* to model the conditional probability. *LaBSE* uses the parallel text or dual text to process all languages through byte pair encoding (BPE) [32]. LaBSE's research shows that this model is effective even in low-resource languages where no data is available during training. The model of the *LaBSE* is shown in Figure 2.

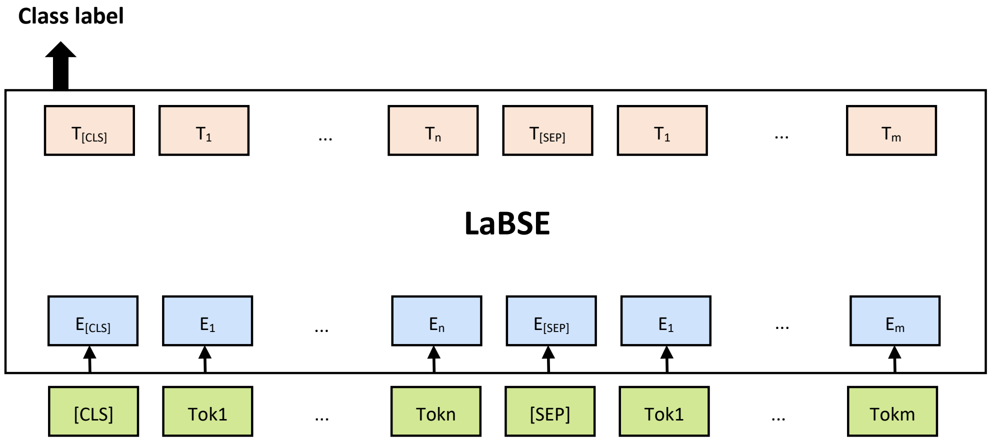


Figure 2. The model of the *LaBSE*.

Shared contents include the same alphabet or the same anchor tokens, such as digits or proper nouns. This method of sharing dictionaries can improve the alignment of languages in the embedded space. We learn the BPE splits on the concatenation of sentences sampled randomly from the monolingual corpora. We sample sentences according to a multinomial distribution with probabilities. Additionally, sentences are sampled according to a probable multinomial distribution $\{q_i\}_{i=1,2,3...n}$, where:

LaBSE adopted the strategy of in-batch negative sampling by training bidirectional dual Encoders with additive margin softmax loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\frac{e^{\phi(x_i,y_i)-m}}{e^{\phi(x_i,y_i)-m}+\sum_{n=1,n\neq i}^{N}e^{\phi(x_i,y_n)}} \tag{1}$$

The embedding space similarity of x and y is given by $\phi(x, y)$. In addition, $\phi(x, y) =$ cosine(x, y), the loss attempts to rank y_i, the true translation of x_i, over all $N - 1$ alternatives in the same batch even when $\phi(x_i, y_i)$ is discounted by margin m.

Notice that $\mathcal{L}$ is symmetric and depends on whether the softmax is over the source or the target. To bi-directional ranking, the final loss function sums the source to target, $\mathcal{L}$, and target to source, $\mathcal{L}'$, losses:

$$\overline{\mathcal{L}} = \mathcal{L} + \mathcal{L}' \tag{2}$$

LaBSE adapted multilingual *BERT* and combined masked language model (MLM) and translation language model (TLM) pretraining with a translation ranking task using bi-directional dual encoders. This distributed sampling method prevents words in low-resource languages from being split at the character level. In particular, it increases the number of tokens associated with low-resource languages and eases the bias toward high-resource languages.

In the pre-training and parameter sharing stage of the model, for an L layer transformer encoder, *LaBSE* used a three-stage progressive stacking algorithm to train, which first learns a $\frac{L}{4}$ layers model, then $\frac{L}{2}$ layers, and finally all L layers. The parameters of the model will be copied to the subsequent training through the learning operation in the early stage.

3.3. BiLSTM

The *BiLSTM* model was proposed by Grave by combining the bidirectional *RNN* models *BRNN* and *LSTM* [5] unit and verified its effectiveness in speech recognition tasks. The core idea of the *BiLSTM* [6] model is that the current input is related to the sequence before and after, and the realization is to input the data sequence into the model from two directions, respectively, and save the data information in two directions, namely the historical information and the future information, through the hidden layer. The recurrent neural network *RNN* is good at capturing long dependent sequence relations. Some outputs of neurons can be transmitted to neurons again as inputs, and the previous information can be effectively utilized. However, in the process of training, the constant multiplication of the function derivative will lead to the problem of "gradient disappearance" and "gradient explosion".

