lazybob at SemEval-2023 Task 9: Quantifying Intimacy of Multilingual Tweets with Multi-Task Learning

Mengfei Yuan
The Ohio State University
Sunwoda Electronic Co., Ltd.
ymf924@gmail.com

Cheng Chen
Ping An Insurance
oscarhscc@gmail.com

Abstract

This study presents a systematic method for analyzing the level of intimacy in tweets across ten different languages, using multi-task learning for SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis. The system begins with the utilization of the official training data, and then we experiment with different fine-tuning tricks and effective strategies, such as data augmentation, multi-task learning, etc. Through additional experiments, the approach is shown to be effective for the task. To enhance the model's robustness, different transformer-based language models and some widely-used plug-and-play priors are incorporated into our system. Our final submission achieved a Pearson R of 0.6160 for the intimacy score on the official test set, placing us at the top of the leader board among 45 teams.

1 Introduction

Intimacy is a fundamental social aspect of language. Understanding the dynamics of intimacy in conversation is important for a variety of practical applications. Research on intimacy in conversations between people, particularly in the context of chance encounters, such as the phenomenon of the "stranger on the train," can provide valuable insights into human communication and relationships (Griffin, 2006; West et al., 2010; Farber, 2003). Studying the content on different social platform can help researchers to better understand how people establish and develop relationships through conversation. Social workers can also benefit from an understanding of how intimacy is established and maintained in conversation to help clients build stronger relationships with others (Labov, 1972; Brown and Levinson, 1978).

The multilingual tweet intimacy task was hosted by the University of Michigan and Snap Inc. This task focuses on predicting the intimacy of tweets in 10 languages (Pei et al., 2022), where 6 languages with training data (English, Spanish, Portuguese, Italian, French, Chinese), and 4 zero-shot languages (Korean, Arabic, Hindi, Dutch). Our submission achieved the first place overall score in these 10 languages.

In this paper, we demonstrate the following contributions: (1) We designed a novel multi-language and multi-task training framework that can simultaneously train data from all languages. (2) Based on 10 million crawled Twitter, book, movie, and Reddit data in various languages, we conducted domain pre-training of multiple pre-training models to enhance the model's prediction ability in that domain. (3) During the training process, we also adopted contrastive learning strategies such as group-layerwise learning rate decay (GLLRD), adversarial weight perturbation (AWP) to enhance the semantic understanding and robustness of language models. (4) For zero-shot language prediction, we tried bidirectional translation data augmentation to enhance the model's prediction ability for small languages in this task.

2 Background

Sociologists and linguists have conducted many studies on how people use language to measure the relationships between individuals. By analyzing conversations, researchers can estimate the relationship between people, such as friendly, neutral, or adversarial relationship (Clark and Schunk, 1980; Weber, 2008; Locher and Graham, 2010). Recently, some scholars have proposed computational models to evaluate the intimacy of a conversation, which greatly reduces the workload of linguistic research (Pei and Jurgens, 2020).

By training on a large amount of annotated data, the proposed model can learn to better understand the semantics of language and accurately predict the intimacy level in multiple languages. This task is significant for comprehending how language is utilized to convey emotions and relationships, and

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 927–933

has practical applications in fields such as psychology, social work, and marketing.

3 System Overview

To better understand the content and semantics of the training and testing sets, which are drawn from Twitter posts and comments, we use large language models trained specifically on tweets, for language encoding. We also use additional datasets from Twitter and Reddit, as well as some from books and movies to pretrain language models. At the same time, we used multiple fine-tuning tricks to improve the model's performance. To make the distribution of the predicted intimacy scores are close to the distribution of the true annotation scores, we designed a multi-task training mode to adjust the distribution of the predicted labels. We also used model fusion methods to generate more robust results. The following section introduce the detail of language model, pretraining method, fine tuning tricks, multi-task setting and ensemble method, etc. were used in this paper.

3.1 Language Models

Considering the source and language of the data, we tested about many pretrained language models in different strategy to understand the semantics of the tex. All of the pretrained models mentioned can be downloaded from Hugging Face [1]. This section introduces some effective pretrained models that were used, along with their basic information and advantages.

Due to the fact that the task dataset came from Twitter, was multilingual, and included zero-shot languages, we mainly used large multilingual models that has been pretrained on tweets. Additionally, we used different models for Chinese and Spanish according to related work (Pei et al., 2022). To enhance the performance of predicting zero-shot languages, we used a data augmentation strategy by using translation. Both the Hindi and Korean which shown a very poor performance in task paper are consider additional training using translation.

