Sea_and_Wine at SemEval-2023 Task 9: A Regression Model with Data Augmentation for Multilingual Intimacy Analysis

Yuxi Chen[1], **Yu Chang**[1], **Yanqing Tao**[1], **and Yanru Zhang**[1,2]

[1]University of Electronic Science and Technology of China(UESTC), China
[2]Shenzhen Institute of Advanced Study, UESTC, China
[1]{yuxi.ch, yuchang}@std.uestc.edu.cn yanruzhang@uestc.edu.cn

Abstract

In Task 9, we are required to analyze the textual intimacy of tweets in 10 languages. We fine-tune XLM-RoBERTa (XLM-R) pretrained model to adapt to this multilingual regression task. After tentative experiments, severe class imbalance is observed in the official released dataset, which may compromise the convergence and weaken the model effect. To tackle such challenge, we take measures in two aspects. On the one hand, we implement data augmentation through machine translation to enlarge the scale of classes with fewer samples. On the other hand, we introduce focal mean square error (MSE) loss to emphasize the contributions of hard samples to total loss, thus further mitigating the impact of class imbalance on model effect. Extensive experiments demonstrate remarkable effectiveness of our strategies, and our model achieves high performance on the Pearson's correlation coefficient (CC) almost above 0.85 on validation dataset.

1 Introduction

The concept of Intimacy dates from social psychology, which is of great value to indicate degree of closeness in the relationships between people. And analysis of social intimacy is beneficial to unveil complicated mechanisms of social interaction. For another, as verbal and textual communications are universally a powerful means to express the affection and nurture social relationships, language plays an indispensable role in unraveling the intimacy between individuals. Thus analyzing textual contents of social media has a potential to unearth the intimacy information in the field of natural language processing (NLP).

Although feasible the task looks like, it faces two core challenges, how to quantify the intimacy of language and how to gauge it according to textual information by computational modelling. To tackle such challenges, Pei and Jurgens (2020) first proposed a computational framework to measure the intimacy of language. However, despite demonstrating the construction of textual intimacy, this work was only based on the form of questions. And there are still gaps with the real situations that contain other forms of language, such as statements and dialogues between people. To enlarge the span of applications and improve the generalization performance of language models, htt proposed a multilingual textual intimacy dataset covering tweets in 10 languages, including English, Spanish, Italian, Portuguese, French, Chinese, as well as Hindi, Arabic, Dutch and Korean, which has been set as the official dataset in SemEval-2023: Task 9 (Pei et al., 2023a,b). Overall, works on textual intimacy remain scarce. In Task 9, we are required to predict the intimacy of textual tweets on the official released dataset. So far, there are several multilingual pre-trained language models, including M-BERT (Pires et al., 2019), XLM-R (Conneau et al., 2019), ERINE-M (Ouyang et al., 2020), which can be utilized to handle multilingual regression tasks.

In our work, we adopt XLM-R (Conneau et al., 2019) model to deal with the task. And its contributions can be summarized as follows: 1) Aiming at a significant challenge of class imbalance in the raw training dataset, we utilize a data augmentation technique fulfilled by machine translation to augment samples of the minority. 2) We introduce a new loss function, focal MSE loss, to the optimization process to further weaken the impact of the majority samples. 3) Experiments demonstrate the effectiveness of our strategies, and our model achieves high performance on Pearson's CC almost above 0.85 on validation dataset. On the official test set, our model gets an average score of 0.65 on 6 languages included in the training set, which ranking 31st overall on the leaderboard, and gets an average score of 0.38 on 4 new languages that are not appearing in the training set, which ranking 27th overall.

text	label	language
Bees vs. Wasps. Http	1.0	English
Here is a nice equation: 0+0-0-0+0=0	1.0	English
@user 真的超级难，欲哭无泪	3.25	Chinese
@liu_xiaoyuan 红外测温仪在室外误差极大。	1.4	Chinese
@user @user Surtout maintenant que ça a été médiatisé	1.6	French
@felipevinha épi	1.5	French
@user Não tô dizendo que tá certo esse pensamento btw	2.0	Portuguese
dios bendiga a how to sell drugs online xq es la única serie q no es un anime q hemos podido ver juntos julián y yo xq no nos gusta nada	3.4	Spanish
@user @capuanogio L'ho criticato spesso...stavolta no	1.4	Italian
@N_ShaniJKT48 Ha	2.0	Italian

