Project Report: Multilingual Tweet Intimacy Analysis

Youheng Fu
youhfu@umich.edu

Abstract

In this digital age, social media serves as a important platform for global communication, improving intimate exchanges across diverse cultures. This study explores the prediction of intimacy levels in 10 languages using tweet data, with intimacy scores ranging from 1 to 5. Other researchers tried different LLMs by simply assembling the results. We employ a novel approach, fine-tuning large language models (LLMs) through regression tasks and employing stacking ensemble techniques with linear regression as the meta-model. The final overall Pearson's r coefficient is 0.598. Our results outperform random and mean label baselines, demonstrating the efficacy of our approach for future research in sentiment analysis and natural language processing.

1 Introduction

In the era of information and internet, social media has become a very common platform where people can share their colorful life to other people they do not come across. Meanwhile, this new communication method connects more and more people with different countries and culture backgrounds. Intimacy, as an elementary role in people's interaction, then emerges in the textual communication more frequently.

By analyzing intimacy revealed in people's daily textual communication such as tweets, we can not only enhance the effectiveness of extracting different expression features in various languages, but can also broaden the boundary of sentiment analysis. Researchers in the future can add the degree of intimacy as a new feature for sentiment analysis or other natural language processing task.

This task aims to use different models to predict the intimacy degree of 10 different languages including English, Chinese, Italian, etc. The degrees of intimacy is from 1 to 5, where 5 is the maximum degree. Jiaxin Pei (2023) provides tweets texts as our input, and we will predict the intimacy scores after training the model. Additionally, we can compare the performance of prediction for different languages. The task is actually the task 9 in SemEval 2023 competition. Other researchers try to solve this task by analyzing performance of different multilingual large language models (LLMs) and augmenting the training data. In this project, I will fine-tune some LLMs by regression task, and use stacking method to ensemble the result of them. The meta model for stacking regressor is linear regression model. The final overall Pearson's r coefficient is , which is much better than our random baseline and mean label baseline. The result of this project shows a different way to ensemble the result for future researches.

2 Data

Jiaxin Pei (2023) provides our training and testing dataset. The dataset contains tweets sampled from 2018 to 2022 and the data has been pre-processed carefully by organizers (Jiaxin Pei, 2023). One of the important process part is annotation. The intimacy annotations are actually labelled by professional humans and designed to decrease bias.

The training dataset contains 9491 data in total. Each data points contains 3 different features: text, label and language. The top 5 data in the dataset is shown as below:

	text	label	language
0	wearing a fake engagement ring so guys won't a...	1.8	English
1	Bees vs. Wasps. http	1.0	English
2	Here is a nice equation: 0+0-0-0+0=0	1.0	English
3	@user @user Enjoy each new day!😊 ᴄᴀ 🐞 🐭	1.6	English
4	I can be having a perfectly good day then I th...	1.6	English

Figure 1: Top 5 data samples in training dataset

For training dataset, there are only 6 languages, including English, Chinese, Portuguese, Spanish,

French, and Italian. The performance of model will be tested by those 6 languages and another external dataset which containing 4 other languages (Hindi, Arabic, Dutch and Korean). The Figure 2 shows the distribution of labels for 6 different languages in traning set. The result indicates that although the labels in different intervals are not distributed uniformly, the distributions are relatively balanced among languages. The majority of labels lies in the interval $[1, 2]$.

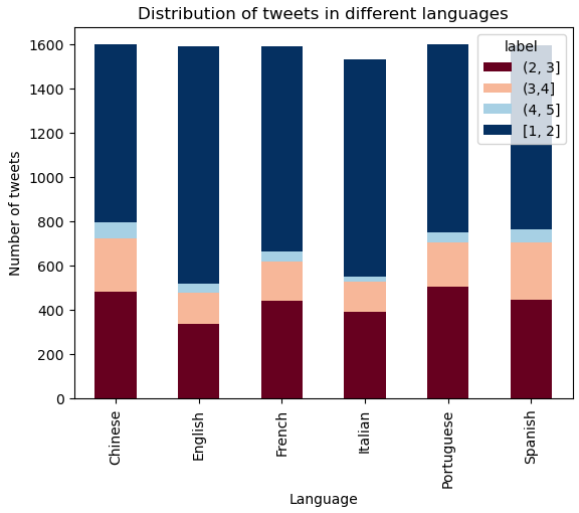

Figure 2: The distribution of intimacy of 6 different languages in training set

I can also find that the size of data points in test set for different language are similar, as shown in Figure 3:

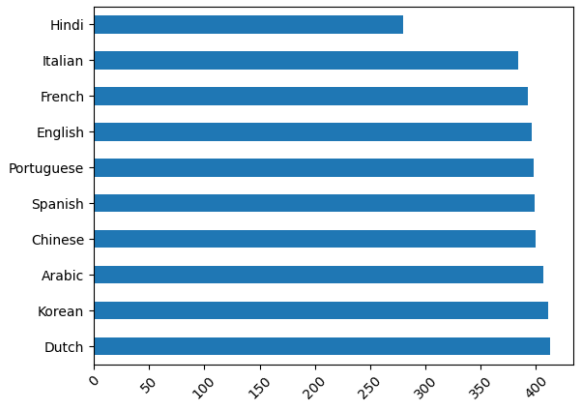

Figure 3: The number of tweets for 10 languages in test set

I also used some simple methods to finish the data pre-processing part. First, I changed all the texts in lowercase for both training and testing set. Additionally, I applied regular expressions to filter irrelevant information in the tweet texts, such as URLs.

3 Related Work

Many other researchers provide their own models to solve this problem. B et al. (2023) proposed ROBERTa (Robustly Optimizer Bidirectional Encoder Representations from Transformations Approach) model to analyze the intimacy. This model utilizes advanced transformer architecture and optimization techniques to achieve robust performance on the task. However, the model cannot perform well when it is used to predict test set. García-Díaz et al. (2023) addressed this task based on data augmentation and 3 large language models (multilingual BERT, XLM and mDeBERTA). By combining these models and augmenting the training data, they finally improved generalization performance of model. The main strategy of Pichardo Estevez et al. (2023) was the undersampling and oversampling methods for training dataset. The authors contribute to the modification of original data, but they do not improve the model structures. Cai et al. (2023) combined XLM-T and bidirectional GRU layers to predict the results, but they only focus on the one potential model (XLM-T). He and Zhang (2023) proposed a new loss function for their model: EPM (Exponential Penalty Mean Squared Loss). All those groups get a relatively satisfying result. In my work, I will compare many different multilingual model, such as XLM-R, XLM-T and ensemble their results. Some researchers also use ensemble methods in their work, but most of them just simply combine and take averages. In this project, I apply stacking method and choose linear regression as meta model, using the predicted labels from other models as features. This ensemble model can produce an more robust result and reduce the risk of overfitting.

4 Methodology

4.1 Data Prepocessing

I apply regular expression to clean the original data and remove irrelevant information, such as URLs. I also make the texts lowercase to capture the same meaning of words with different cases. Then I use tokenizer to transfer tweets into tokens. Here we use different tokenizers for different models, but since all the models I use in the next parts are BERT based, I can get both the token ids and attention mask matrix as outputs of tokenizers.

4.2 Large Language Models

Since the texts in the task are all tweets and consist of many different languages. I use multilingual LLMs or multilingual based version of models for this and below models.

- **Multilingual DistillBERT**: The model is trained using knowledge distillation, which significantly decreases the number of parameters but still keep a satisfying result compared to original BERT model (Sanh et al., 2019). Due to the smaller size, the model may perform as good as BERT model for training set, but it has a stronger ability for zero-shot task.
- **mDeBERTa**: This model is the multilingual version of DeBERTa model, which uses disentangled attention and enhanced mask decoder to improve the performance (He et al., 2021).
- **XLM-R**: This model combines the advantages of two LLMs: XLM and RoBERTa, and leverages cross-lingual transfer learning, meaning it can transfer knowledge learned from one language to improve performance on tasks in other languages (Conneau et al., 2019). I choose this model for its better performance on zero-shot task, such as prediction for Hindi, which is not shown in our training set.
- **XLM-T**: It is a XLM-R model training on 198M multilingual tweets (Barbieri et al., 2022). Since our task is based on the tweets, this model can resolve the in-context knowledge and perform well.

4.3 Training

To training these large language models, we first use the tokenizer corresponding to each model to get the token ids and the attention mask matrix for each sequence as the input of model. Then I need to construct a regressor to predict the result. Since those models cannot give us a pooler result as the original BERT model does, I construct the regressor by sequence classification model with number of labels being the hidden size. The final output is generated by a dropout layer with rate 0.1 and a linear layer mapping the result to one predicted number. Due to the time limit, I did not try many combinations of hyperparameters. The chosen learning rate is 1e-5, the loss function is MSE loss, the optimizer is AdamW, the batch size is 4, and I train those models for 10 epochs.