LSTM cleverly introduces a "gate" mechanism to solve this problem. *LSTM* is composed of a series of repeated timing modules, each of which contains three gates and a memory cell, namely, the forgetting gate, input gate, and output gate. The forgetting gate determines what information the cell will discard, reads h_{t-1} and x_t, and outputs a value between 0 and 1 to each in cell state C_{t-1}.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3}$$

The input gate determines what information is stored in the cell state, and there are two parts to this. First, a sigmoid neural network layer determines what values will be updated, called the "input gate layer." Then, a tanh layer creates a new candidate vector C_t, which is added to the state.

$$\begin{gathered} i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \end{gathered} \tag{4}$$

When the unit information is updated, the old state is multiplied by f_t, irrelevant information is discarded, and $i_t * \tilde{C}_t$ is added to form A new candidate value.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$

The output gate determines which part of the cell state will be output by running a sigmoid layer, and then processes the cell state through the tanh function to get a value between -1 and 1, which is multiplied by the output of the sigmoid gate, and finally outputs only part of the determined output.

$$\begin{gathered} O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \\ h_t = o_t * \tanh(C_t) \end{gathered} \tag{6}$$

where: $\tanh()$ represents the activation function. σ represents the neural network layer of sigmoid. x_t is the unit state input at time of t. f_t, i_t and O_t are the settlement results of forgetting gate, input gate and output gate, respectively. W_f, W_i, W_o and W_c respectively represent the weight of forgotten gate, input gate, output gate, and the updated weight. b_f, b_i, b_o and b_c are the corresponding offset quantities.

In the process of text classification, *BiLSTM* is used to make full use of the context information of text, which is to combine the two *LSTM* models with opposite sequences. That is, add a layer of reverse *LSTM* layer to the original forward *LSTM* network layer and combine them, so the output can be expressed as:

$$H_t = h_l \oplus h_r \tag{7}$$

H_t for *BiLSTM* model outputs the text feature vector. Where h_l is the output of the forward *LSTM* and h_r is the output of the reverse *LSTM*.

In this experiment, we added a dropout layer to prevent the model from overfitting. At each iteration, neurons are randomly output to zero at the specified ratio. In the experiment, the output of *BiLSTM* is denoted as H_T, and W_{hy} stands for the ratio of the dropout layer.

$$y = W_{hy} H_T \tag{8}$$

Finally, the linear layer performs a linear transformation of the output of dropout, where A is the weights and b is the bias.

$$z = yA^T + b \tag{9}$$

4. Experiments

4.1. Datasets

For our experiments, we made use of a low-resource language dataset—Uyghur. The experiment of a corpus containing Uyghur sentiment analysis from public comments on the hotel's website (https://www.dianping.com/) (accessed on 23 November 2021) grab the comment text (datasets) and a crawl Tianshan net and Xinhua Uyghur channel sentiment analysis of the dataset, under the guidance of Uyghur experts, to complete the corpus tagging work. The hotel evaluation data includes 2011 data of positive and negative emotions, and the emotional analysis includes five emotions, such as happy, surprised, sad, angry, and neutral, with 10,000 data. We divided our data into 80% training, 10% validation, and 10% test sets.

The specific proportion of data is shown in Figures 3 and 4.

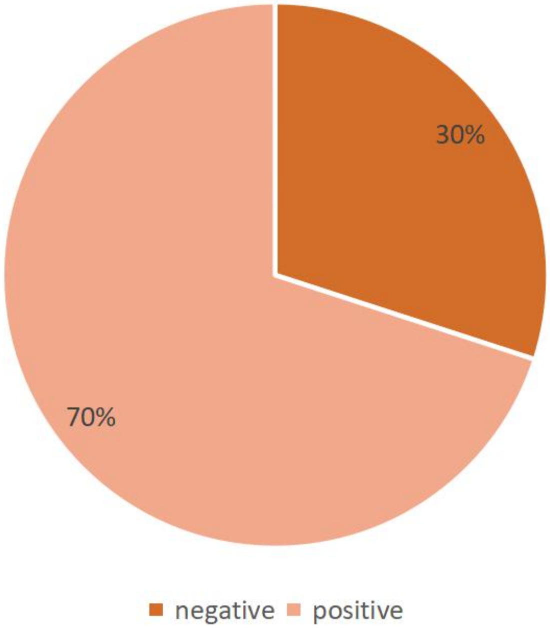


Figure 3. The specific proportion of the two datasets in binary datasets.