XLM-RoBERTa [2]: XLM-R is a multilingual language model that builds on the RoBERTa architecture and outperforming BERT on many multilingual NLP tasks, including language modeling, text classification, and machine translation (Devlin et al., 2018; Conneau et al., 2019). It performs well on text encoding and generation on many different languages, including low-resource languages that have little training data available. The XLM-RoBERTa model is pre-trained on a large corpus of text from more than 100 languages. It is trained using a self-supervised learning approach, which means it learns to represent language by predicting masked words and the order of sentences in a given text.

DeBERTa [3]: The main advantage of DeBERTa over BERT is that it uses disentangled attention, which allows the model to focus on different aspects of the input sequence in a more fine-grained way (He et al., 2020). Based on many previous tasks, DeBERTa can capture complex relationships and dependencies between words and characters, especially on English text, resulting in improved performance on a wide range of NLP tasks.

Twitter-XLM-Roberta [4]: XLM-T trained start from XLM-Roberta base and continue pre-training on a large corpus of Twitter in multiple languages (Barbieri et al., 2021). For text augmented by translation strategy, we applied the latest version of this model which trained only by English datasets [5]. The Twitter-XLM-Roberta model has performed very well on Twitter-related corpora. This is likely due to the fact that it was pre-trained using a large amount of data in domain.

TwHin-BERT [6]: TwHin-BERT is based on the BERT architecture and is designed to encode the meaning of tweets in multiple languages, taking into account the specific linguistic and social features of tweets (Zhang et al., 2022). The main advantage of TwHin-BERT compared to other BERT-related models is its ability to capture the social and linguistic features that are unique to tweets, such as hashtags, emoticons, and slang words. These features are often used to convey sentiment, humor, or irony in tweets, and can significantly impact the meaning of a tweet.

3.2 External data and Pretrained Models

In development stage, we realized the domian-based data can improve the model preformance. We select TwHin-BERT, which shown a best performance in validation stage, as a base model to

[1] https://huggingface.co/models?sort=downloads
[2] https://huggingface.co/xlm-roberta-large
[3] https://huggingface.co/microsoft/deberta-v3-large
[4] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment
[5] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest
[6] https://huggingface.co/Twitter/twhin-bert-large

pretrain a specific model for this task. The external data are processed and provided by the organizer (Pei and Jurgens, 2020). However, due to the time and cost, we only sample some part data (10M out of 80.5M) from the external data to pretrain the TwHin-BERT model. The external data we used to pretrain include 1M English tweet [7] and 9M from Reddit, movies and books (Pei and Jurgens, 2020; Louviere et al., 2015; Kiritchenko and Mohammad, 2017). We also test the performance of this self-trained model. In the following section, the self-pretrained model is referred as "Intimacy-TwHin-Bert".

3.3 Fine-tuning Strategies

In this task, we applied some popular deep learning fine-tuning techniques to enhance the model robustness, such as weight sampling, constrastive learning, dropout, attention mechanism, etc. The performance and effectiveness of different trick are also test in the section of Experimental setup.

3.3.1 Weighted Random Sampler

In deep learning, a weighted random sample is used for giving more importance to certain samples (He and Garcia, 2009). The advantage of using a weighted random sample over a simple random sample is that it allows the model to focus more on interested samples. It also can help to mitigate the negtive effects of noisy or irrelevant data. By assigning lower weights to such examples, the model is less likely to be biased.

3.3.2 Attention

In this task, we applied attention on text embedding output from a pretrained model (Vaswani et al., 2017). A set of attention weights is calculated from the output embedding using attention mechanism. It learns to assign weights to different parts of the input data based on their importance or relevance to the task at hand. Attention focus on different parts of the input data based on their relevance or importance, enabling the model to achieve better accuracy.

3.3.3 Multi-sample Dropout

Dropout is a regularization technique that is commonly used to prevent overfitting in neural networks (Srivastava et al., 2014). It randomly drops some nodes in a neural network during training to prevent the network from relying too heavily on any single feature. However, dropout technique shows less effective for regression tasks than for classification tasks. In this task we set a dropout with a rate of 0 for regression tasks (Bahdanau et al., 2014).

For classification problem in multi-task learning, we incorporate the multi-sample dropout layer technique into the output of pretrained language models to improve the model generalization. Multi-sample dropout is a regularization technique which can accelerate training convergence and improve the model generalization compared to the network structure with a traditional dropout layer (Inoue, 2019). Four dropout layers with a dropout rate equals 0.4 were applied to the pooling layer output from the pretrained model.

3.3.4 Adversarial weight perturbation

Adversarial weight perturbation (AWD) is a contrastive learning technique that aims to improve the quality of learned representations by adding perturbations to the weights of the neural network. The basic idea is to introduce small and designed perturbations to the weights of the neural network in order to make the learned representations more robust to adversarial attacks (Wu et al., 2020; Liu et al., 2020; Tian et al., 2019). This technique usually can applied to improved performance on downstream tasks.

