electronics

MDPI

Article

Enhancing Cross-Lingual Sarcasm Detection by a Prompt Learning Framework with Data Augmentation and Contrastive Learning

Tianbo An [1,2,3,†], Pingping Yan [1,2,3,†], Jiaai Zuo [4], Xing Jin [5,6], Mingliang Liu [1] and Jingrui Wang [1,2,3,*]

[1] College of Computer Science and Technology, Changchun University, Changchun 130022, China; antb@ccu.edu.cn (T.A.); 220702288@mails.ccu.edu.cn (P.Y.); 041940324@mails.ccu.edu.cn (M.L.)
[2] Key Laboratory of Intelligent Rehabilitation and Barrier-Free for the Disabled, Ministry of Education, Changchun University, Changchun 130022, China
[3] Jilin Provincial Key Laboratory of Human Health Status Identification & Function Enhancement, Changchun 130022, China
[4] College of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China; 19837955958@163.com
[5] School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310018, China; jinxing@hdu.edu.cn
[6] Experimental Center of Data Science and Intelligent Decision-Making, Hangzhou Dianzi University, Hangzhou 310018, China
* Correspondence: wangjingrui530@gmail.com
† These authors contributed equally to this work.

Abstract: Given their intricate nature and inherent ambiguity, sarcastic texts often mask deeper emotions, making it challenging to discern the genuine feelings behind the words. The proposal of the sarcasm detection task is to assist us with more accurately understanding the true intention of the speaker. Advanced methods, such as deep learning and neural networks, are widely used in the field of sarcasm detection. However, most research mainly focuses on sarcastic texts in English, as other languages lack corpora and annotated datasets. To address the challenge of low-resource languages in sarcasm detection tasks, a zero-shot cross-lingual transfer learning method is proposed in this paper. The proposed approach is based on prompt learning and aims to assist the model with understanding downstream tasks through prompts. Specifically, the model uses prompt templates to construct training data into cloze-style questions and then trains them using a pre-trained cross-lingual language model. Combining data augmentation and contrastive learning can further improve the capacity of the model for cross-lingual transfer learning. To evaluate the performance of the proposed model, we utilize a publicly accessible sarcasm dataset in English as training data in a zero-shot cross-lingual setting. When tested with Chinese as the target language for transfer, our model achieves F1-scores of 72.14% and 76.7% on two test datasets, outperforming the strong baselines by significant margins.

Keywords: cross-lingual; prompt learning; sarcasm detection; Chinese

Citation: An, T.; Yan, P.; Zuo, J.; Jin, X.; Liu, M.; Wang, J. Enhancing Cross-Lingual Sarcasm Detection by a Prompt Learning Framework with Data Augmentation and Contrastive Learning. *Electronics* **2024**, *13*, 2163. https://doi.org/10.3390/electronics13112163

Academic Editor: Ping-Feng Pai

Received: 16 May 2024
Revised: 29 May 2024
Accepted: 30 May 2024
Published: 1 June 2024

1. Introduction

As Internet technology swiftly advances, social media has emerged as the primary platform for individuals to voice their opinions on trending events, political matters, and societal issues. Sarcasm, which is a prevalent rhetorical device in daily communication, has significantly permeated social media [1]. Sometimes, sarcasm [2] is used to make jokes, express annoyance, or degrade someone. These offensive and malicious sarcastic remarks may lead to disputes and conflicts. However, sarcastic expression can obscure the genuine emotional tendencies of the text, making them difficult to discern. Accurate sarcasm detection can enhance our understanding of users' true intentions. This not only improves the performance of natural language processing tasks such as opinion mining and sentiment

analysis, but it also enhances the effectiveness of fake news detection and hate speech recognition for safeguarding the harmony and health of the online environment [3–7].

Sarcasm detection is generally considered a classification task [8] that aims to identify if a text is ironic. Initial research in this field employed rule-based methods [1,9], which often relied on fixed rules or dictionaries to identify irony. Subsequently, researchers began adopting machine learning methods [10] and manually designing features for sarcasm detection. In recent years, the rapid advancement of deep learning has led to its widespread adoption in numerous studies aimed at solving the problem of sarcasm detection. Deep learning methods can automatically extract textual features and understand contextual relationships, allowing for more precise identification of sarcastic texts [8,11]. To further enhance the performance of methods such as deep learning and neural networks in sarcasm detection tasks, researchers requires the support of large-scale annotation datasets [12]. The sarcastic content on social media platforms covers multiple languages, but the existing sarcastic corpora are primarily in English [13–16]; researchers lack datasets in other languages (for example, Chinese), which to a certain extent limits the application and development of sarcasm detection technology in multilingual environments.

Researchers typically use a cross-lingual transfer strategy to transfer the knowledge gained in high-resource languages to low-resource languages in order to address the issue of low-resource languages [17–19]. The key to cross-lingual tasks is the effective mapping of representations from different languages into a shared space. In this research trend, pre-trained multilingual language models, including mBERT [20], XLM [21], and XLM-RoBERTa [22], have shown outstanding performance. Notably, at a time when annotation resources are expensive, cross-lingual zero-shot learning is beginning to emerge. This method, which can be trained without any labeled data in the target language, has gained widespread popularity in the academic community. For instance, adaptation learning was used to fine-tune mBERT and XLM-R pre-trained models on English and Dravidian datasets, addressing the low-resource sentence-level fake news classification problem in Dravidian languages [23]. Cross-lingual XLM-R transformer models are used for sentiment classification tasks in the low-resource Hindi language [24]. Recently, in order to guide the model to better understand the task, prompt learning was introduced when solving cross-lingual tasks using pre-trained multilingual language models [25–27]. For example, Qi et al. [28] introduced a prompt learning framework that generates augmented questions using multilingual templates and then uses a loss of consistency to enhance the correspondence between different languages in the original and augmented questions. However, due to the relatively small differences between the enhanced questions generated using cross-lingual templates and the original questions, reducing their representation differences through consistency loss does not significantly enhance the performance of the model.

