ZBL2W at SemEval-2023 Task 9: A Multilingual Fine-tuning Model with Data Augmentation for Tweet Intimacy Analysis

Hao Zhang[*], Youlin Wu[*], Junyu Lu, Zewen Bai, Jiangming Wu, Hongfei Lin, Shaowu Zhang[†]
School of Computer Science and Technology, Dalian University of Technology, China
{zh373911345,wuyoulin,dutljy,986427968,
wjm374733751}@mail.dlut.edu.cn
{hflin,zhangsw}@dlut.edu.cn

Abstract

This paper describes our system used in the SemEval-2023 Task 9 Multilingual Tweet Intimacy Analysis. There are two key challenges in this task: the complexity of multilingual and zero-shot cross-lingual learning, and the difficulty of semantic mining of tweet intimacy. To solve the above problems, our system extracts contextual representations from the pretrained language models, XLM-T, and employs various optimization methods, including adversarial training, data augmentation, ordinal regression loss and special training strategy. Our system ranked 14th out of 54 participating teams on the leaderboard and ranked 10th on predicting languages not in the training data. Our code is available on Github [1].

1 Introduction

A theoretical approach to understanding the level of intimacy, human emotion, and social connections in communication is called intimacy in language. Quantifying the intimacy expressed in language plays a key role in revealing important social norms in various context (Pei and Jurgens, 2020). SemEval-2023 Task 9 aims at mining intimacy in tweets. Pei et al. (2023) provides a new multilingual intimacy analysis dataset covering 13,372 tweets in 10 languages. Text intimacy is labeled on a continuum scale, in the range of [1, 5]. Among them, the participants are only given the training set in six languages (English, Spanish, Italian, Portuguese, French, and Chinese), and the model performance will be evaluated on the test set of ten languages, including 4 languages not in the training set (Hindi, Arabic, Dutch and Korean).

Given that the SemEval-2023 Task 9 is a complex challenge that integrates language transfer and semantic mining, we propose the intimacy prediction system based on pre-trained representations and the fusion of multiple optimization methods. First, we chose XLM-T (Barbieri et al., 2021) as the encoder of the model, which is pre-trained on millions of tweets in multiple languages. Then, according to the form of annotated dataset, we change the calculation method of model loss and adjust the training strategy of the model to make it more suitable for the intimacy prediction of zero-shot language. Finally, we introduce adversarial training and data augmentation to enhance the model generalization. Our system obtained Pearson correlation scores of 0.7029 on seen language section and 0.4359 on unseen language section.

2 Related Work

Intimacy computing is an emerging field of natural language processing. Pei et al. (2023) propose to encode social messages of intimacy through topics and other more subtle cues such as swear words. A new computational framework for studying the expression of intimacy in language is introduced, accompanied by multilingual datasets, MINT, and deep learning models.

With the development of pre-training technology, multi-lingual pre-train model based methods are becoming the new trend to solve cross-lingual tasks. Conneau and Lample (2019) extend generative pre-training to multiple languages and show the effectiveness of cross-lingual pre-training for the first time. Knowledge distillation methods (Hinton et al., 2015) have been widely used in cross-language model research, but DistillBert (Sanh et al., 2019) and MiniLM (Wang et al., 2020) do not perform well for intimacy tasks. Our work builds on the work of Barbieri et al. (2021), which takes into account the unique expression of tweets and the generalization of multiple languages.

[*] These authors contributed equally to this work.

[†] Corresponding author.

[1] https://github.com/HelloHank-Hub/Multilingual-Tweet-Intimacy-Analysis

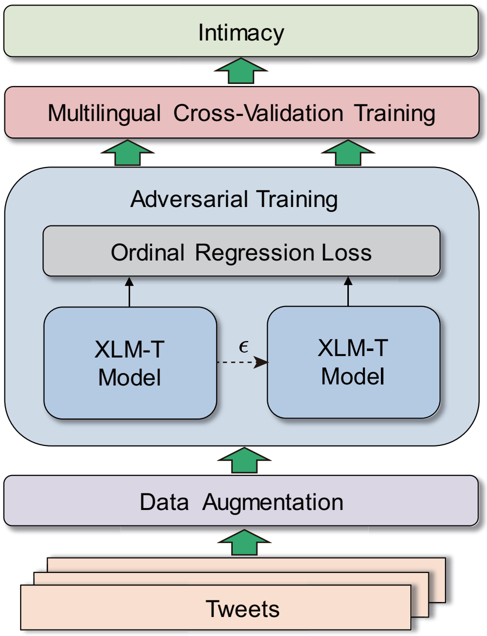

Figure 1: The overall architecture of our proposed system in SemEval-2023 Task 9.