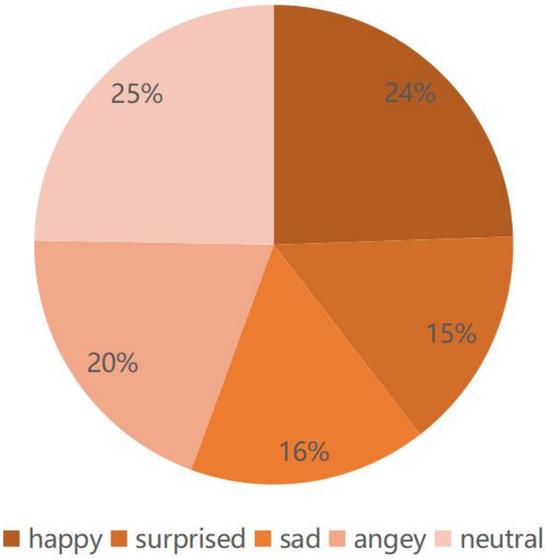


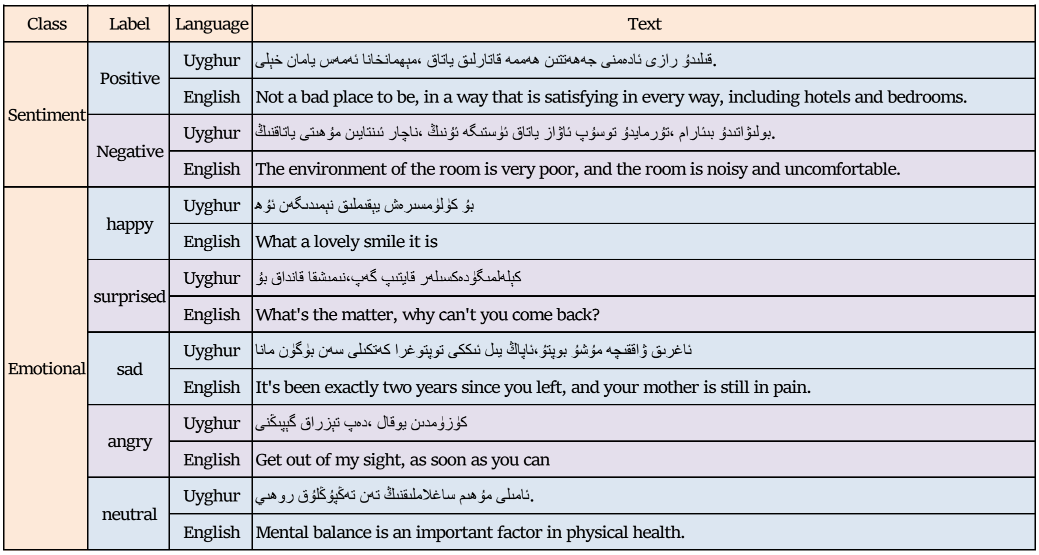
Figure 4. The specific proportion of the two datasets in five category datasets.

In this experiment, the data size of the training set before and after data augmentation is shown in Table 1.

Table 1. The data size of the training set before and after data augmentation.

The Dataset	Before Data Augmentation	After Data Augmentation
Binary	7999	12,833
Five categories	1608	4824

The details of the two datasets are shown in the Figure 5. We illustrated with detailed examples of data.



Class	Label	Language	Text
Sentiment	Positive	Uyghur	قىلىدۇ رازى ئادەمنى جەھەتتىن ھەممە قاتارلىق ياتاق ،مېھمانخانا ئەمەس يامان خېلى.
		English	Not a bad place to be, in a way that is satisfying in every way, including hotels and bedrooms.
	Negative	Uyghur	بولىۋاتىدۇ بىئارام ،تۇرمايدۇ توسۇپ ئاۋاز ياتاق ئۇستىگە ئۇنىڭ ،ناچار ئىنتايىن مۇھىتى ياتاقنىڭ.
		English	The environment of the room is very poor, and the room is noisy and uncomfortable.
Emotional	happy	Uyghur	بۇ كۈلۈمسىرەش يېقىملىق نېمىدىگەن ئۇھ
		English	What a lovely smile it is
	surprised	Uyghur	كېلەلمىگۈدەكسىلەر قايتىپ گەپ،نىمىشقا قانداق بۇ
		English	What's the matter, why can't you come back?
	sad	Uyghur	ئاغرىق ۋاققىچە مۇشۇ بوپتۇ،ئاپاڭ يىل ئىككى توپتوغرا كەتكىلى سەن بۈگۈن مانا
		English	It's been exactly two years since you left, and your mother is still in pain.
	angry	Uyghur	كۆزۈمدىن يوقال ،دەپ تېزراق گېپىڭنى
		English	Get out of my sight, as soon as you can
	neutral	Uyghur	ئامىلى مۇھىم ساغلاملىقنىڭ تەن تەڭپۇڭلۇق روھىي.
		English	Mental balance is an important factor in physical health.

Figure 5. Examples of two dataset annotation methods in Uyghur and English sentiment analysis.

4.2. Baseline Models

In recent years, research on sentiment analysis has become more and more popular. We analyzed traditional machine learning methods and found them unsatisfactory. Therefore, we mainly used a cross-language pre-training model to compare the effects of this experiment.

Before us, only a few people have done Uyghur sentiment analysis on cross-language pre-training model. We compared our *AB-LaBSE* model with several baseline models in different categories, including *mBERT* [33], *Sentence-bert* [34], *XLM-R* [35] and *XLM-align* [36], each of which is described below.

mBERT: *Multilingual BERT* is a transformer model pretrained on a large corpus of multilingual data in a self-supervised fashion, *mBERT* follows the same model architecture and training procedure as *BERT*, except that it is pre-trained on concatenated Wikipedia data of 104 languages. For tokenization, *mBERT* uses WordPiece embeddings with a 110,000-word shared vocabulary to facilitate embedding space alignment across different languages. This means it was pretrained on raw texts only, with no humans labeling them in any way (which is why it can use lots of publicly available data), instead of using an automatic process to generate inputs and labels from those texts.