In summary, existing sarcasm detection work mainly focuses on classifying text based on contextual information or addressing multimodal issues [29,30] and neglects the challenges of low-resource languages. Currently, prompt learning can be used to enhance the performance of pre-trained models in cross-lingual classification tasks [28]. However, there are still shortcomings in aligning different languages into the same space in these works. Therefore, we hope to further improve the model based on prompt learning for application in cross-lingual sarcasm detection tasks. Specifically, we propose a zero-shot cross-lingual transfer learning framework for Chinese in this paper. The framework is based on Prompt learning with Data augmentation and Contrastive learning (PDC), and we further enhance its cross-lingual transfer capability. We first introduce data augmentation to expand the training data and construct the original data and augmented data into cloze-style questions with <mask> using prompt templates, which are used as inputs for the pre-trained model. Then, a contrastive learning strategy is adopted to make the representations of the original questions and the augmented questions closer. Finally, we use consistency loss to constrain the probability distributions of answers in different languages. The proposed method is trained by minimizing the total loss, which includes the cross-entropy term, the contrastive loss term, and the consistency loss term. In this paper, we utilize PDC to strengthen the

pre-trained multilingual language model XLM-R [22]. The experimental results on test datasets show that PDC can effectively enhance the pre-trained model for the cross-lingual sarcasm detection task in the zero-shot cross-lingual setting. The contributions of this work are as follows:

- A cross-lingual sarcasm detection problem that facilitates research on sarcastic content in low-resource languages is implemented in this paper.
- A novel cross-lingual sarcasm detection framework based on prompt learning that combines data augmentation and contrastive learning is proposed to improve cross-lingual transfer learning ability, and consistency loss is used to make the probability distribution of answers in different languages as similar as possible.
- Experiments are conducted in a zero-shot cross-lingual setting, where the model is trained on the English sarcasm dataset and subsequently evaluated on the Chinese dataset. The experimental outcomes indicate that the model in this paper achieves outstanding performance.

2. Related Works

In this section, we comprehensively sort out the research context of sarcasm detection and cross-lingual tasks; then, we trace the development processes of their research methods.

2.1. Sarcasm Detection

Early sarcasm detection mainly relied on fixed patterns or guidelines such as preset grammar rules, specific labels, and metaphorical rules. The tag provided in the tweet is used to determine whether the emotions of the rest of the tweet conflict with the tag [1]. Metaphor rules and Google searches can be used together to distinguish sarcastic tendencies [31]. In statistical methods [32], the semantic correlations between words can be measured based on the similarity of Wordnet to extract features [33]. Later, traditional machine learning methods [34,35], including SVM [36], decision trees [37], and Naïve Bayes [38], were applied to solve sarcasm detection tasks.

In recent years, researchers have deeply explored the application of deep learning methods in the field of sarcasm detection. Poria et al. [39] first proposed using deep learning for sarcasm detection in 2016, using CNN for feature extraction followed by feeding these features into a support vector machine for classification. To more effectively extract features of sarcastic text and improve the accuracy of sarcasm detection, many studies have combined various deep learning techniques [40,41] such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms to construct sarcasm detection models. With the introduction of the pre-trained model BERT, researchers have expanded on it by incorporating emotional features and contextual information for sarcasm detection, further advancing the field [3,42]. Later, graph convolutional networks (GCNs) were widely applied to text classification tasks. There is a study that uses graphs to represent the emotional and syntactic relationships in the text, thereby connecting the information in the entire text [43]. To avoid the noise introduced by manually constructing static graphs, a method of iterative incongruity graph learning has been proposed to alleviate this problem [44]. Additionally, some studies have utilized external knowledge and information to enhance the model's semantic understanding of sarcastic text [45,46].

2.2. Cross-Lingual Tasks

The purpose of cross-lingual transfer learning is to utilize the knowledge from the source language to enhance the performance of the task in the target language, particularly when the target language lacks extensive annotated data [17–19]. Cross-lingual word embedding embeds vocabulary from different languages into a shared embedding space, allowing different languages to share lexical and semantic information. MUSE [47] employs a domain-adversarial setting to address the challenge of insufficient supervision with the aim of aligning multilingual word vectors into a shared vector space effectively. Moreover, LASER [48] utilizes a single BiLSTM encoder to construct sentence embeddings for all

languages; the encoder is combined with an auxiliary decoder. With the remarkable achievements of pre-trained models in natural language processing tasks, researchers have proposed pre-trained multilingual language models such as XLM [21], mBERT [20], and XLM-RoBERTa [22]. These methods are all based on the transformer architecture and are pre-trained on large-scale text data. They can achieve zero-shot learning [49,50] and perform excellently in cross-lingual tasks. Notably, XLM-RoBERTa has been pre-trained on large-scale data in 100 languages, is widely adopted in cross-lingual classification tasks [51,52], and has achieved state-of-the-art results.

Considering the diversity of languages on social media platforms, cross-lingual transfer techniques have been tried in order to solve low-resource issues in NLP tasks like natural language inference [53], fake news detection [23], sentiment analysis [24], and hate speech detection [54]. Recent cross-lingual research has employed some strategies to improve the performance of pre-trained multilingual language models on cross-lingual tasks. For example, data augmentation has frequently been used to expand training data and enhance the generalization ability of models in diverse linguistic contexts [55,56]. To further bridge the gap between different languages, strategies such as adversarial training [57,58] and contrastive learning [54] have been used to improve the capability of the model for cross-lingual processing. Among them, adversarial training introduces adversarial perturbations to enhance the robustness of the model, while contrastive learning utilizes contrastive loss to effectively discern the similarities and disparities across languages. Moreover, the advent of a prompt learning framework based on cross-lingual templates provided new perspectives and methodologies for solving cross-lingual tasks.

3. Proposed Method

3.1. Framework

The PDC framework proposed in this paper is clearly depicted in Figure 1. Our framework is mainly divided into four modules, including data augmentation, prompt learning, contrastive learning, and the overall training objective. The augmented target language dataset is generated from the source language dataset through the translation module of data augmentation. The prompt learning module embeds the input data into the prompt templates, thereby constructing the cloze-style questions with <mask>. Then, we utilize the contrastive learning module to align the embedded representations of the original question and the augmented question as much as possible. Finally, the masking layer of the pre-trained model is responsible for calculating the probability distribution of the masked token, which is then regularized for consistency. The model is trained by minimizing the overall loss, which includes the cross-entropy loss that assesses the accuracy of categorization, the contrastive loss for reducing differences in representations across languages, and the Kullback–Leibler divergence (KLD) loss to ensure answer consistency. The main components of the framework are elaborated on below.