3 Methodology

3.1 System Architecture

The overall architecture of our final system is illustrated in Figure 1. In order to mine the semantic features related to intimacy in multilingual texts, we chose XLM-T as the encoder of the whole system. The input to the XLM-T was randomly obtained from tweets in different languages, and the last hidden state of the output was converted into a semantic pooling vector through the pooling layer. Then, the pooled vector and the output `[CLS]` vector were concatenated and fed to the feedforward network consisting of two layers of linear networks to predict the final results. Finally, we trained the model using the ordinal regression loss.

3.2 Adversarial Training

Adversarial training (Goodfellow et al., 2015) is an efficient regularization technique for classifiers to increase robustness to small, approximately worst-case perturbations. In Task 9, we introduced the Fast Gradient Method (FGM) (Miyato et al., 2016), a novel method for adversarial training, enhancing the generalization of the model in the intimacy analysis.

According to FGM, we applied tiny perturbations to sentence embeddings, and the adversarial perturbation r_{adv} on s is defined as:

$$r_{adv} = \epsilon \cdot g/\|g\|_2 \; where \; g = \nabla_s L(s, y) \quad (1)$$

where ϵ is a hyperparameter controlling the strength of the adversarial perturbations.

We employed an overall loss function to combine the information learned from the original and adversarial samples:

$$L = L(s, y) + L_{adv}(s + r_{adv}, y) \quad (2)$$

3.3 Ordinal Regression

Though intimacy in MINT is represented as a real number after post-processing, the original annotation comes from Likert Scale (Pei et al., 2023), which is a close-ended, forced-choice scale, whose options listed from 1 to 5 have strong inner order. That inspired us to introduce ordinal regression into our model.

Ordinal regression dedicates to learn a rule to predict labels from an ordinal scale by ensuring that predictions farther from the true label receive a greater penalty than those closer to the right label (Pedregosa et al., 2017; Rennie and Srebro, 2005).

If we have k ordinal labels $\mathcal{Y} = \{y_1, \ldots, y_k\}$, these labels are separated by $k+1$ boundary values $\mathcal{A} = \{\alpha_0 \ldots \alpha_k\}$, to be more generalized, extend the bound to infinity, i.e. $\alpha_0 = -\infty$ and $\alpha_k = +\infty$. The probability of label y_i could be defined as:

$$p(y_i) = \Pi(\alpha_i) - \Pi(\alpha_{i-1}), 1 \leq i \leq k \quad (3)$$

where Π is the Cumulative Distribution Function (CDF), and we used `sigmoid` in our code. To describe the error between ground-truth and predicted labels, we used a "inter-boundary" method, i.e.

$$p(y_i) = \Pi(\alpha_i - z) - \Pi(\alpha_{i-1} - z), 1 \leq i \leq k \quad (4)$$

where $z = z(x)$ is the model's predict value based on input dataset X, and the ordinal regression loss function is defined as:

$$\mathrm{loss}(z(x); \mathcal{Y}, \mathcal{A}) = -\sum_{i=1}^{k} \log p(y_i). \quad (5)$$

$z(x)$ starts up with a random number, and $\mathcal{A}$ could be treated as a hyperparameter.

3.4 Data Augmentation

In this task, data augmentation is crucial considering of the zero-shot cross-lingual prediction target, and also the dataset is relatively small for large pretrained language models, we employed two levels of data augmentation in our system.

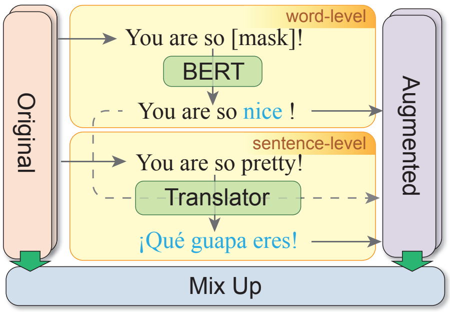

Figure 2: Two levels of augmentation. Word-altered sentence could also be translated.

