Zhegu at SemEval-2023 Task 9: Exponential Penalty Mean Squared Loss for Multilingual Tweet Intimacy Analysis

Pan He[1], **Yanru Zhang**[1,2*]
[1]University of Electronic Science and Technology of China, Chengdu
[2]Shenzhen Institute for Advanced Study, UESTC
newtonysls@gmail.com
yanruzhang@uestc.edu.cn

Abstract

We present the system description of our team Zhegu in SemEval-2023 Task 9 Multilingual Tweet Intimacy Analysis. We propose **EPM** (**E**xponential **P**enalty **M**ean Squared Loss) for the purpose of enhancing the ability of learning difficult samples during the training process. Meanwhile, we also apply several methods (frozen Tuning & contrastive learning based on Language) on the XLM-R multilingual language model for fine-tuning and model ensemble. The results in our experiments provide strong faithful evidence of the effectiveness of our methods. Eventually, we achieved a Pearson score of 0.567 on the test set.

1 Introduction

Text intimacy is considered as the essential component in emotional communication and social relationships (Pei and Jurgens, 2020; Hovy and Yang, 2021; Sullivan, 2013). At the same time, intimacy exists in any language in the world. However, both data sets and methods of this area are still rare. Valuable research on the multilingual regression task is necessary to facilitate the sentiment analysis of language. To further investigate the intimacy in text, Pei et al. propose the MINT, a new Multilingual intimacy analysis dataset. In this case, we describe the model system of our team Zhegu in SemEval-2023 Task 9.

Since the mean of intimacy is different from the "positive" or "negative" in sentiment analysis, it is difficult to have a clear definition. There exist many hard samples, which can not be learned by the model easily. Lin et al. propose Focal loss for learning the hard samples for model only in classification. Therefore, we present **EPM** (**E**xponential **P**enalty **M**ean Squared Loss) for the regression task, which set a penalty term for hard samples during calculating the training loss to accelerate the model focus on those hard samples. We also construct a language-based contrastive loss function to draw closer the distance of the text in the same language and distinguish in the different languages. The test set in this task includes four new languages text data additionally, which do not exist in the training data. In order to achieve better Zero-shot inference on new languages data, we also adopt the frozen tuning to fine tune the model, which means freezing some parameters of the model for training. All in all, the main contributions of our paper are as follows.

- We first propose EPM, which is an exponential penalty mean squared loss function for better learning hard sample in regression.
- We introduce a language-based contrastive loss for the multilingual problem.
- We apply frozen tuning for increasing the capability of model generalization.

All methods we mentioned above will be presented in detail in the Section 3 and will be validated in Section 4. In Section 5, we will give the conclusion of our research and future work.

2 Dataset

The data sets of Task 9 is divided into the labeled set and the unlabeled testing set. Some samples of the labeled set are shown in Table 1. The value of intimacy is range from 1 to 5. The training set contains 9,491 labeled data of six languages, including English (en), Chinese (zh), French (fr), Italian (it), Portuguese (po) and Spanish (sp). There are a total of 13,697 unlabeled data in the test set of ten languages with additional four new languages, including Arabic (ar), Dutch (du), Hindi (hi) and Korean (ko). Details of the data sets can refer to Table 2.

According to Table 2, the amount of text data for each language is relatively balanced to each other

[*]Corresponding author

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 318–323
July 13-14, 2023 ©2023 Association for Computational Linguistics

Text	Language	Intimacy
@user @user Enjoy each new day!	English	1.6
“If you trust them they will always be here for us too”	English	3.0
"Buenas, recién de Desperté después de un turno de noche de que me perdí."	Spanish	3.6
@user @user @user amo você demais minha princesa ♡	Portuguese	4.0

Table 1: The multilingual training set. The score of label is range 1 to 5.