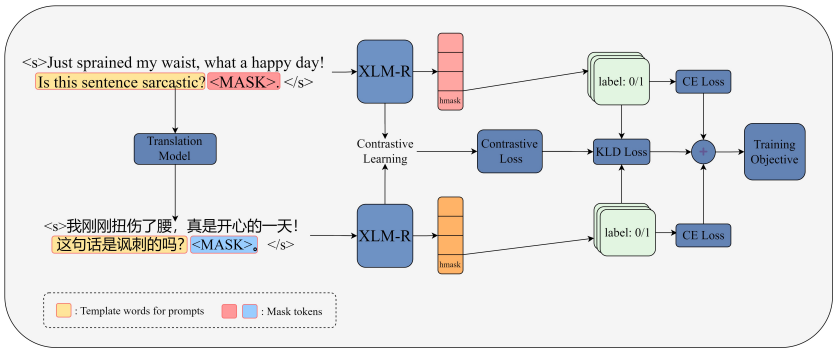

Figure 1. The framework of PDC. The leftmost input sentences are the original and the augmented questions constructed by the prompt templates. The upper and lower sides of the framework are the training processes of the original question and the augmented question, respectively. The training objective is the combined total of cross-entropy loss, contrastive loss, and KLD loss.

3.2. Data Augmentation

A data augmentation strategy can effectively augment training data, increasing the model's generalization ability and robustness. They mainly operate through synonym replacement, random insertion, translation, or other methods. Among them, translation can utilize a bilingual dictionary to randomly select words from the source language sentences and substitute them with their counterparts in the target language [55] or use machine translation to generate pseudo-parallel data [54]. To be able to better handle the differences between different languages, we use translation as a means of data augmentation to translate the source language into the target language through machine translation. Specifically, we used Google Translate, which offers high accuracy and extensive language support, as the primary translation module to translate the source language into the target language. Given a source language dataset $\mathcal{S}$, where $\mathcal{S} = \{t_1^s, t_2^s, \ldots, t_i^s\}$. The translation strategy translates each text $t_i^s \in \mathcal{S}$ into the target text t_i^t, as obtained through Equation (1).

$$t_i^t = Translate(t_i^s) \tag{1}$$

3.3. Prompt Learning

Prompt learning emerged with the launch of GPT-3 [59], which, together with pre-training, constitutes the fourth paradigm of NLP: the prediction paradigm. In recent years, prompt learning has garnered significant attention due to its ability to help models understand given tasks through prompt templates [27,60]. There have been many studies focused on exploring methods for constructing prompt templates. Schick et al. [25] employed prompt learning with manually defined templates for text classification tasks. The template uses a masking layer to transform the input into a sentence with a special mask token and then uses a verbalizer (a mapping function) to map each label to a single word in the given vocabulary. To address cross-lingual tasks, Qi et al. [28] translated the prompt words in the source language template with those in other languages. Inspired by these works [25,28], we determined a template, and examples of applying the prompt template are shown in Figure 2. Following this, we embedded the training data into the corresponding prompt template to construct the cloze-style original question X_i^s and the augmented question X_i^t, which are computed by Equations (2) and (3), respectively.

$$X_i^s = Prompt(t_i^s) \tag{2}$$

$$X_i^t = Prompt(t_i^t) \tag{3}$$

The embedded representations E_i^s and E_i^t of texts in different languages are obtained through the pre-trained model XLM-RoBERTa and can be calculated by Equations (4) and (5).

$$E_i^s = XLM\text{-}R(X_i^s) \tag{4}$$

$$E_i^t = XLM\text{-}R(X_i^t) \tag{5}$$

Figure 2. Applying prompt templates to examples of original text and augmented text, where the augmented text is generated by translating the source language text into the target language. The verbalizer demonstrates label mapping. The uncolored parts in the figure represent sarcastic texts, the yellow parts represent template words for prompts, and the red and blue parts represent masked tokens for the original question and augmented question, respectively.

3.4. Contrastive Learning

Contrastive learning seeks to enable the model to better learn the characteristics of the data by comparing the relationships between different samples, thereby generating more discriminative representation vectors. It mainly optimizes the representational learning ability of the model by constructing positive and negative samples, reducing the distance between positive samples and expanding the distance between negative samples [61]. In this paper, we consider the representations of the same sample in different languages (E_i^s, E_i^t) as positive pairs, while the combinations of other samples' representations in the same batch are considered negative pairs. The contrastive loss $\mathcal{L}_{CLC}$ can be obtained through Equations (6) and (7), where $sim(u, v)$ calculates the cosine similarity for each pair of embedded representations, and τ represents a temperature parameter.

$$sim(u, v) = \frac{u \cdot v}{\| u \| \| v \|} \tag{6}$$

$$\mathcal{L}_{CLC} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{sim(E_i^s, E_i^t)/\tau}}{\sum_{j=1}^{N}(e^{sim(E_i^s, E_j^t)/\tau} + e^{sim(E_j^s, E_i^t)/\tau})} \tag{7}$$

3.5. Training Objective

The original and augmented questions serve as inputs to the pre-trained model, and the probability distribution of masked token answers is calculated through the MLM layer of the pre-trained cross-lingual model. Then, the labels of the data are mapped to the indices of the corresponding answer words in the given vocabulary using the function $\mathcal{M}$, which maps from $\mathcal{Y}$ to $\mathcal{V}$. The probability distribution of the answer to the original question y^s is computed using Equation (8).

$$y^s = softmax(Wh^s_{<mask>}) \tag{8}$$

where the parameters of the pre-trained MLM layer are represented by $W \in \mathbb{R}^{l \times d}$, and $h^s_{<mask>}$ is used to represent the contextual representation of the <mask> token. In the same manner, the answer probability distribution for the augmented question y^t is computed.

The training data are organized into triplets, represented as (X_i^s, X_i^t, Y_i), where $Y_i \in \mathcal{Y}$ denotes the index of the label. The cross-entropy losses of the original question X_i^s and the augmented question X_i^t are obtained through Equations (9) and (10), respectively.

$$\mathcal{L}_X^s = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{l} I(j = \mathcal{M}(Y_i)) \log y_{i,j}^{X_i^s} \tag{9}$$

$$\mathcal{L}_X^t = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{l} I(j = \mathcal{M}(Y_i)) \log y_{i,j}^{X_i^t} \tag{10}$$

where I represents the batch. If C is true, then $I(C)$ is 1, otherwise it is 0. Additionally, $y_{i,j}^{X_i^s}$ represents the j-th element of the answer probability distribution y_i^s for the original question X_i^s (the same applies to $y_{i,j}^{X_i^t}$).