Table 1: Details of dataset

2 Background

2.1 Dataset Description

In the official released training dataset (Pei et al., 2023b), there are 9491 textual tweets in 6 languages, including English, Spanish, Italian, Portuguese, French, and Chinese. And they are all sampled from dialogues on social media Tweet from 2018 to 2021 and annotated by scores ranging from 1 to 5, where a higher score implicates more intimacy between users. Then we are required to predict the intimacy scores according to series of given unlabeled texts. Details of the dataset are shown in Table 1

2.2 Related work

2.2.1 Class Imbalance

In the dataset of Task 9, there is a serious problem of class imbalance observed in our tentative experiments, where one class has many more samples than the other. And samples of the majority class are referred to as easy samples, while another ones are called hard samples. Due to that hard samples are extremely few in numbers, easy samples overwhelmingly dominate the gradient updating during optimization, which may keep the gradient heavily skewed towards the majority class and away from the optimal solution. Generally speaking, as a key point to this phenomenon is mainly about the distribution of the class, methods of enlarging the scale of the hard samples through data augmentation should have the highest priority to be conducted intuitively. For another, emphasizing the contributions of hard samples to the total loss by modifying structure of the loss function proves to be an effective way to improve the performance of models (Lin et al., 2017). Therefore, in our work, we take more efforts in this two aspects to mitigate the negative impact of class imbalance on training models.

2.2.2 Data Augmentation in NLP

Data augmentation has been widely accepted to enlarge the scale of datasets and meanwhile improve theirs quality. Current solutions include data noising techniques, e.g. easy data augmentation (EDA) (Wei and Zou, 2019), back translation(BT) (Sennrich et al., 2015), and adversarial training (Kusner and Hernández-Lobato, 2016; Zhang et al., 2016), etc. Although lots of methods of data augmentation, they do not perform quite well when applied in NLP tasks, and some challenges remain noteworthy. On the one hand, natural language data itself contains semantic information, which may be destroyed if the structure of the text is changed, while simpler operations, e.g. replacement of synonyms, may bring overfitting of models during training process to some extent. On the other hand, metrics to appropriately reflect the effectiveness of the generated language data are hard to specify.

In our work, noticing the particularity of multilingual data, to preserve the semantic information of original sentences and simultaneously maintain theirs distribution as best, we employ the data augmentation techniques through machine translation to translate some texts into other languages, thus mitigating the class imbalance. More details of the implementation will be delineated in section 3.

2.2.3 Focal Loss

Focal loss was first proposed by Lin et al. (2017) to address the challenge of class imbalance. Its main idea is to weaken the contributions of easy samples to the total loss to achieve an equilibrium between easy samples and hard samples, as shown in Eq. 1

as follows:

$$L = \sum_i -(1-p_i)^{\gamma} log(p_i) \tag{1}$$

where p_i represents predicted probability of the ith sample. The coefficient $(1-p_i)^{\gamma}$, named focal factor, can be seen as a decay weight of the loss of each sample, where γ is a focusing parameter to adjust the the rate that how easy samples are down-weighted. For an easy sample, its predicted probability p_i is clearly large, and $(1-p_i)$ must be a small value approaching to 0. When $\gamma = 0$, the focal loss is the same as cross entropy (CE) loss. And as γ increases ranging above 1, the focal factor becomes further small due to that the value of $(1-p_i)$ is not larger than 1. In this way, easy samples contribute less to the total loss than before, which in turn raises the impact of hard samples. Inspired by this original focal loss, in our work, we propose focal MSE loss modified from the MSE loss to prompt the hard samples to stand out.

3 System Overview

3.1 Model Structure

In our work, we fine-tune XLM-R, a pre-trained multilingual RoBERTa model, to adapt to the specified regression tasks. Specifically, data augmentation is conducted based on the original dataset at first. Due to that the process of translation in the data augmentation may cause some unpredictable string, converting a punctuation mark into "&39;" for example, we then conduct some data cleansing on the augmented texts. Thus the final input of the model is a sequence of word or sub-word preprocessed from the augmented dataset, and the output is a sequence of theirs embedding vectors, as well as the predicted values of intimacy. It is also worth noting that we add a LayerNorm layer between the Pooler layer and the fully-connected layer during the training of cross validation to avoid a potential problem of vanishing gradient. The overall framework is demonstrated as Fig. 1.