Sentence-BERT: *Sentence-BERT* can be used to calculate sentence/text embedding for over 100 languages and is a deep learning model for state-of-the-art sentence, text, and image embedding. These embeddings can be compared with cosine similarity to find sentences with similar meanings, which can be used for semantic text similarity, semantic search, or free translation mining. The model makes sense for tasks such as clustering or semantic search, and is based on PyTorch and Transformer, providing a large collection of pre-trained models for a variety of tasks.

XLM-R: The *XLM-R* uses filtered common-crawled data (more than 2 TB) to demonstrate that using a large-scale multi-language pretraining model can significantly improve the performance of cross-language migration tasks. The *XLM-R* uses the XLM [37] model in combination with Robert to significantly improve the performance of the pre-trained model.

XLM-align: The *XLM-ALIGN* cross-language model uses denoising word alignment as a new cross-language pre-training task, which improves cross-language portability across different datasets, especially for top-level tasks such as question answering and structured prediction.

4.3. Experiment Setting

In order to learn more about the Uyghur language, we use the *LaBSE* model, which uses 17 billion monolingual sentences and 6 billion pairs of bilingual sentences to train, using cosine distance to find the nearest neighbor translation for a given sentence. We ran these typical models on a 4386Tesla K80 gpu and built the experimental model using pytorch version 1.9.0. The parameters set in the experiment are shown in Table 2.

Table 2. The hyperparameters of *AB-LaBSE*.

Hyperparameters	**Uyghur**
Max sequence length	64
Max epochs	10
Batch size	32
Learning rate	2×10^{-5}
Gradient accumulation steps	1
Warm up proportion	0.0
Dropout	0.5
The dimension of the word vector	768
The number of hidden neurons in the BiLSTM layer	384

4.4. Results and Analysis

In this experiment, we used the “strict” standard to evaluate the model’s results. In this experiment, P (precise rate), R (recall rate), and F (F1-Measure) value were used as evaluation indexes to evaluate the effectiveness of the experimental method. P refers to the ratio of the number of correctly classified comment texts to the total number of comment texts. R refers to the ratio of the number of correctly classified texts belonging to a certain emotional orientation to the number of truly emotional orientation comments in the review texts. F value is the harmonic average of precise and recall rate.

In this section, by comparing the accuracy and recall rates of these models, we can prove that the *AB-LaBSE* model used in this paper has the best effect. As can be seen from Tables 3 and 4, the *AB-LaBSE* model proposed by us has significantly improved in the above three indicators, reaching the optimal value in the comparison model, and its F1 value is also the highest, so the performance of the model is the most stable. By comparing the *mBERT* model and *Sentence-BERT* model, the *mBERT* model does not add Uyghur language data for training, so our five categorical emotion analysis dataset has a poor effect on this model. However, with the increase of data volume, The *mBERT* model performs much better than the *Sentence-BERT* model on the binary dataset of hotel reviews. It uses WordPiece embedding to promote the spatial alignment of embedding across languages. The *BERT* model maps sentences to a vector space. This vector space is not suitable for cosine similarity, however, the *Sentence-BERT* model overcomes this shortcoming, so in the case of a relatively small dataset, the effect of using the *Sentence-BERT* pre-training model is better than *mBERT*.

Table 3. Results of different algorithms for two classification datasets of sentiment analysis.