3.3.5 Group-layer wise learning rate decay

Group-layer wise learning rate decay (GLLRD) is a technique used in contrastive learning to adjust the learning rate of different layers in a neural network during training (Tenney et al., 2019; Tan and Le, 2019). It divide the layers of the neural network into different groups based on their depth or complexity, and then apply a different learning rate schedule to each group. The advantage of using GLLRD is ro prevent overfitting and improve the training efficiency and effectiveness of the neural network.

3.4 Ensemble

In the final version of our submission, we utilized a combination of fusion strategies. The version that we ultimately submitted involved merging 25 sets of results from the Multi-learning Task, all of which were trained using Twhin-BERT. We based this decision on our previous experimentation and validation scores obtained during the evaluation stage. Furthermore, we substituted the Chinese

[7] https://developer.twitter.com/

results by training the model using XLM-R, while the Spanish results were replaced with a model that was trained using XLM-T.

4 Experimental setup

In this competition, we applied 3 strategy with different languages models and datasets to predict the intimacy score of tweets in 10 languages. Based on the organizers' description, this task is more like a classification task since the initial task of the annotators was to label the data into one of the five levels of intimacy on a scale of 1-5. However, the true training data provided is the average value of multiple types of annotation data, which is not a standard Multi-level classification.

Considering the final evaluation metric for this competition is Pearson correlation coefficient (Pearson R) (Yeager, 2019), we consider this task as a regression problem for prediction. The goal of the task is to predict a continuous value of intimacy score rather than a discrete class label, and the Pearson r evaluates the correlation between the predicted values and the true values.

The three strategies are described as below. Table 1 shows the overall performance using different strategies. Table 2 shows the performance in each single language with different languages models. TwHin-BERT-large trained using multi-task learning provide the best performance all over all the method.

Strategy 1 (Single task on official data):

The first strategy we used is to treat this task as a pure regression problem, using the raw data provided by the organizer for training. Since the data was already cleaned, we only performed some basic text processing on features unique to tweets. We then added an attention layer and fine-tuned the model using various language models.

Strategy 2 (Single task on augmented data):

Due to limited size of our training dataset, we perform data augmentation through Google Translation Cloud API [8] to translate all data into English. We then fine-tuned a well-performing English pretrained language model using the translated data. This approach allowed the model to better handle the linguistic features, particularly in zero-shot tasks.

Strategy 3 (Multi task with Pseudo labels):

The label distribution from 1 to 5 is unbalanced and has very few labels greater than 3 (See Table 3 in Appendix). When we treat this task as a pure regression problem, we used Mean Squared Error (MSE) as the loss function and Pearson R as the evaluation metric. This can lead to a general problem caused by the convergence of the MSE loss function, where the model will pursue a similar average of the true and predicted values for each batch to achieve a higher Pearson R score during validation.

To address this issue, we firstly tried removing the weighted sampler in the batch sampling process. However, this strategy did not work and actually had a negative impact. Therefore, we ultimately adopted a multi-task learning approach to train the model. We converted the official labels into pseudo-labels based on their specific numerical values. Specifically, we divided the labels into four categories based on four intervals: [1-2), [2, 3), [3, 4), and [4,5]. In Multi-task learning, we trained the regression with MSE loss and classification with cross entropy loss at a same time. In validation stage, we still only applied the Pearson R as the metric. Figure 1 shows the label prediction using multi task learning is closer to the annotated distribution rather than single task distribution.

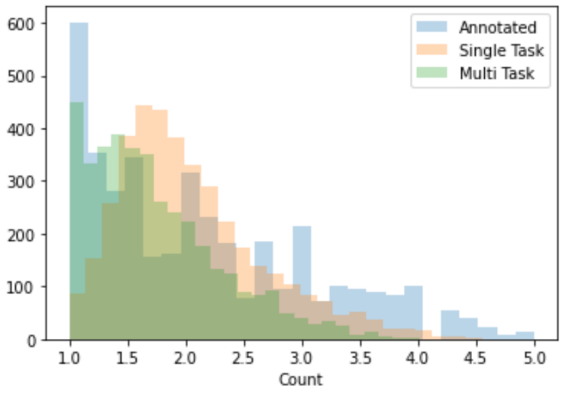

Figure 1: Prediction distribution for Single and Multi-task learning

5 Result and Discussion

In conclusion, Twhin-BERT is the best language model that can understand the semantic of competition tweets, both in zero-shot and trained languages. Data augmentation by translation helps improve the accuracy of prediction of zero-shot languages, such as Dutch and Arabic. Multi-task learning help to amend the distribution of label prediction and improve the prediction accuracy. Contrastive learning strategies can slight improve the performance of