In cross-lingual tasks, there may be a certain bias in the answer probability distribution of the source and target languages. To enhance the consistency for the probability distribution of answers between the original and augmented questions $(y_i^{X_i^s}, y_i^{X_i^t})$, we adopt the Kullback–Leibler divergence (KLD) loss. $\mathcal{L}_{KLD}$ is calculated as in Equation (11) below.

$$\mathcal{L}_{KLD} = \frac{1}{N}\sum_{i=1}^{N}(KL(y_i^{X_i^s} y_i^{X_i^t}) + KL(y_i^{X_i^t} y_i^{X_i^s})) \tag{11}$$

In this paper, we obtain the total loss by weighted summing of the cross-entropy loss of the training data for source and target languages, the KLD loss, and the contrastive loss.

The training objective is the total loss $\mathcal{L}$, which is calculated by Equation (12), where λ is the weight parameter.

$$\mathcal{L} = \frac{1-\lambda}{3}\mathcal{L}_X^s + \frac{1-\lambda}{3}\mathcal{L}_X^t + \frac{1-\lambda}{3}\mathcal{L}_{KLD} + \lambda\mathcal{L}_{CLC} \tag{12}$$

4. Experimental Settings

This section details the datasets, baseline models, and implementation details.

4.1. Datasets

Our experiments involve three different datasets (dataset statistics are shown in Table 1): specifically, including a news headline sarcasm dataset (English) and two Chinese datasets. Table 2 shows examples of sarcastic and non-sarcastic texts from the three datasets. These datasets are detailed below:

Table 1. Dataset statistics.

Language	Dataset	Ironic	Non-Ironic	Total
English	News Headlines Dataset [62]	11,724	12,779	24,503
Chinese	A Balanced Chinese Internet Sarcasm Corpus [63]	1000	1000	2000
	A New Benchmark Dataset [64]	968	7734	8702

Table 2. Examples of ironic and non-ironic texts in English and Chinese.

Language	Dataset	Example	Label
English	News Headlines Dataset [62]	"bus-stop ad has more legal protections than average citizen"	ironic
		"psychologists push for smartphone warning labels"	non-ironic
Chinese	A Balanced Chinese Internet Sarcasm Corpus [63]	"实用，可以一边走路一边泡脚了"	ironic
		"早安！美好的一天！"	non-ironic
	A New Benchmark Dataset [64]	"晕。小桔你不够专业呀，亏你工龄还那么长了。"	ironic
		"再过一会儿，十二点整在新台的第一次播出就将开始啦！"	non-ironic

- News Headlines Dataset [62]: The dataset is collected from two news websites—The Onion and HuffPost—containing about 28 K headlines, of which 13 K are sarcastic. The Onion aims to produce sarcastic news, from which all headlines in the "News Briefing" and "Photo News" categories have been collected; HuffPost publishes real (non-sarcastic) news. Compared to existing Twitter datasets, the news dataset contains high-quality labels with much less noise.
- A Balanced Chinese Internet Sarcasm Corpus [63]: The corpus contains 2 K Chinese annotated texts and balances sarcasm and non-sarcasm and longer and shorter texts, both in a 1:1 ratio. The shorter sarcastic texts mainly originate from the corpus constructed by Tang and Chen [65], while the longer texts are from Onion the Storyteller (a Weibo account). The shorter non-sarcastic texts are collected from the comment section of the official WeChat public account, while the longer texts are mainly information crawled from some official accounts on Weibo and WeChat. The balanced dataset is constructed by randomly selecting a portion of the data from each source.
- A New Benchmark Dataset [64]: The benchmark dataset is a fine-grained dataset that includes five levels of sarcasm: 1 (no sarcasm), 2 (unlikely to be sarcastic), 3 (indeterminate), 4 (mild sarcasm), and 5 (strong sarcasm). The data were randomly collected

from Weibo posts, comprising 8.7k posts, with each annotated by five annotators. Ironic data (including 4 and 5) only account for 11.1% of the labeled data.

4.2. Baseline Models

We conduct a comparative analysis between the proposed model and several baseline methods. The baseline methods are described below:

- LASER+SVM [66]: LASER is an architecture designed for embedding multilingual sentences that uses a single, language-agnostic BiLSTM encoder. The SVM classifier is then trained for sarcasm detection.
- LASER+LR: In this method, LASER is used for text embeddings, followed by training a logistic regression (LR) classifier for sarcasm detection.
- LASER+RNN: The same as above—LASER technology is utilized to extract the embedded representation of the text, followed by the application of an RNN model from deep learning for the classification of sarcasm.
- mBERT: Multilingual BERT is a variant of BERT designed for multilingual tasks. It has been pre-trained on the Wikipedia corpora and covers 102 languages.
- XLM-R: The XLM model is based on the transformer architecture, similar to the BERT model, and it has been pre-trained on Wikipedia with 100 languages. Compared to XLM, XLM-R extends the training languages and increases the scale of the training corpus.
- PCT-XLM-R: PCT is a cross-lingual framework that is based on prompt learning with cross-lingual templates. PCT-XLM-R is obtained by applying PCT to XLM-R.

4.3. Implementation Details

Our model is implemented using TensorFlow 2.4.0 and trained on an RTX 3090. We apply PDC to the pre-trained multilingual language model XLM-R, which consists of 12 transformer layers and outputs 768-dimensional tensors. During the experiments, the RMSProp optimizer [67] is used to adaptively adjust the learning rate for each parameter, with the initial learning rate set to 2×10^{-5}. The number of training epochs for our model is 10, the batch size of the training data is 8, and the dropout rate is set to 0.1. In addition, the temperature parameter for contrastive learning τ is set to 0.5, and the weight parameter λ for the total loss of the training target is set to 0.55.

5. Experiment Results

5.1. Main Results

The experiments are conducted in a zero-shot cross-lingual setting, where models are trained on the source language (English) dataset and then tested on the target language dataset (Chinese). Additionally, PDC is evaluated based on XLM-R. The principal benchmarking metrics, including precision, accuracy, and F1-score, are used to evaluate the performance of the models. These metrics can be computed as in Equations (13)–(16).

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{13}$$

$$Precision = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i} \tag{14}$$

$$Recall = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i} \tag{15}$$

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{16}$$

The ROC (receiver operating characteristic) curve is a tool used to evaluate the classification performance of a model. The curve is drawn with the true positive rate (TPR) as the ordinate and the false positive rate (FPR) as the abscissa and shows the performance of the model under different thresholds. The AUC (area under the curve) is defined as the area under the ROC curve, which is used as a numerical value to assess the performance of the model. Balanced datasets can better reflect the classification performance of models across all categories. Therefore, in this work, we employ the ROC-AUC to evaluate the classification effectiveness of the model on the Balanced Chinese Internet Sarcasm Corpus.