Language	Train	Test
Chinese (zh)	1596	1354
English (en)	1587	1396
French (fr)	1588	1382
Italian (it)	1532	1352
Portuguese (po)	1596	1390
Spanish (sp)	1592	1396
Arabic (ar)	0	1368
Dutch (du)	0	1389
Hindi (hi)	0	1260
Korean (ko)	0	1410
All	9,491	13,697

Table 2: The number of text data of each language in the labeled set and test set.

in both the training set and test set. Meanwhile, the distribution of the number of text data with different labeled values in the labeled set is shown in Figure 1. It is obvious that the distribution shows

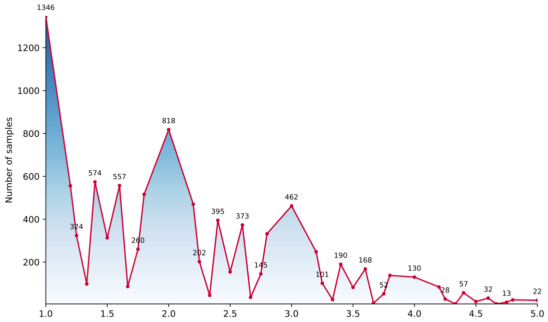

Figure 1: The distribution of the training set according to the label. The number of those data with lower values is far away bigger than those with higher values.

the decreasing trend, principally like the long-tail distribution. The number of those data with 1.0 is close to 1400, which is nearly one hundred times of the value of 5.0.

3 Methodology

In this section, we describe in detail the methods we used on task 9. We first introduce a loss function for regression task we proposed, and we named it **EPM**. After that, we present contrastive loss and frozen tuning. The reasons for their ability to improve model performance will be described in detail. And We take XLM-RoBERTa-large (Conneau et al., 2019) as the backbone of our model.

3.1 EPM

We denote the training data $\mathcal{D} = [(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)], y_i \in [1.0, 5.0]$. N represents the number of the training set. And the model function is

$$\hat{y}_i = f(x_i) = \boldsymbol{w}x_i \tag{1}$$

where $\hat{y}_i$ is the prediction of the model and $\boldsymbol{w}$ is the parameters. The calculation of MSE (Mean Squared Error) loss in regression is

$$\mathcal{L}_{MSE}(y_i, \hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2}$$

The MSE loss function is used to measure the difference between the label and the prediction. And we define our **EPM** loss function as

$$\mathcal{L}_{EPM}(y_i, \hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 e^{|y_i - \hat{y}_i|} \tag{3}$$

EPM adds a penalty term $e^{|y_i - \hat{y}_i|}$ on side of MSE, which can further penalize those samples with huge deviations. EPM will degrade to the standard MSE loss function when $\hat{y}_i$ equals to y_i. The derivative of the MSE loss function with respect to $\boldsymbol{w}$ is

$$\frac{\partial(\mathcal{L}_{MSE})}{\partial \boldsymbol{w}} = \frac{2}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)x_i \tag{4}$$

The derivative of the EMP loss function with respect to $\boldsymbol{w}$ is

$$\frac{\partial \mathcal{L}_{EPM}}{\partial \boldsymbol{w}} = \frac{1}{N}\sum_{i=1}^{N} e^{|\hat{y}_i - y_i|}[2(\hat{y}_i - y_i) + (\hat{y}_i - y_i)^2]x_i \tag{5}$$

$$= \frac{\partial(\mathcal{L}_{MSE})}{\partial w} e^{|\hat{y}_i - y_i|} + \mathcal{L}_{EPM}x_i, \hat{y}_i \geq y_i \tag{6}$$

$$\frac{\partial \mathcal{L}_{EPM}}{\partial \boldsymbol{w}} = \frac{1}{N}\sum_{i=1}^{N} e^{|\hat{y_i}-y_i|}[2(\hat{y_i}-y_i)-(\hat{y_i}-y_i)^2]x_i \tag{7}$$

$$= \frac{\partial(\mathcal{L}_{MSE})}{\partial w} e^{|\hat{y_i}-y_i|} - \mathcal{L}_{EPM} x_i, \hat{y_i} < y_i \tag{8}$$

The value of Equation 5 equals to Equation 7 when $\hat{y_i}$ equals to y_i. We can easily know that the derivative function of EPM is continuous by referred to Equation A. The EPM derivative function reveals the $e^{|y_i-\hat{y_i}|}$ penalty term amplifies the gradient isometrically during the process of back-propagation. And $(\hat{y_i}-y_i)^2$ penalizes the gradient once again. The larger $|y_i-\hat{y_i}|$, the larger the penalty for that one hard sample. Focal loss (Lin et al., 2017) also has capacity to penalize hard samples for classification tasks. Nevertheless, EPM does not reduce the attention of the learning of simple samples, while Focal loss does. For those easy samples that they could be easy to learn by the model, $\hat{y_i}$ converges to y_i so much that $\mathcal{L}_{EPM}$ highly approximates to $\mathcal{L}_{MSE}$.