3.4.1 Word-level: Substitution

By tweaking words, faint noises are introduced to the model, which may challenge the training process and improve model robustness. We used the `bert-base-multilingual-uncased` (Devlin et al., 2018) as a masked language model to randomly mask out and substitute word-pieces in tweets (Jiao et al., 2019). But thinking of the essence of tweets-concise, we refrained from deleting or inserting random words since which may drastically alter the literal meaning.

3.4.2 Sentence-level: Translation

Although XLM-T is pre-trained using multi-lingual corpus and XLM has a translation language modeling (TLM) pre-train task, considering zero-shot learning is an extreme case in transfer learning which may only work well under the ideal situation, in other words, such theory is only feasible unless additional information has been exploited during training (Socher et al., 2013), we should provide unseen language samples.

In such semantic comprehension tasks, there is no doubt that the richer semantic information is, the better the model performs (Xu et al., 2022). By calling Azure Translator API[2], we randomly chose tweets from the original dataset and translated them into (a) 6 seen languages, to enrich the original training set; (b) 4 unseen languages, to provide additional information for training.

Level 1 and level 2 could be applied separately or simultaneously, after which, an enhanced dataset was obtained. During training, a portion of augmented data was chosen and mixed up with the original dataset.

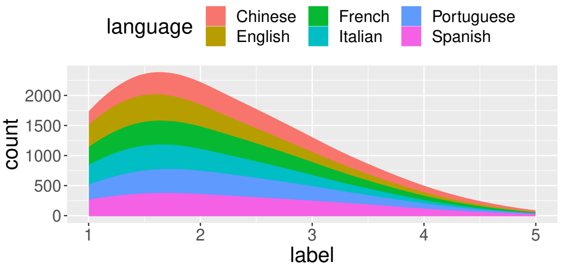

Figure 3: Label distribution of training set. Intimacy labels around 1.5 are the most among all the languages.

3.5 Training Strategy

In the model training stage, we tried to randomly split 10% of the training set as the validation set to fine-tune the model, and we chose one language from the six seen languages and added them all to the validation set to verify the model's predictive performance for text intimacy in unseen languages. In order to unleash the potential of the model's unseen language prediction and eliminate the diversity of intimacy conveyed in different languages, we proposed the strategy of multilingual cross-validation training. During the training process, the composition of the validation set and the training set was changed every 13 epochs. The six languages were each used as the unseen language in turn, with a total of 78 epochs for training across all languages. All parameters in XLM-T were not frozen during training.

4 Experiments

4.1 Dataset and Evaluation

The intimacy dataset, MINT (Pei et al., 2023), used in this competition covers 13,372 in 10 languages tweets sampled from 2018 to 2022. The label distribution of datasets are shown in Figure 3.

As mentioned in the official task description, we use Pearson's r to evaluate the level of linear correlation between predictions and ground-truth. The Pearson correlation coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations:

$$r = \frac{\sum (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum (x_i - \bar{x}_i)^2 \sum (y_i - \bar{y}_i)^2}} \tag{6}$$

4.2 Experimental Settings

We implemented the intimacy analysis model based on XLM-T (Barbieri et al., 2021) with Pytorch (Paszke et al., 2019) and the Huggingface Transformers library (Wolf et al., 2020). The AdamW

[2] https://azure.microsoft.com/products/cognitive-services/translator/

Model	English	French	Dutch	Arabic	Seen	Unseen	Overall
Our Model	0.7126	0.6990	0.5765	0.6248	0.7184	0.4158	**0.5866**
OR loss^9	0.6925	0.6712	0.5929	0.5810	0.6962	0.4297	0.5778
MSE loss	0.5570	0.5795	0.4632	0.4137	0.5821	0.4073	0.5014
- Adversarial Training	0.6983	0.6721	0.6191	0.5601	0.7037	0.4119	0.5744
- Training Strategy	0.7049	0.7016	0.5882	0.6461	0.7018	0.4219	0.5753
Data Augmentation							
4 languages (substitution)	0.7080	0.6713	0.6168	0.6156	0.7118	0.3844	0.5619
4 languages	0.7042	0.6950	0.5990	0.6425	0.7139	0.4308	0.5865
6 languages	0.7063	0.6965	0.5865	0.6200	0.7159	0.3781	0.5737
Submission Results	0.6992	0.6674	0.6353	0.5890	0.7029	0.4359	0.5808