The precision, accuracy, and F1-score results for all models as tested on the two Chinese datasets are shown in Table 3. Figure 3 shows the ROC curves and calculated AUC values for all models on the balanced test dataset. We first observe the results of traditional machine learning classifiers and a deep learning model. From Table 3, on the balanced test dataset, LASER+LR surpasses the others in terms of precision, accuracy, and F1-score among these three methods. Additionally, as shown in Figure 3, LASER+RNN has the highest ROC-AUC value, indicating that its classification performance is better than that of LASER+SVM and LASER+LR. On the imbalanced test dataset, LASER+RNN outperformed LASER+SVM and LASER+LR, achieving the best results across the three metrics. We then employed the pre-trained models mBERT, XLM-R, and PCT-XLM-R as baseline models. On the balanced dataset, we found that the F1-score of LASER+LR is higher than that of the pre-trained models mBERT and XLM-R. From this result, it can be observed that LASER is capable of generating better embeddings for cross-lingual tasks. At the same time, compared with other baseline models, the PCT-XLM-R model has achieved significant improvements in precision, accuracy, and F1-score. The ROC-AUC value of PCT-XLM-R also surpasses that of other baseline models. This suggests that the PCT framework can significantly enhance the performance of XLM-R in cross-lingual sarcasm detection. On the imbalanced dataset, mBERT achieves the best accuracy and F1-score among the baseline models.

Compared to the best-performing baseline model, PCT-XLM-R, on the balanced dataset, the proposed model improves precision by approximately 5.9%. In addition, both accuracy and F1-score increased by around 5%. From Figure 3, it can be seen that our model has a higher ROC-AUC value than other baseline models, reaching above 75.2%, which indicates the best classification performance of our model. On the imbalanced dataset, our model improved accuracy by around 1.4% and F1-score by about 1.1%, with PCT-XLM-R achieving the best precision value. These results demonstrate that our model outperforms other baseline models and can better perform cross-lingual sarcasm detection tasks.

Table 3. Performance evaluation of different models based on precision, accuracy, and F1-score.

Model	**A Balanced Chinese Internet Sarcasm Corpus**			**A New Benchmark Dataset**		
	Precision	**Accuracy**	**F1-Score**	**Precision**	**Accuracy**	**F1-Score**
LASER+SVM	55.25	54.80	53.82	78.85	56.34	64.44
LASER+LR	56.34	55.85	54.99	78.61	56.26	64.37
LASER+RNN	54.90	54.55	53.71	79.47	57.66	65.50
M-Bert	60.03	56.10	52.13	80.00	72.14	75.64
XLM-R	68.34	58.49	52.65	80.62	65.62	71.53
PCT-XLM-R	71.60	67.60	67.41	**83.33**	64.86	71.07
Ours	**77.51**	**72.50**	**72.14**	82.62	**73.52**	**76.70**

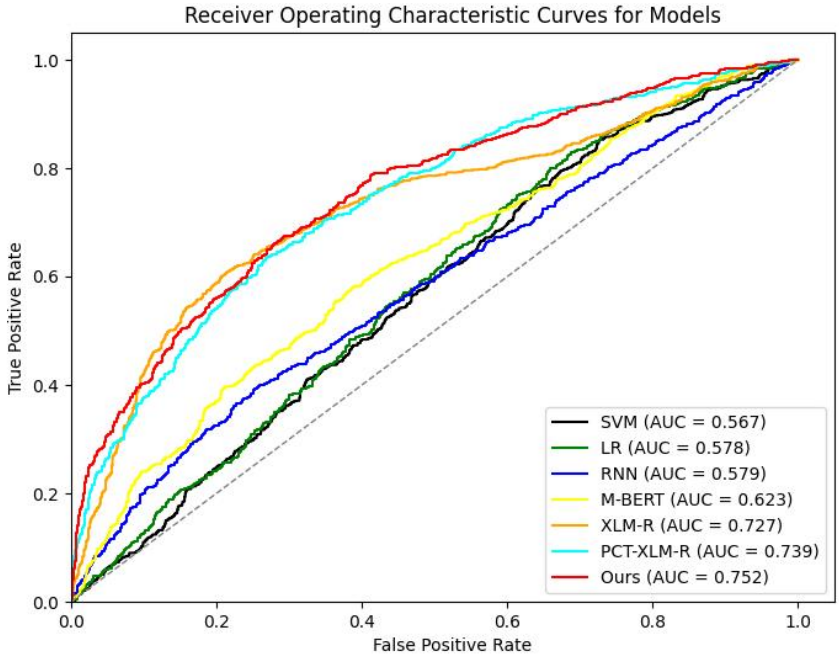

Figure 3. The ROC curves of the models and the corresponding AUC values.

5.2. Ablation Study

To evaluate the impact of each crucial element of the model, we conduct ablation studies in a zero-shot cross-lingual setting. When the data augmentation (DA) component is eliminated, contrastive learning remains effective because our model utilizes cross-lingual templates. When the cross-lingual contrastive learning (CLC) component is omitted, the model no longer calculates contrastive loss. All of the models for ablation experiments are based on XLM-R. During the experiments, we find that the computational cost of the ablation experiments does not differ significantly from the comparative experiments, with a maximum difference in computation time cost of no more than ten minutes. Therefore, we do not show the computation time in the table.

The experimental results of our model and the ablation models on two Chinese test datasets are shown in Table 4. These results are evaluated based on three key performance metrics: precision, accuracy, and F1-score. On the balanced test dataset, compared to the model without data augmentation, the complete model improves precision by 8.6%, accuracy by 8.2%, and F1-score significantly by 11%. Compared to the model without contrastive learning, the complete model improves precision by 3%, while accuracy and F1-score improve by approximately 3.4%. The results on the balanced dataset indicate that data augmentation plays a more important role in improving performance on balanced data than contrastive learning. The translation strategy used for data augmentation generates target language data, which helps the model better learn the features of different language data.

Table 4. Precision, accuracy, and F1-score comparisons of ablation models.