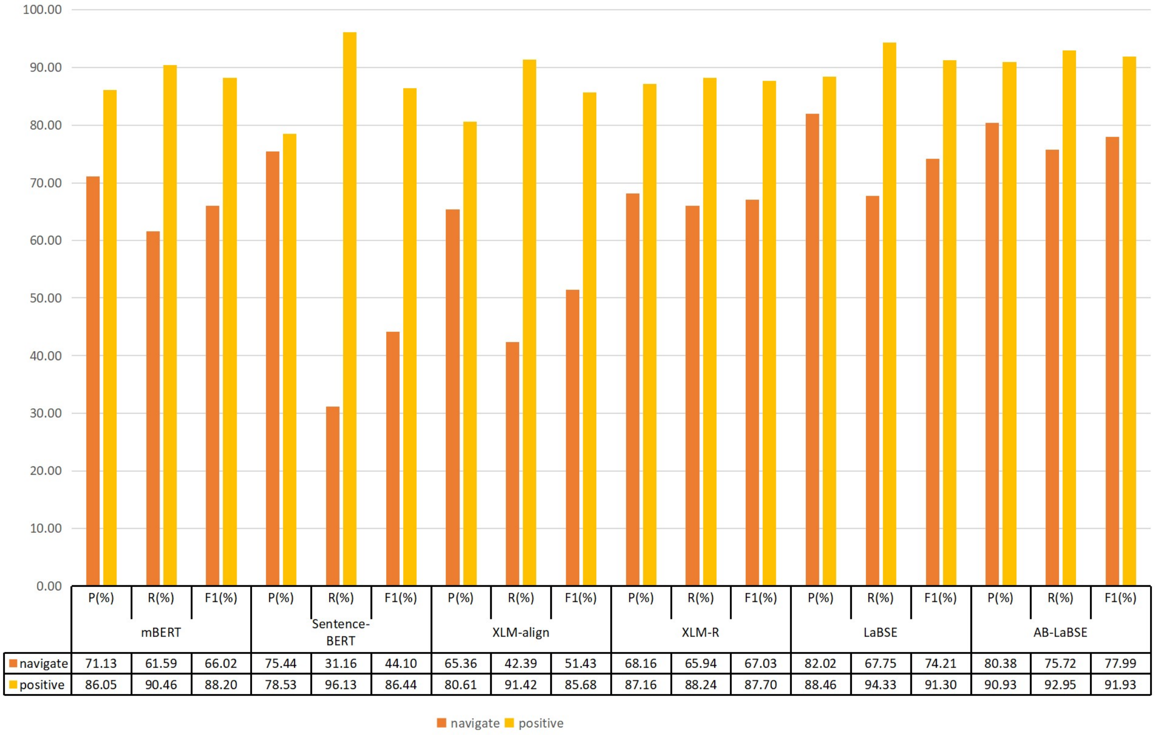
Model	P(%)	R(%)	F1(%)
mBERT	78.59	76.03	77.11
Sentence-BERT	76.98	63.64	65.27
XLM-align	72.99	66.91	68.55
XLM-R	77.66	77.09	77.37
AB-LaBSE	**85.66**	**84.34**	**84.96**

Table 4. Results of different algorithms in five categorical datasets of emotion analysis.

Model	P(%)	R(%)	F1(%)
mBERT	48.14	48.13	46.57
Sentence-BERT	51.47	53.12	51.04
XLM-align	61.13	58.8	58.34
XLM-R	62.08	56.93	58.00
AB-LaBSE	**76.78**	**77.19**	**76.04**

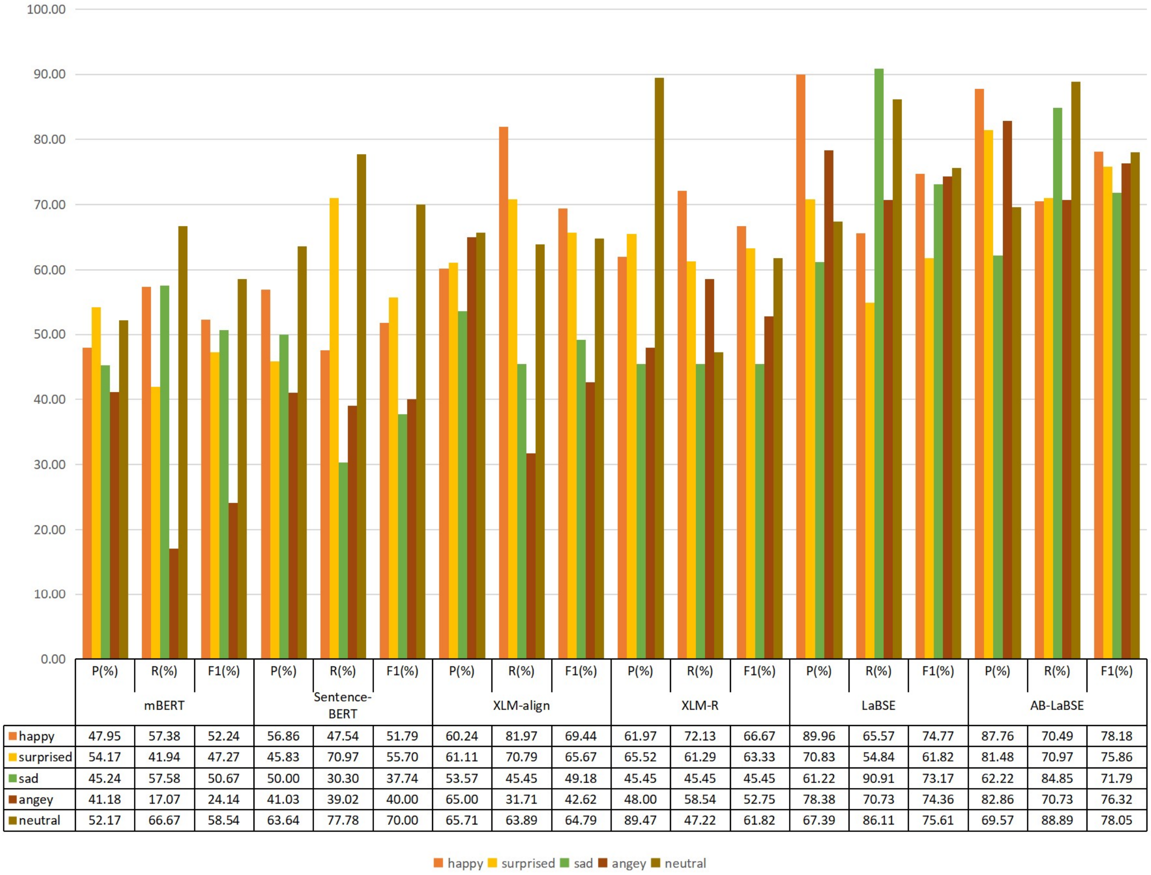
By comparing the data in Figures 6 and 7, we can see that the recognition effect of *XLM* class model is generally better than that of the *BERT* class model. A large part of the reason is that the *XLM* class model now uses a large shared statement block model to mark strings and adds the Uyghur language dataset to the model. *XLM-aligin* also introduces denoising Word alignment before cross-language training, which is better than the above two models. It can also be seen from the experimental results that the F1 of *XLM-R* is higher than that of *XLM-aligin*. In the two-category experiment, the F1 value of negative increased by 15.6%, and the F1 value of positive increased by 2.02%. This is because *XLM-R* combines *XLM*'s cross-language and Roberta's improved optimization function and larger model parameter number, so as to improve the accuracy of recognition.