3.2 Contrastive Loss

Contrastive learning can help model to distinguish the negative samples and converge the positive samples by constructing appropriate positive and negative samples. In this section, we introduce a language-based contrastive loss.

In contrastive learning, the construction of positive and negative samples is crucial. According to the analysis of data in the previous Section 2, the labeled set contains text data from six languages. For the multilingual task, we take XLM-R (Conneau et al., 2019) as the backbone of our model. We believe that the distance between those text data from the same language should be closer than those from different languages. Based on the above analysis, the construction of the positive and negative samples of contrastive learning in our paper are based on the languages. The process of contrastive learning for a batch of training data is shown in Figure 2.

Let v_i be the sentence vector obtained by inputting x_i to the model, l_i denote the language of $x_i, l_i \in \{zh, en, fr, it, po, sp\}$. For a batch of data $(x_1, x_2, ..., x_M)$, $(v_1, v_2, ..., v_M)$ are the sentence vectors. M is batch size. Then KL (Kullback-Leibler) divergence is calculated between every two samples to get the matrix $\boldsymbol{KL} \in \mathbb{R}^{M\times M}$. The reason of adopting KL divergence rather than cosine similarity is that cosine similarity usually is used to calculate the similarity of the meaning between the text, while KL divergence obtains the distance of two distributions. KL divergence is regarded as the metric appropriately because the different texts in the same language do not contain the same meaning. Denote the contrastive learning loss function as $\mathcal{L}_{CL}$.

$$\mathcal{L}_{CL} = -\frac{1}{2}\ln\frac{\sum_{l_i=lj} e^{KL[i][j]/\tau}}{\sum_{l_i\neq lj} e^{KL[i][j]/\tau}}, i,j \in [1, M] \tag{9}$$

where τ is temperature hyper-parameter. The final loss function of the model is the sum of EPM loss and contrastive loss. The model learns the objective function of regression while converging the distribution of the text data from the same language and distinguishing from the different languages.

$$\mathcal{L} = \mathcal{L}_{EMP} + \mathcal{L}_{CL} \tag{10}$$

3.3 Frozen Tuning

The test set contains text data from four languages additionally, which the training set do not have. The model of this task is required to have the capability of generalization for inferring the new language text data as possible. During the training process, the parameters of the model are fitted to the training set containing only six languages. To enhance the generalization of our model on new language text data, the Frozen Tuning is used in our paper. We named the parameters of the standard XLM-R **General parameters**, which is pre-trained on the general pre-trained corpus. And we also named the parameters of fine-tuned XLM-R with the training set **Domain parameters**. Frozen Tuning freezes some of the parameters of the XLM-R during fitting in the training set, as shown in Figure 3. The languages of the pre-trained corpus of XLM-R contain all the languages in the training set and the test set. The frozen parameters in the pre-training model remain the same generalization capacity for all the languages, while the fine-tuned parameters can be considered as the classifier learned on the training set. Meanwhile, the updatable parameters of the model during the training process can be reduced by frozen tuning to improve the generalization performance.

4 Experiments & Results

In this section, we describe the experimental procedures and results in detail. The evidence of those methods we mentioned above will be provided by results and analysis.