Model	A Balanced Chinese Internet Sarcasm Corpus			A New Benchmark Dataset		
	Precision	**Accuracy**	**F1-Score**	**Precision**	**Accuracy**	**F1-Score**
w/o DA	68.88	64.35	61.10	82.85	69.19	73.95
w/o CLC	74.52	69.10	68.74	**83.97**	64.57	70.93
Ours	**77.51**	**72.50**	**72.14**	82.62	**73.52**	**76.70**

On the imbalanced test dataset, the complete model achieves a 4.3% increase in accuracy and an approximately 2.8% increase in F1-score compared to the model without data augmentation. Compared to the model without contrastive learning, the complete model achieves

a significant 9% increase in accuracy and a 5.8% increase in F1-score. The results on the imbalanced dataset indicate that contrastive learning plays a more crucial role in improving model performance. By comparing the similarity and dissimilarity between different samples, contrastive learning helps the model more effectively identify minority class samples. These results indicate that both contrastive learning and data augmentation can effectively enhance the performance of a cross-lingual framework based on prompt learning.

5.3. Visualization Analysis

To provide a more intuitive presentation of the model's classification performance in the target language (Chinese) while ensuring the credibility and comparability of the visualization results and avoid potential sample bias from an imbalanced dataset, we use UMAP to visualize representations of text embeddings generated by the model on the balanced test dataset. UMAP [68] (uniform manifold approximation and projection for dimension reduction) is a manifold learning algorithm that can be used for dimensionality reduction and data visualization. The visualization results are shown in Figure 4. In this figure, the red dots denote sarcastic texts, while the blue dots denote non-sarcastic texts. It can be intuitively seen from Figure 4 that there is a clear polarization between non-ironic and ironic texts, and the red and blue points only have less overlap, indicating that our model can effectively classify balanced Chinese sarcasm data.

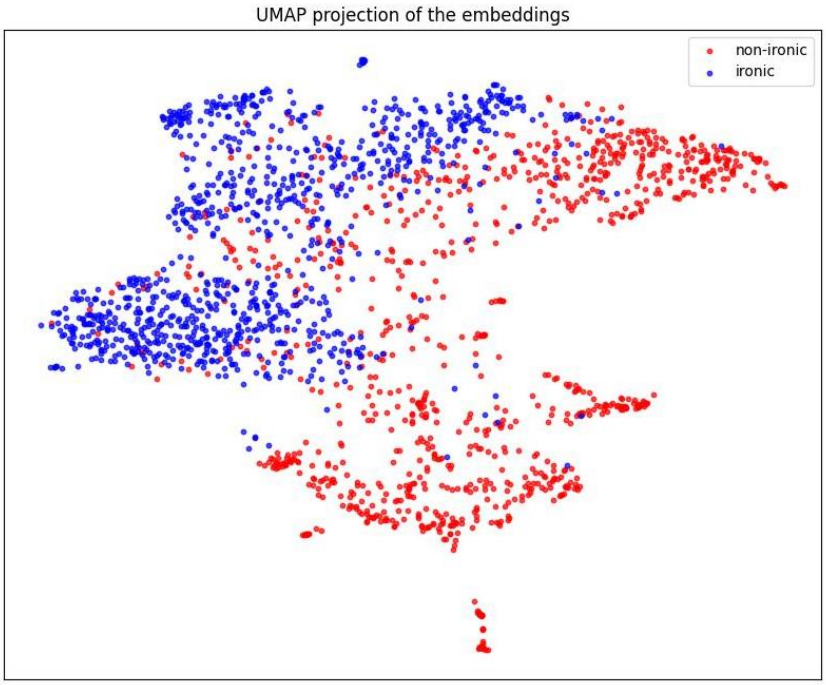

Figure 4. UMAP visualizes the embedded representations of the test data in Chinese, where the blue and red points represent ironic and non-ironic texts, respectively.

6. Conclusions and Discussion

Accurately identifying sarcastic text is instrumental for capturing the underlying emotions in the text, which, in turn, enables a deeper understanding of the speaker's attitude or intention. Therefore, sarcasm detection plays a crucial role in the field of NLP. However, the majority of research on sarcasm detection is predominantly focused on English texts, and there is a lack of Chinese datasets. To mitigate this issue of Chinese as a low-resource language in sarcasm detection, we employ the strategy of cross-lingual transfer learning.

In this work, we propose a framework named PDC and based on prompt learning that combines data augmentation and contrastive learning for cross-lingual sarcasm detection. The data augmentation strategy can augment training data by generating target language data through a translation module, enhancing the generalization ability of the model in

different languages. To better understand the downstream task, the original and augmented data are constructed as cloze-style questions with <mask> tokens using prompt templates. Then, contrastive learning is employed to reduce the difference between the embedded representations of the original and augmented questions, enhancing the capability of cross-lingual knowledge transfer for the model. Results from testing on two Chinese datasets indicate that the accuracy and F1-score of our method outperform the baseline models in the zero-shot cross-lingual setting. In the ROC curve that considers the model performance at different classification thresholds, our model has the highest ROC-AUC value on the balanced test dataset compared to all other evaluated models. Ablation experiments and the visualization result further suggest that our study contributes to research on sarcasm detection in low-resource languages (Chinese).

The proposed PDC framework has achieved certain results in cross-lingual sarcasm detection tasks, but there are still some limitations. The diverse forms of sarcasm make it difficult to capture some sarcastic expressions. Furthermore, our research has primarily focused on knowledge transfer from English to Chinese and has not yet explored its effectiveness in other languages. In future work, we should improve the model by further deepening our understanding of sarcastic texts. At the same time, it is worth further study to expand the model's support for more languages. Notably, more effective cross-lingual transfer strategies should be explored to enhance the transfer effects of the framework. Through these efforts, we hope to further advance the research on cross-lingual sarcasm detection.

Author Contributions: Conceptualization, T.A. and J.W.; Methodology, X.J. and J.W.; Software, P.Y. and J.Z.; Validation, P.Y. and X.J.; Formal analysis, M.L.; Investigation, T.A., P.Y. and M.L.; Resources, J.Z. and X.J.; Data curation, X.J.; Writing—original draft, T.A. and P.Y.; Supervision, J.W.; Project administration, J.W.; Funding acquisition, T.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Jilin Provincial Department of Science and Technology (YDZJ202301ZYTS496, YDZJ202303CGZH010, 20240305048YY, YDZJ202201ZYTS549) and the Education Department of Jilin Province (JJKH20230673KJ, JJKH20220597KJ).

Data Availability Statement: Data are contained within the article.