Its language processing for low-resource languages is also very good. We can see that the pre-training model has achieved the best results in the downstream tasks, and the F1 value has improved a lot.



	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
	mBERT			Sentence-BERT			XLM-align			XLM-R			LaBSE			AB-LaBSE		
navigate	71.13	61.59	66.02	75.44	31.16	44.10	65.36	42.39	51.43	68.16	65.94	67.03	82.02	67.75	74.21	80.38	75.72	77.99
positive	86.05	90.46	88.20	78.53	96.13	86.44	80.61	91.42	85.68	87.16	88.24	87.70	88.46	94.33	91.30	90.93	92.95	91.93

Figure 6. The performance of different models on the hotel review binary datasets.



	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
	mBERT			Sentence-BERT			XLM-align			XLM-R			LaBSE			AB-LaBSE		
happy	47.95	57.38	52.24	56.86	47.54	51.79	60.24	81.97	69.44	61.97	72.13	66.67	89.96	65.57	74.77	87.76	70.49	78.18
surprised	54.17	41.94	47.27	45.83	70.97	55.70	61.11	70.79	65.67	65.52	61.29	63.33	70.83	54.84	61.82	81.48	70.97	75.86
sad	45.24	57.58	50.67	50.00	30.30	37.74	53.57	45.45	49.18	45.45	45.45	45.45	61.22	90.91	73.17	62.22	84.85	71.79
angey	41.18	17.07	24.14	41.03	39.02	40.00	65.00	31.71	42.62	48.00	58.54	52.75	78.38	70.73	74.36	82.86	70.73	76.32
neutral	52.17	66.67	58.54	63.64	77.78	70.00	65.71	63.89	64.79	89.47	47.22	61.82	67.39	86.11	75.61	69.57	88.89	78.05

Figure 7. The performance of different models on the five categorical datasets.

In order to further verify the influence of *BiLSTM* on the effect of dichotomous datasets and five categorical datasets, two versions are used for comparison, as shown in Tables 5 and 6, showing the scores of *LaBSE* algorithm in the two datasets. In the sentiment analysis of hotel review dichotomous datasets, with the addition of *BiLSTM* layer, the overall effect is getting better and better. The results showed that P, R, and F1 values were 76.78%, 77.19%, and 76.04%, respectively. In the emotion analysis of five categorical datasets, with the addition of *BiLSTM* layer, the overall effect is getting better and better. The results showed that P, R, and F1 values were 85.66%, 84.34%, and 84.96%, respectively.

Table 5. Results of hotel reviews binary datasets for sentiment analysis.

Model	P(%)	R(%)	F1(%)
LSTM	83.54	82.85	83.18
BiLSTM	85.66	84.34	84.96

Table 6. Results of five categorical datasets for emotion analysis.

Model	P(%)	R(%)	F1(%)
LSTM	74.82	76.05	74.70
BiLSTM	76.78	77.19	76.04

4.5. Ablation Study

In order to evaluate the role of key factors in the used methods, we used the *LaBSE* pre-training model to train and test the hotel review binary and emotion analysis of five categories of Uyghur datasets for methods with and without data enhancement and methods with and without the *BilSTM* layer, as shown in Figures 8 and 9.