Table 1: The table illustrates Pearson correlation scores of applying different optimization methods to XLM-T. We applied all effective optimization methods to **Our Model** including translation augmentation of 6 seen language texts (**6 languages**), randomly masking out and substituting word-pieces in tweets (**substitution**), adversarial training, specific training strategy and ordinal regression loss (k = 17). **4 languages** means translation augmentation of 4 unseen language samples. In addition, we tested a variety of losses during model training, among which **OR loss**^{9} represented the ordinal regression loss with the k value of 9.

optimizer was used for model training. Before starting to train the model, we processed the data, specifically, we filtered user tags and some invalid characters. We kept the URL, which had a positive effect on intimacy analysis. During the training phase, we evaluated the performance of the model every 100 steps and retained the parameters of the model that performed best on the validation set. The hyperparameters settings adopted are shown in Table 2. All models were trained on NVIDIA Geforce GTX 3090 GPU.

Hyperparameters	Value
k	17
Epochs	13
Dropout	0.3
Batch size	32
Hidden dim	1024
Learning rate	1e-5
Sequence length	50

Table 2: The hyperparameters of the experiment.

4.3 Results and Discussions

4.3.1 Adversarial Training Analysis

We set up the experiment to verify the effectiveness of adversarial training, as shown in Table 1. From the experimental results, we can observe that the performance of the model's intimacy prediction is improved after adding FGM. It shows that adversarial training can improve the robustness of the model.

4.3.2 Ordinal Regression Analysis

The results in Table 1 demonstrate the effect of the ordinal regression loss. Compared with MSE loss, the system performance is significantly improved by using ordinal regression loss, which further confirms that using ordinal regression as the loss function to help model capturing the interval label order rather than fit the continuous value will benefit more for predicting multiple-level discrete ordinal labels. At the same time, it can also be observed from the results that finer-grained interval division can improve the performance of the model.

4.3.3 Data Augmentation Analysis

In order to analyze whether data augmentation will improve the generalization of the model, we designed three experiments **6 languages**, **4 languages** and **4 languages (Substitution)** to compare our model, and the details are shown in Table 1. Translation augmentation and word-level noise on seen language data can bring observable benefits to the model. However, the model's performance will suffer if noise is added while expanding unseen language data. This may be because using training sets to extend unseen languages introduces the distribution error, and adding perturbations to the data will cause the error to progressively grow.

4.3.4 Training Strategy Analysis

Since the text features of intimacy between different languages are not the same, the model will have deviations in learning performance due to the different text language combinations in the training set during model training. As the results in the Table 1, compared with the single combination of training set languages, it is better to alternately use combination methods of different languages in one training process, which proves that the training strategy we designed is more reasonable.

5 Conclusion and Future Work

In this paper, we present an intimacy analysis system by deploying various optimization methods, including adversarial training, data augmentation, ordinal regression loss and special training methods, to SemEval-2023 Task 9.

Compared with the method of sharing the same task in multiple languages, building a multi-task model for different languages is a more worthy of analysis. Multilingual research based on pretraining is more dependent on the diversity of languages contained in the pre-training corpus. For low-resource languages, the performance of the model will get stuck in a bottleneck. In the future, we will further explore the relationships between different linguistic features and build a stronger multilingual multitasking system for the intimacy analysis.

Acknowledgement

We thank our anonymous reviewers for their helpful comments. This work was supported by grant from the National Social Science Fund of China (Grant No. 21BXW047).

References

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *arXiv: Machine Learning*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. 2017. On the consistency of ordinal regression methods. *The Journal of Machine Learning Research*, 18(1):1769–1803.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Jason DM Rennie and Nathan Srebro. 2005. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*, volume 1. AAAI Press, Menlo Park, CA.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv: Computation and Language*.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Neural Information Processing Systems*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Zihang Xu, Ziqing Yang, Yiming Cui, and Zhigang Chen. 2022. HFL at SemEval-2022 task 8: A linguistics-inspired regression model with data augmentation for multilingual news similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1114–1120, Seattle, United States. Association for Computational Linguistics.