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

References

1. Maynard, D.G.; Greenwood, M.A. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In Proceedings of the Lrec 2014 Proceedings, ELRA, Reykjavik, Iceland, 26–31 May 2014.
2. Merriam-Webster, I. *The Merriam-Webster Dictionary*; Merriam-Webster: Springfield, MA, USA, 1995.
3. Eke, C.I.; Norman, A.A.; Shuib, L. Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and BERT model. *IEEE Access* **2021**, *9*, 48501–48518.
4. Majumder, N.; Poria, S.; Peng, H.; Chhaya, N.; Cambria, E.; Gelbukh, A. Sentiment and sarcasm classification with multitask learning. *IEEE Intell. Syst.* **2019**, *34*, 38–43.
5. Ghorbanali, A.; Sohrabi, M.K.; Yaghmaee, F. Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. *Inf. Process. Manag.* **2022**, *59*, 102929.
6. Maladry, A.; Lefever, E.; Van Hee, C.; Hoste, V. Irony detection for dutch: A venture into the implicit. In Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, Dublin, Ireland, 26 May 2022; pp. 172–181.
7. Reyes, A.; Saldívar, R. Linguistic-based Approach for Recognizing Implicit Language in Hate Speech: Exploratory Insights. *Comput. Sist.* **2022**, *26*, 101–111.
8. Wen, Z.; Gui, L.; Wang, Q.; Guo, M.; Yu, X.; Du, J.; Xu, R. Sememe knowledge and auxiliary information enhanced approach for sarcasm detection. *Inf. Process. Manag.* **2022**, *59*, 102883.
9. Reyes, A.; Rosso, P.; Veale, T. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.* **2013**, *47*, 239–268.
10. Joshi, A.; Bhattacharyya, P.; Carman, M.J. Automatic sarcasm detection: A survey. *Acm Comput. Surv.* **2017**, *50*, 1–22.
11. Zhang, S.; Zhang, X.; Chan, J.; Rosso, P. Irony detection via sentiment-based transfer learning. *Inf. Process. Manag.* **2019**, *56*, 1633–1644.

12. Ranasinghe, T.; Zampieri, M. Multilingual offensive language identification with cross-lingual embeddings. *arXiv* **2020**, arXiv:2010.05324.
13. Walker, M.A.; Tree, J.E.F.; Anand, P.; Abbott, R.; King, J. A Corpus for Research on Deliberation and Debate. In Proceedings of the LREC, Istanbul, Turkey, 23–25 May 2012; Volume 12, pp. 812–817.
14. Joshi, A.; Sharma, V.; Bhattacharyya, P. Harnessing context incongruity for sarcasm detection. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, 26–31 July 2015; pp. 757–762.
15. Oraby, S.; Harrison, V.; Reed, L.; Hernandez, E.; Riloff, E.; Walker, M. Creating and characterizing a diverse corpus of sarcasm in dialogue. *arXiv* **2017**, arXiv:1709.05404.
16. Khodak, M.; Saunshi, N.; Vodrahalli, K. A large self-annotated corpus for sarcasm. *arXiv* **2017**, arXiv:1704.05579.
17. Schuster, T.; Ram, O.; Barzilay, R.; Globerson, A. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *arXiv* **2019**, arXiv:1902.09492.
18. Pant, K.; Dadu, T. Cross-lingual inductive transfer to detect offensive language. *arXiv* **2020**, arXiv:2007.03771.
19. Taghizadeh, N.; Faili, H. Cross-lingual transfer learning for relation extraction using universal dependencies. *Comput. Speech Lang.* **2022**, *71*, 101265.
20. Pires, T.; Schlinger, E.; Garrette, D. How multilingual is multilingual BERT? *arXiv* **2019**, arXiv:1906.01502.
21. Lample, G.; Conneau, A. Cross-lingual language model pretraining. *arXiv* **2019**, arXiv:1901.07291.
22. 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.
23. Raja, E.; Soni, B.; Borgohain, S.K. Fake news detection in Dravidian languages using transfer learning with adaptive finetuning. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106877.
24. Kumar, A.; Albuquerque, V.H.C. Sentiment analysis using XLM-R transformer and zero-shot transfer learning on resource-poor Indian language. *Trans. Asian-Low-Resour. Lang. Inf. Process.* **2021**, *20*, 1–13.
25. Schick, T.; Schütze, H. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv* **2020**, arXiv:2001.07676.
26. Shin, T.; Razeghi, Y.; Logan, R.L., IV; Wallace, E.; Singh, S. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv* **2020**, arXiv:2010.15980.
27. Huang, L.; Ma, S.; Zhang, D.; Wei, F.; Wang, H. Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt. *arXiv* **2022**, arXiv:2202.11451.
28. Qi, K.; Wan, H.; Du, J.; Chen, H. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 1910–1923.
29. Li, Y.; Li, Y.; Zhang, S.; Liu, G.; Chen, Y.; Shang, R.; Jiao, L. An attention-based, context-aware multimodal fusion method for sarcasm detection using inter-modality inconsistency. *Knowl.-Based Syst.* **2024**, *287*, 111457.
30. Liu, H.; Wei, R.; Tu, G.; Lin, J.; Liu, C.; Jiang, D. Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection. *Inf. Fusion* **2024**, *108*, 102353.
31. Veale, T.; Hao, Y. Detecting ironic intent in creative comparisons. In *ECAI 2010*; IOS Press: Amsterdam, The Netherlands, 2010; pp. 765–770.
32. Wang, J.; Zhang, H.; An, T.; Jin, X.; Wang, C.; Zhao, J.; Wang, Z. Effect of vaccine efficacy on vaccination behavior with adaptive perception. *Appl. Math. Comput.* **2024**, *469*, 128543.
33. Hernández-Farías, I.; Benedí, J.M.; Rosso, P. Applying basic features from sentiment analysis for automatic irony detection. In Proceedings of the Pattern Recognition and Image Analysis: 7th Iberian Conference, IbPRIA 2015, Santiago de Compostela, Spain, 17–19 June 2015; Proceedings 7; Springer: Berlin/Heidelberg, Germany, 2015, pp. 337–344.
34. Wang, Y.; Li, Y.; Wang, J.; Lv, H. An optical flow estimation method based on multiscale anisotropic convolution. *Appl. Intell.* **2024**, *54*, 398–413.
35. Zhang, H.; An, T.; Yan, P.; Hu, K.; An, J.; Shi, L.; Zhao, J.; Wang, J. Exploring cooperative evolution with tunable payoff's loners using reinforcement learning. *Chaos Solitons Fractals* **2024**, *178*, 114358.
36. Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; Huang, R. Sarcasm as contrast between a positive sentiment and negative situation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 704–714.
37. Reyes, A.; Rosso, P.; Buscaldi, D. From humor recognition to irony detection: The figurative language of social media. *Data Knowl. Eng.* **2012**, *74*, 1–12.
38. Mukherjee, S.; Bala, P.K. Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering. *Technol. Soc.* **2017**, *48*, 19–27.
39. Poria, S.; Cambria, E.; Hazarika, D.; Vij, P. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv* **2016**, arXiv:1610.08815.
40. Kumar, A.; Narapareddy, V.T.; Srikanth, V.A.; Malapati, A.; Neti, L.B.M. Sarcasm detection using multi-head attention based bidirectional LSTM. *IEEE Access* **2020**, *8*, 6388–6397.
41. Jamil, R.; Ashraf, I.; Rustam, F.; Saad, E.; Mehmood, A.; Choi, G.S. Detecting sarcasm in multi-domain datasets using convolutional neural networks and long short term memory network model. *PeerJ Comput. Sci.* **2021**, *7*, e645.