3.2 Data Augmentation

When conducting tentative experiments on the raw dataset, serious class imbalance is observed to exert negative effects on training the model. As shown in Fig. 2(a), we can see that samples with labels above 3.7 seem quite few in number, thus causing significant uneven distribution. To mitigate the influence from such distribution, we take measures to augment the dataset through machine translation, provided by Google translator. Details of the experimental setting are delineated in Appendix A.1. Then the distribution of the augmented dataset is shown in Fig. 2(b), in which the imbalance has been remarkably alleviated and at the same time the distribution of 6 languages is maintained as evenly as possible, demonstrated in Fig. 3. Besides, considering that there are 4 new languages (Hindi, Arabic, Dutch and Korean) that are not included in the training dataset, and to simulate specific distribution in the test dataset better, we also translate some items into this 4 languages. In the end, there are 22279 items in all to be put into training.

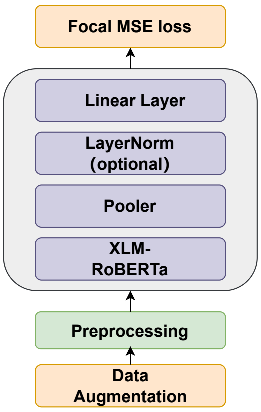


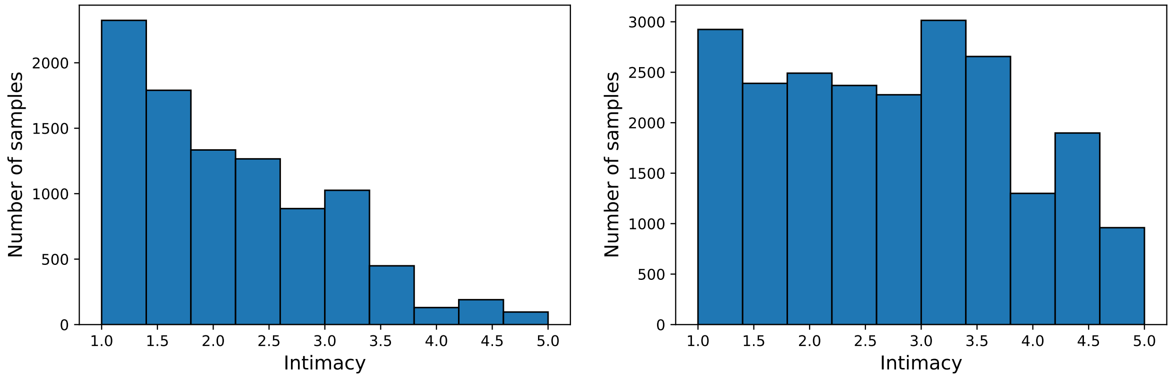
Figure 1: The overall framework of our model. Before fed into the model, data is preprocessed including cleansing in advance. To mitigate the class imbalance, part of the classes are augmented through machine translation. Then XLM-R model absorbs the data. After a series of calculating, focal MSE loss is utilized in optimization process to reduce the prediction error.

3.3 Focal MSE

When calculating MSE loss during optimization, although the loss value of a single hard sample is large, the accumulation of all of them causes slight effect on the total loss because of their small quantities, thus leading to the overwhelming predominance of the easy samples. In our work, we design a loss function based on the original MSE loss to raise the impact of hard samples on the total loss. We introduce a focal factor, defined as Eq. 2, to adjust the contributions of hard samples.

$$\alpha_i = \begin{cases} (-sim(\hat{y}_i, y_i))^{\gamma}, & sim(\hat{y}_i, y_i) < 0 \\ 1, & sim(\hat{y}_i, y_i) \geq 0 \end{cases} \tag{2}$$

$$MSE(\hat{y}_i, y_i) = \frac{1}{k} \sum_i |\hat{y}_i - y_i|^2 . \tag{3}$$



(a) The distribution of classes on the raw data.