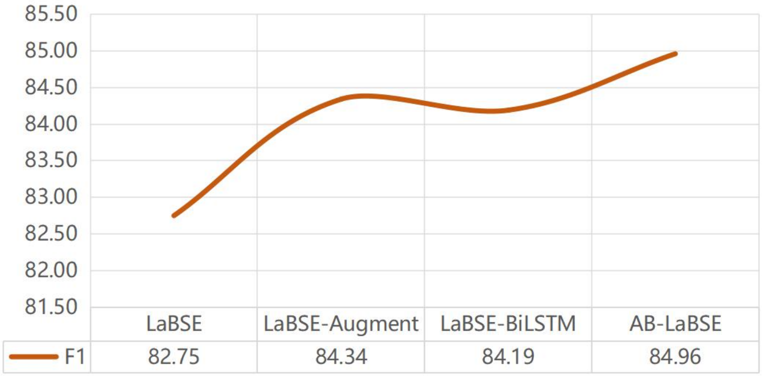


Figure 8. Two ablation studies of the hotel review binary Uyghur datasets using *LaBSE*.

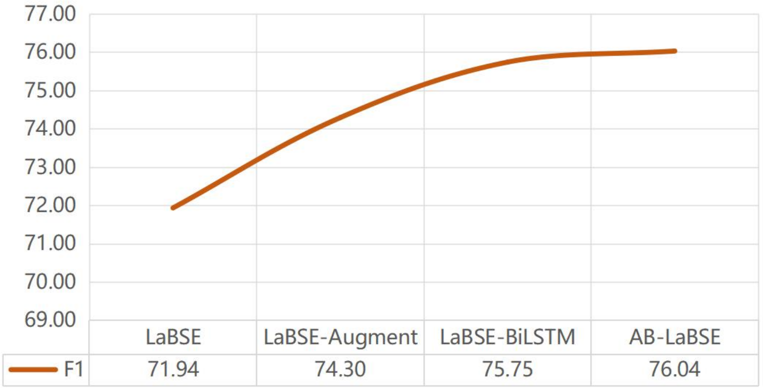


Figure 9. Two ablation studies of the emotion analysis five categories Uyghur datasets using *LaBSE*.

The effect of augmentation. For the binary dataset of hotel review, in the process of the experiment, we only performed data augmentation for the data items marked negative in view of the uneven distribution of training data in this dataset. Therefore, the performance of the negative F1 value attribute is improved by 3.1%, and the total F1 value is also improved by 1.59%. For the five categories of emotion analysis dataset, the data volume of this dataset is small. As a result, we performed random data augmentation on the data of the training set to improve the final performance. In Figure 9, we can see the result after data augmentation, and the overall effect has been improved by 2.36%. Experimental results show that this method can improve the performance of tasks with low resource datasets and enable us to obtain more semantic vector representations of original texts using pre-training models.

The effect of *BiLSTM*. By comparing the F1 values of various methods, we analyzed the two datasets separately, in the two categories datasets, the result improved by 1.44% after adding the *BiLSTM* layer. In the five categories dataset, the result improved by 3.81% after adding the *BiLSTM* layer. The *BiLSTM* layer not only considers the forward text sequence information, but also considers the reverse text sequence information, mining the semantic relationship between texts at a deeper level, thus greatly improving the performance of our model.

5. Conclusions and Future Work

The research of sentiment analysis is mature in language with abundant resources, but there is a lot of research space in language with scarce resources. This paper presents a method of Uyghur sentiment analysis based on a pre-training model with cross-language sentiment analysis. Different from previous studies, this method fully considers the insufficient data volume of low-resource languages and uses a cross-language pre-training model for modeling. In addition, long short-term memory (*BiLSTM*), a special type of *RNN*, is introduced into the model design. *BiLSTM* takes full advantage of its long-term dependence on learning. Experimental results show that the proposed model can improve the classification performance, has good robustness, and can effectively improve the ability of learning and understand the semantics of the model.

At present, we have used the method proposed by us in the project of Xinjiang Multilingual Information Technology Laboratory. In the future, we expect to make some advances in other directions in the field of NLP or other low-resource languages.

Author Contributions: Datasets, S.C., Z.K. and W.S.; methodology, Y.P. and S.C.; writing—original draft, S.C. and Y.P.; writing—review and editing, S.C. and Y.P.; data curation, S.C. and Q.G.; supervision, Z.K. and Q.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China under grant 2017YFC0820700; in part by the National Language Commission Research Project under grant ZDI135-96; and in part by the China Academy of Electronics and Information Technology, National Engineering Laboratory for Risk Perception and Prevention (NEL-RPP).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are very thankful to the editor and referees for their valuable comments and suggestions for improving the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. [CrossRef]
2. Bravo-Marquez, F.; Khanchandani, A.; Pfahringer, B. Incremental Word Vectors for Time-Evolving Sentiment Lexicon Induction. *Cogn. Comput.* **2022**, *14*, 425–441. [CrossRef]