[8] https://cloud.google.com/translate

Language	TwHin(S)	TwHin(M)	TwHin(T)	Twitter-R(T)	XLM-T	XLM-R
English	0.754085	**0.755503**	0.745862	0.729933	0.716733	0.701392
Spanish	0.778559	**0.781324**	0.744150	0.714292	0.742690	0.727252
Portuguese	0.676013	**0.688518**	0.637796	0.635662	0.687433	0.653045
Italian	0.729114	**0.739180**	0.696133	0.681582	0.706044	0.699579
French	0.717390	**0.719804**	0.697370	0.687506	0.703939	0.704679
Chinese	0.741089	**0.748144**	0.687744	0.667377	0.713368	0.745640
Korean	0.387832	**0.427624**	0.360079	0.338618	0.401420	0.414930
Hindi	**0.268562**	0.264481	0.238121	0.250108	0.198754	0.239456
Dutch	0.607502	0.629129	**0.648211**	0.639731	0.627170	0.650921
Arabic	0.622513	0.655984	0.659934	**0.671837**	0.625884	0.615393

Table 1: Performance of each language using different model. The bottom four rows are tested under the zero-shot setting. The special note S (Single task), R (Multi task), and T (Translation task) stands for different strategies.

Strategy	Language	Pre-trained model	Dropout	Overall	Known	Unknown
Single Task	XLM	XLM-R	-	0.589161	0.716019	0.437579
		XLM-T		0.582964	0.722472	0.454854
		TwHin-BERT		0.575757	0.743396	0.410383
	En	Twitter-R		0.571658	0.689799	0.437219
		DeBERTa		0.569497	0.696589	0.420566
		TwHin-BERT		0.568751	0.705585	0.405864
		Intimacy-TwHin-BERT		0.578476	0.708268	0.425555
Multi Task	XLM	TwHin	0.1	**0.618196**	0.749388	**0.499055**
			0.2	0.596449	**0.753423**	0.440531
			0.4	0.616463	0.750014	0.498403

Table 2: Pearson R score using different strategies and language model

label prediction but it requires a lot more effort on training.

Additionally, we also tried other approaches which not demonstrated in details. For instance, we tried training each language separately, but due to the small size of the training data, the validation score was relatively low. We also tried another multi-task learning approach, where each language was treated as a separate task and trained separately as a regression problem. However, the performance of this approach was even worse than that of the single task, probably because of the lack of training data. Finally, to ensure the prediction performance for zero-shot languages, we also tried translating all of the training data into Korean and Hindi to improve the performance of these two languages. However, the results were not as good as using the multi-language strategy directly, possibly due to the noise introduced by translation.

6 Conclusion and Future Work

Overall, research on intimacy in conversations between people can provide valuable insights into social relationships. This work developed a pretrained and multilingual language model to determine the intimacy level of tweets. The model perform well on both trained language and zero-shot languages. In the future, research can consider adding feature engineering and tuning more deep learning tricks to improve the performance of the model prediction. Additionally, this study demonstrated that pretraining a model with domain knowledge can improve its robustness, so it is important to consider pretraining the model with relevant content before using it to predict intimacy scores for specific types of conversations.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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

Penelope Brown and Stephen C Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, pages 56–311. Cambridge University Press.

Herbert H Clark and Dale H Schunk. 1980. Polite responses to polite requests. *Cognition*, 8(2):111–143.

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.

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*.

Barry A Farber. 2003. Patient self-disclosure: A review of the research. *Journal of clinical psychology*, 59(5):589–600.

EM Griffin. 2006. *A first look at communication theory*. McGraw-hill.

Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Svetlana Kiritchenko and Saif M Mohammad. 2017. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. *arXiv preprint arXiv:1712.01741*.

William Labov. 1972. *Language in the inner city: Studies in the Black English vernacular*. 3. University of Pennsylvania Press.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Miriam A Locher and Sage L Graham. 2010. *Interpersonal pragmatics*, volume 6. Walter de Gruyter.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozon Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2022. Semeval 2023 task 9: Multilingual tweet intimacy analysis. *arXiv preprint arXiv:2210.01108*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tilo Weber. 2008. *Handbook of interpersonal communication*, volume 2. Walter de Gruyter.

Richard L West, Lynn H Turner, and Gang Zhao. 2010. *Introducing communication theory: Analysis and application*, volume 2. McGraw-Hill New York, NY.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969.

Kristin Yeager. 2019. Libguides: Spss tutorials: pearson correlation. *kent. edu*.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.

A Appendix

Language	all	[1,2)	[2,3)	[3,4)	[4,5]
Chinese	1596	647	552	294	103
English	1587	945	404	182	56
French	1588	774	524	223	67
Italian	1532	846	458	190	38
Portuguese	1596	715	536	279	66
Spanish	1592	703	496	307	86

Table 3: Distribution of training dataset label. The label are categorized in to four class based in values.