42. Babanejad, N.; Davoudi, H.; An, A.; Papagelis, M. Affective and contextual embedding for sarcasm detection. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 225–243.
43. Lou, C.; Liang, B.; Gui, L.; He, Y.; Dang, Y.; Xu, R. Affective dependency graph for sarcasm detection. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 11–15 July 2021; pp. 1844–1849.
44. Wang, X.; Dong, Y.; Jin, D.; Li, Y.; Wang, L.; Dang, J. Augmenting affective dependency graph via iterative incongruity graph learning for sarcasm detection. In Proceedings of the AAAI conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2023; Volume 37, pp. 4702–4710.
45. Ren, Y.; Wang, Z.; Peng, Q.; Ji, D. A knowledge-augmented neural network model for sarcasm detection. *Inf. Process. Manag.* **2023**, *60*, 103521.
46. Yu, Z.; Jin, D.; Wang, X.; Li, Y.; Wang, L.; Dang, J. Commonsense knowledge enhanced sentiment dependency graph for sarcasm detection. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macao, China, 19–25 August 2023; pp. 2423–2431.
47. Singh, P.; Lefever, E. Sentiment analysis for hinglish code-mixed tweets by means of cross-lingual word embeddings. In Proceedings of the 4th Workshop on Computational Approaches to Code Switching, Marseille, France, 11–16 May 2020; pp. 45–51.
48. Mao, Z.; Gupta, P.; Wang, P.; Chu, C.; Jaggi, M.; Kurohashi, S. Lightweight cross-lingual sentence representation learning. *arXiv* **2021**, arXiv:2105.13856.
49. Li, I.; Sen, P.; Zhu, H.; Li, Y.; Radev, D. Improving cross-lingual text classification with zero-shot instance-weighting. In Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), Bangkok, Thailand, 6 August 2021; pp. 1–7.
50. Yang, Z.; Cui, Y.; Chen, Z.; Wang, S. Cross-lingual text classification with multilingual distillation and zero-shot-aware training. *arXiv* **2022**, arXiv:2202.13654.
51. Bukhari, S.H.H.; Zubair, A.; Arshad, M.U. Humor detection in english-urdu code-mixed language. In Proceedings of the 2023 3rd International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 22–23 February 2023; pp. 26–31.
52. Ghayoomi, M. Enriching contextualized semantic representation with textual information transmission for COVID-19 fake news detection: A study on English and Persian. *Digit. Scholarsh. Humanit.* **2023**, *38*, 99–110.
53. Ding, K.; Liu, W.; Fang, Y.; Mao, W.; Zhao, Z.; Zhu, T.; Liu, H.; Tian, R.; Chen, Y. A simple and effective method to improve zero-shot cross-lingual transfer learning. *arXiv* **2022**, arXiv:2210.09934.
54. Liu, L.; Xu, D.; Zhao, P.; Zeng, D.D.; Hu, P.J.H.; Zhang, Q.; Luo, Y.; Cao, Z. A cross-lingual transfer learning method for online COVID-19-related hate speech detection. *Expert Syst. Appl.* **2023**, *234*, 121031.
55. Qin, L.; Ni, M.; Zhang, Y.; Che, W. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv* **2020**, arXiv:2006.06402.
56. Zhu, Z.; Cheng, X.; Chen, D.; Huang, Z.; Li, H.; Zou, Y. Mix before align: Towards zero-shot cross-lingual sentiment analysis via soft-mix and multi-view learning. In Proceedings of the INTERSPEECH, Dublin, Ireland, 20–24 August 2023.
57. Lin, H.; Ma, J.; Chen, L.; Yang, Z.; Cheng, M.; Chen, G. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. *arXiv* **2022**, arXiv:2204.08143.
58. Shi, X.; Liu, X.; Xu, C.; Huang, Y.; Chen, F.; Zhu, S. Cross-lingual offensive speech identification with transfer learning for low-resource languages. *Comput. Electr. Eng.* **2022**, *101*, 108005.
59. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
60. Schick, T.; Schütze, H. It's not just size that matters: Small language models are also few-shot learners. *arXiv* **2020**, arXiv:2009.07118.
61. Lin, N.; Fu, Y.; Lin, X.; Zhou, D.; Yang, A.; Jiang, S. Cl-xabsa: Contrastive learning for cross-lingual aspect-based sentiment analysis. *arXiv* **2023**, arXiv:2204.00791 .
62. Misra, R. News headlines dataset for sarcasm detection. *arXiv* **2022**, arXiv:2212.06035.
63. Zhu, Y. Open Chinese Internet Sarcasm Corpus Construction: An Approach. *Front. Comput. Intell. Syst.* **2022**, *2*, 7–9.
64. Xiang, R.; Gao, X.; Long, Y.; Li, A.; Chersoni, E.; Lu, Q.; Huang, C.R. Ciron: A new benchmark dataset for Chinese irony detection. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 5714–5720.
65. Tang, Y.j.; Chen, H.H. Chinese irony corpus construction and ironic structure analysis. In Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014; pp. 1269–1278.
66. Artetxe, M.; Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 597–610.
67. Dauphin, Y.; De Vries, H.; Bengio, Y. Equilibrated adaptive learning rates for non-convex optimization. *arXiv* **2015**, arXiv:1502.04390.
68. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.