3. Ahmad, M.; Aftab, S.; Bashir, M.S.; Hameed, N.; Ali, I.; Nawaz, Z. SVM optimization for sentiment analysis. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 393–398. [CrossRef]
4. Mathapati, S.; Nafeesa, A.; Manjula, S.; Venugopal, K. OTAWE-Optimized topic-adaptive word expansion for cross domain sentiment classification on tweets. In *Advances in Machine Learning and Data Science*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 213–224.
5. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef] [PubMed]
6. Nguyen, T.H.; Grishman, R. Event detection and domain adaptation with convolutional neural networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 365–371.
7. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259.
8. Rehman, A.U.; Malik, A.K.; Raza, B.; Ali, W. A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. *Multimed. Tools Appl.* **2019**, *78*, 26597–26613. [CrossRef]
9. Xu, H.; Liu, B.; Shu, L.; Yu, P.S. Dombert: Domain-oriented language model for aspect-based sentiment analysis. *arXiv* **2020**, arXiv:2004.13816.
10. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
11. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
12. Wu, Z.; Ong, D.C. Context-Guided BERT for Targeted Aspect-Based Sentiment Analysis. *arXiv* **2020**, arXiv:2010.07523.
13. Mao, Y.; Shen, Y.; Yu, C.; Cai, L. A Joint Training Dual-MRC Framework for Aspect Based Sentiment Analysis. *arXiv* **2021**, arXiv:2101.00816.
14. Li, X.; Fu, X.; Xu, G.; Yang, Y.; Xiang, T. Enhancing BERT Representation With Context-aware Embedding For Aspect-Based Sentiment Analysis. *IEEE Access* **2020**, *8*, 46868–46876. [CrossRef]
15. Ain, Q.T.; Ali, M.; Riaz, A.; Noureen, A.; Kamran, M.; Hayat, B.; Rehman, A. Sentiment analysis using deep learning techniques: A review. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 424.
16. Karimi, A.; Rossi, L.; Prati, A. AEDA: An Easier Data Augmentation Technique for Text Classification. *arXiv* **2021**, arXiv:2108.13230.
17. Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; Wang, W. Language-agnostic bert sentence embedding. *arXiv* **2020**, arXiv:2007.01852.
18. Shleifer, S. Low resource text classification with ulmfit and backtranslation. *arXiv* **2019**, arXiv:1903.09244.
19. Sun, L.; Xia, C.; Yin, W.; Liang, T.; Yu, P.S.; He, L. Mixup-Transformer: Dynamic Data Augmentation for NLP Tasks. *arXiv* **2020**, arXiv:2010.02394.
20. Bari, M.S.; Mohiuddin, T.; Joty, S. Multimix: A robust data augmentation framework for cross-lingual nlp. *arXiv* **2020**, arXiv:2004.13240.
21. Dymetman, M.; Bouchard, G.; Carter, S. Optimization and Sampling for NLP from a Unified Viewpoint. In Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology, Patna, India, 16 December 2012; pp. 79–94.
22. Chen, Y. Convolutional Neural Network for Sentence Classification. Master's Thesis, University of Waterloo, Waterloo, ON, Canada, 2015.
23. Li, S.; Wang, Z.; Zhou, G.; Lee, S.Y.M. Semi-supervised learning for imbalanced sentiment classification. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011.
24. Dwivedi, R.K.; Aggarwal, M.; Keshari, S.K.; Kumar, A. Sentiment analysis and feature extraction using rule-based model (RBM). In Proceedings of the International Conference on Innovative Computing and Communications, Valladolid, Spain, 19–20 February 2019; pp. 57–63.
25. Can, E.F.; Ezen-Can, A.; Can, F. Multilingual sentiment analysis: An rnn-based framework for limited data. *arXiv* **2018**, arXiv:1806.04511.
26. Wang, Y.; Huang, M.; Zhu, X.; Zhao, L. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 606–615.
27. Chen, P.; Sun, Z.; Bing, L.; Yang, W. Recurrent attention network on memory for aspect sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 452–461.
28. Sangeetha, K.; Prabha, D. Sentiment analysis of student feedback using multi-head attention fusion model of word and context embedding for LSTM. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 4117–4126. [CrossRef]
29. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv* **2019**, arXiv:1903.09588.
30. Yin, D.; Meng, T.; Chang, K.W. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. *arXiv* **2020**, arXiv:2005.04114.

31. Wei, J.; Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv* **2019**, arXiv:1901.11196.
32. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv* **2015**, arXiv:1508.07909.
33. Artetxe, M.; Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 597–610. [CrossRef]
34. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv* **2019**, arXiv:1908.10084.
35. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv* **2019**, arXiv:1911.02116.
36. Chi, Z.; Dong, L.; Zheng, B.; Huang, S.; Mao, X.L.; Huang, H.; Wei, F. Improving Pretrained Cross-Lingual Language Models via Self-Labeled Word Alignment. *arXiv* **2021**, arXiv:2106.06381.
37. Lample, G.; Conneau, A. Cross-lingual language model pretraining. *arXiv* **2019**, arXiv:1901.07291.