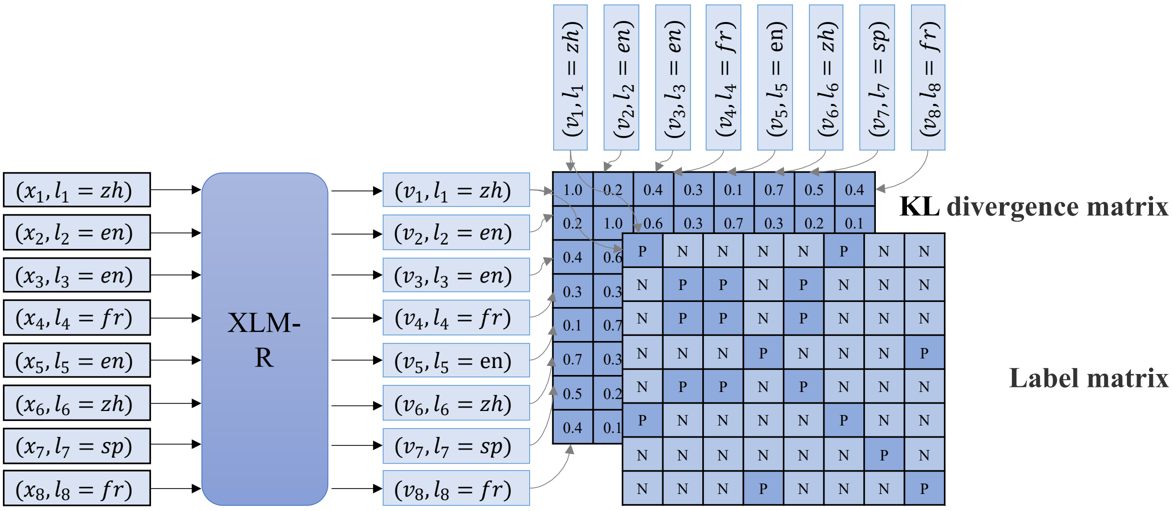

Figure 2: The procedure of contrastive learning based on languages during training process for a batch of data, while the batch size M is 8. **P** and **N** represent the positive and negative sample in the label matrix.

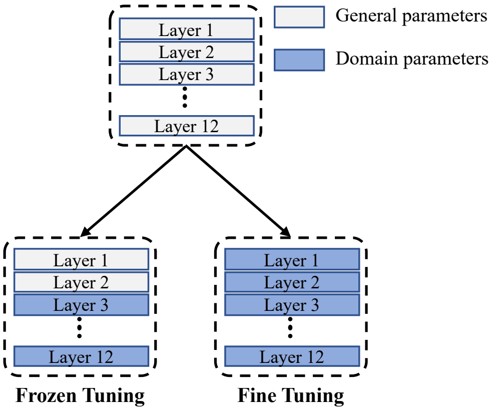

Figure 3: The process of frozen tuning. The number of hidden layers is 12 in this figure.

4.1 Data Split

The labeled data and the unlabeled test set are provided in Task 9. Two strategies of dividing the validation set for model selection were adopted in our experiments.

Dev1 : We directly sample one-fifth of the labeled data as the validation set and the remaining data as the training set.

Dev2 : In order to simulate the distribution of the test set as much as possible, the validation set should also contain unseen languages' data. Therefore, we first sample two languages text data from the labeled data containing six languages. Subsequently, we concatenate one-fifth of the remaining four languages' data and two unseen languages data as the final validation set.

4.2 Performance & Analysis

All experiments in this paper were conducted on the XLM-R pre-trained languages model. The detailed hyperparameters can be found in Table 4 in the Appendix. The final results are shown in Table 3. Figure 4 shows the PCC (Pearson correlation coefficient) of the training set with using different loss function. For fair comparison, we used the MSE loss function as the reference group. With the help of the EPM, the model could converge faster than MSE on the training data. Meanwhile, the fluctuation of PCC of EPM in the optimization process is much smoother than MSE. And the results of the validation sets in Table 3 show that EPM not only let the model converge faster, but also can obtain better performances than MSE.

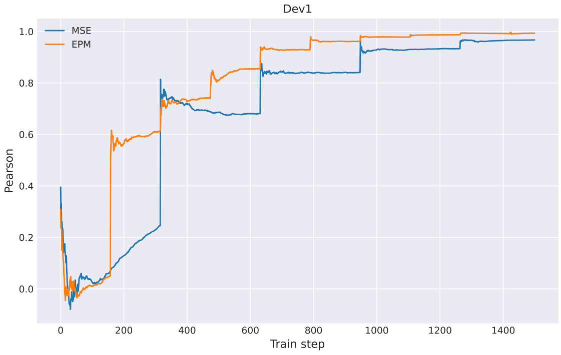

Figure 4: The PCC of the training data during the training process when we set the EPM and MSE as the loss function of the model.

Figure 5 presents the PCC of the training data using EPM and MSE loss functions respectively

Method	Dev1	Dev2	Test	
			Seen Languages	Unseen Languages
XLM-R + MSE	68.66	69.13	67.26	37.39
+ EPM	68.86	70.47	69.61	44.85
+ EPM + CL	69.15	71.36	70.02	46.86
+ EPM + FT	68.98	70.22	70.04	45.98
+ EPM + CL + FT	68.59	69.63	69.81	45.14
Baseline			65.33	41.25
Ensemble			**72.00**	**46.92**

Table 3: The PCC of dev and test datasets. We report the performance of baseline from (Pei et al., 2023) directly. **CL**: contrastive learning. **FT**: Frozen Tuning. The bolded and underlined: the SOTA results of all methods. The underlined: the second best.