(b) The distribution of classes after data augmentation.

Figure 2: (a) There is an obvious class imbalance observed from the distribution, where class with labels above 3.7 appears extremely scarce. (b) After implementing data augmentation on the raw data, the class imbalance has been apparently eased.

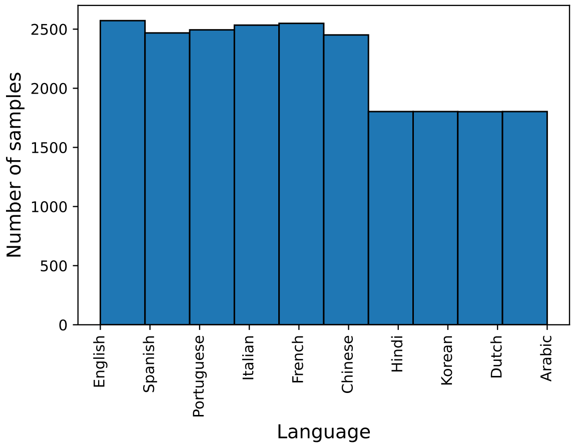


Figure 3: The distribution of languages in the dataset with data augmentation. In the dataset, 4 languages with the least number of samples are totally generated. And during data augmentation, the distribution is maintained as even as possible.

In the Eq. 2, $sim(\cdot)$ is a cosine similarity function, which measures the similarity between two vectors. If the value is positive, then it means that the two vectors are of high similarity. $\hat{y_i}$ and y_i represent the prediction and the label of the ith sample, respectively. And γ, an odd number greater than 1, is a modulating factor to regulate the degree of amplification of the discrepancy. Suppose that there are k samples in all, according to the well-known MSE loss function described as Eq. 3, our focal MSE loss function can be defined as Eq. 4 as follows:

$$L = \alpha MSE(\hat{y_i}, y_i). \tag{4}$$

In the focal factor α_i, if the value of the cosine similarity is less than 0, then their must be huge discrepancy between the predictions and the labels. That is to say, the samples are more likely hard samples. Then we take the opposite as a penalty to amplify the original MSE loss. And as the modulating factor γ ranges in value above 1, the final loss L can be further enlarged. On the contrary, if the cosine similarity is 0 or more, which means that the discrepancy between the two vectors appears not so far, then we take the focal factor as 1 to set the final loss L as MSE itself. As a result, our focal MSE loss works only on the hard samples.

4 Experimental Setup

To validate the effectiveness of our model and strategies, we conduct extensive experiments both on the raw data and the dataset with data augmentation. During the training process, we make attempts at 5-fold and 10-fold cross validations, as well as training without cross validation, respectively. After K-fold training, the model with the best score of the K models will be trained again with dataset including training set and validation set, and examined by the test dataset split in advance to get a final score. When it comes to the loss function, we also compare the performances of the original MSE loss and our focal MSE loss. More details are delineated as follows.

4.1 Dataset Split

In the situations without cross validation, we randomly pick up 2 from the 6 languages in the raw training dataset and take theirs corresponding items as validation dataset, leaving the rest in the augmented dataset as the training dataset. And during training, items in a batch are randomly sampled in the training set.

When using the strategy of K-fold cross validation, we split the augmented dataset in proportion of $8:2$, where the part with lesser data are set aside as the final test dataset. Subsequently, the

Validation	Model(XLM-R)	Pearson's CC	Loss
w/o cross validation	+MSE loss	0.576	0.88
	+Focal MSE	0.675	0.51

Table 2: Performances on the raw dataset

Validation	Model(XLM-R)	Pearson's CC	Loss
w/o cross validation	+Focal MSE	0.858 (on Chinese & Arabic)	0.33
		0.885 (on English & Portuguese)	0.34
		0.748 (on Spanish & French)	0.51
5-fold cross validation		0.855	0.36
10-fold cross validation		0.861	0.35

Table 3: Performances on the dataset with data augmentation

rest is further split into training set and validation set by StratifiedKFold strategy. Considering that such strategy is only suitable for classification tasks, there is a preprocessing of encoding target labels with integers ranging above 0, to simulate the operation of classification tasks.