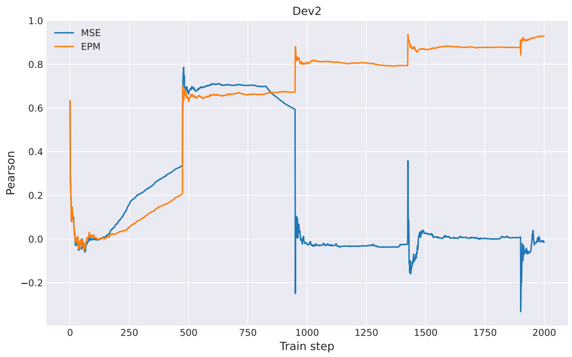

Figure 5: The PCC of the training set when we conduct contrastive learning task with using EPM and MSE loss functions respectively.

when we conduct contrastive learning. The model also converges normally when we combine EPM and language-based contrastive learning. Although slightly slower optimization in the early stage of training, the model still can converge on the training set. The reason of slower optimization is that the model optimizes both $\mathcal{L}_{EPM}$ and $\mathcal{L}_{CL}$ at the same time. Nevertheless, the PCC of the training set suddenly drops or even becomes negative in the middle of the training process when the final loss is $\mathcal{L}_{MSE} + \mathcal{L}_{CL}$. This phenomenon is caused by the gradient disappearance during the back-propagation process. In the training process, the regression and the contrastive learning task are optimized at the same time. And it may make two gradients of $\mathcal{L}_{MSE}$ and $\mathcal{L}_{CL}$ add up to a small amount and cause disappearing of the gradient. However, the penalty term of EPM could amplify the gradient, which is able to avoid the gradient vanishing.

Table 3 presents EPM brings improvement both on the validation and test sets. And the improvement on the seen languages and unseen languages are nearly 2% and 7%. Meanwhile, the contrast learning based on languages outperforms other methods on zero-shot prediction of the unseen languages, which obtains the promotion of 2% on the basis of EPM. Although the effectiveness is not much as both EPM and contrastive learning, Frozen tuning also brings a surprising improvement of 1% in the unseen languages' data, which may cause by the reduction of the training parameters. It is difficult to optimize $\mathcal{L}_{EPM}$ and $\mathcal{L}_{CL}$ simultaneously when freeze some layers of the pre-trained model. Finally, the best result is obtained by multi excellent models ensemble, which selected by the validation set. The decent improvement is 7% in the seen languages and 5% improvement of the unseen languages compared to baseline. The results strongly provide faithful evidence to our methods.

5 Conclusion

We introduce the system description of our team Zhegu in SemEval-2023 Task 9. We first propose a loss function, we named **EMP** (**E**xponential **P**enalty **M**ean Squared Loss), which could increase the gradient of those hard samples by adding the penalty term. EPM could be used in any regression task. And we conduct a contrastive learning task based on languages to reduce the distance of those in the text from the same languages. The experiments prove the effectiveness of our proposed methods. Meanwhile, EPM could avoid the disappearance of gradient than MSE when we conduct contrastive learning. Frozen tuning is also useful for improving the performance. In future work, the effectiveness of EPM should also be proved in others natural language regression tasks.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. *arXiv preprint arXiv:2011.03020*.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozon Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2023. Semeval 2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Harry Stack Sullivan. 2013. *The interpersonal theory of psychiatry*. Routledge.

A Appendix

$$lim_{\hat{y}_i \to y_i^+} \frac{1}{N} \sum_{i=1}^{N} e^{|\hat{y}_i - y_i|} [2(\hat{y}_i - y_i) + (\hat{y}_i - y_i)^2] x_i = lim_{\hat{y}_i \to y_i^-} \frac{1}{N} \sum_{i=1}^{N} e^{|\hat{y}_i - y_i|} [2(\hat{y}_i - y_i) - (\hat{y}_i - y_i)^2] x_i$$

Number of freeze layer	3
Max sequence length	128
Batch size	64
Epochs	11
Optimizer	AdamW
Adam epsilon	1e-6
Weight decay	0.1
Scheduler	linear warmup
Warmup rate	0.1
Max grad norm	1
τ	0.2
XLM-R learning rate	2e-5
Classifier learning rate	1e-4

Table 4: The hyperparameters of various strategies for model training.