4.2 Network Initialization

We initialize the XLM-R model with hyper-parameters shown in Table 4 in Appendix. It is worth noting that the focal MSE loss introduces a new hyper-parameter γ. Experiments and theoretical analysis show that the model will get higher performance when γ ranges in value above 1. Then in our model, we choose 2 as its value.

4.3 Metrics

Following the requirements of Task 9, we choose Pearson's CC as the metrics to evaluate our trained model. And at the same time, during the training process we also retain the indicator MSE loss between the predictions and the labels, which demonstrates the model effect more intuitively and is beneficial for us to recognize more potential matters to be settled.

5 Results

As shown in Table 2, the Pearson's CC goes up remarkably when our focal MSE loss takes the place of the original MSE loss, which indicates the effectiveness of such modification on loss function. Likewise, more conspicuous rise on the performance is observed after data augmentation, as shown in Table 2 and 3.

In addition to following the superficial scores, the results can also unearth that when the training occurs under the conditions of no use of cross validation, due to that languages in the validation set are never seen in the training set, the corresponding scores can reflect zero-shot transfer performance of the model on different languages, which gauge its task-learning capability to some extent. And we can infer that such capability varies in languages, as is indicated by a higher Pearson's CC on English and Portuguese than on Chinese and Arabic. For the ways of K-fold cross validation, we hold that although the scores are not far off from that of the ways without cross validation, it may reflect the model effect in actual situations better, as theirs corresponding strategies of the data partition are based on distributions rather than language categories.

6 Conclusion

In Task 9, we fine-tune XLM-R pre-trained model to figure out the multilingual regression task of predicting the textual intimacy. We recognize severe class imbalance in the raw dataset as a primary obstacle that impede the attempts of the model to achieve high performance. To tackle this challenge, we conduct data augmentaion through machine translation provided by Google translator to enlarge the scale of the minority classes. Meanwhile, we introduce a new loss function, focal MSE loss, to further punch above the weight of hard samples on the total loss. And extensive experiments demonstrate the effectiveness of our model and strategies by high performance on Pearson's CC. However, there is no denying that overfitting on the augmented dataset still exist, and we intend to explore more solutions to such limitation in the future.

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. *CoRR*, abs/1911.02116.

Matt J Kusner and José Miguel Hernández-Lobato. 2016. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora.

Jiaxin Pei, Francesco Barbieri, Vítor Silva, Maarten Bos, Yozen Liu, Leonardo Neves, and David Jurgens. 2023a. Semeval 2023 task 9: Multilingual tweet intimacy analysis. `https://sites.google.com/umich.edu/semeval-2023-tweet-intimacy/home`.

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. 2023b. Semeval 2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Yizhe Zhang, Zhe Gan, and Lawrence Carin. 2016. Generating text via adversarial training. In *NIPS workshop on Adversarial Training*, volume 21, pages 21–32.

A Appendix

A.1 Details of Data Augmentation

To mitigate the class imbalance of the training set, we utilize Google translator to translate some items into languages with few samples. Specifically, we select all the samples with labels above 3.7, to be translated into another 5 languages in the training set. For the samples with labels between 1.8 and 3.7, we select 1700 items, including 400 items with labels from 1.8 to 2.6 and 3.4 to 3.8, respectively, 550 items with labels from 2.6 to 3.0, and 350 items with labels from 3 to 3.4, and translate them into other randomly picked 2 languages. Besides, in order to simulate the distribution of the test set with 4 new languages, we also randomly pick out 1700 items, including 150 items with labels from 1 to 1.4, 1.4 to 1.8, and 3 to 3.4, respectively, 400 items with labels from 1.8 to 2.6, 200 items from 2.6 to 3 and 3.4 to 3.8, respectively, and all items with labels above 3.8, into this 4 languages.

A.2 Hyperparameters

Table 4 provides the details of initial hyperparameters of our model.

Hyperparameters	Value/Range
Bert seq length	128
Bert learning rate	4e-5
Learning rate	1e-4
Batch size	16
Max epochs	9-11
Class label	1
Gamma of Focal MSE	above 1

Table 4: Main initial hyperparameters