4.4 Stacking

I apply stacking method to combine the results of different LLMs. I choose linear regression as my meta model, and utilize the predicted labels of other models as features. Since the model is evaluated by Pearson's r coefficient, which reflects the linear relationship of original and predicted labels, the simple linear regression model can enhance the linear relationship and enhance the performance to the final result. The figure 5 shows the insight behind the stacking ensemble algorithm

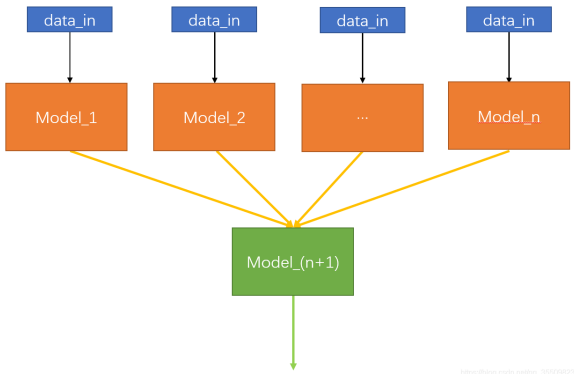

Figure 4: Stacking Ensemble Algorithm

5 Evaluation and Results

The evaluation method for this task is simple: we try to use Pearson's r value as the metric according to Jiaxin Pei (2023). The formula of calculating Pearson's r value is:

$$r = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}}$$

where X, Y are the true intimacy scores and the predicted values, respectively. As the value of r getting closer to 1, the model is better.

I have some baseline methods to make simple analysis for the datasets.

5.1 Baseline 1

I randomly pick the labels for the data in training set. The result of Pearson's r value is shown as below:

5.2 Baseline 2

In this baseline method, I use the mean of all labels of a dataset as its predictive label. The result is shown as table 2:

The positive Pearson's r value means the pseudo "predictive" result is positive correlated with the

Language	Pearson's r
English	0.012701
Spanish	0.081860
Portuguese	-0.000828
Italian	0.025090
French	0.010133
Chinese	0.016256

Table 1: Pearson's r value for 6 languages with random label

Language	Pearson's r
English	1.365650e-16
Spanish	2.392165e-16
Portuguese	-1.300971e-16
Italian	-1.388904e-16
French	1.199435e-17
Chinese	2.565374e-17

Table 2: Pearson's r value for 6 languages with mean label

true label, which indicates the good performance of model. If the Pearson's r value is close to 0, it reveals that the result is independent with the ground truth. The results of baselines cannot reveal the relationship between predictions and true labels.

5.3 Multilingual DistillBERT

The loss of model which is accumulated in 10 steps is shown as below:

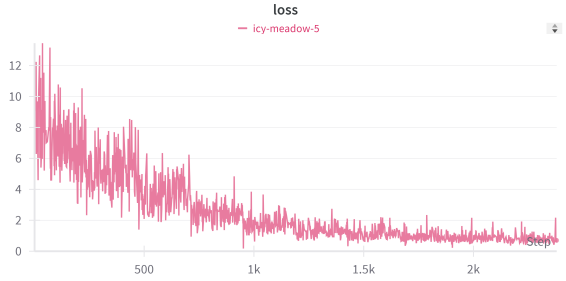

Figure 5: Loss for every 10 steps in training process

Table 3 shows the result of multilingual DistillBERT model:

5.4 mDeBERTa

Table 4 shows the result of mDeBERTa model:

5.5 XLM-R

Table 5 shows the result of XLM-R model:

Language	Pearson's r
English	0.57150
Spanish	0.60630
Portuguese	0.54381
Italian	0.58105
French	0.57072
Chinese	0.64005
Hindi	0.13724
Korean	0.28107
Arabic	0.33117
Dutch	0.43789

Table 3: Pearson's r value for 10 languages for mDistilBERT

Language	Pearson's r
English	0.67333
Spanish	0.68207
Portuguese	0.63406
Italian	0.68051
French	0.66721
Chinese	0.71694
Hindi	0.25727
Korean	0.35612
Arabic	0.55122
Dutch	0.54952

Table 4: Pearson's r value for 10 languages for mDeBERTa

5.6 XLM-T

Table 6 shows the result of XLM-T model:

5.7 Stacking

The overall result is 0.598. Table 7 shows the result after stacking of all results of above 4 models:

6 Discussion

As stated in section 5, the baseline method 1 which use randomly assigned labels to calculate Pearson's r coefficient has a low performance, which is not surprising. The randomly picked results do not have actual meaningful relationship with the truth data and do not capture any information. For baseline method 2, the constant mean labels nearly give a zero output. The reason is that the mean label has no variance and also cannot reflect any patterns or information with the truth. Hence, we do need models to discover the internal relationship in the text and multilingual contexts.

Comparing the four large language models I used, the same thing is that they all perform bet-

Language	Pearson's r
English	0.63972
Spanish	0.66047
Portuguese	0.61908
Italian	0.65615
French	0.63320
Chinese	0.69370
Hindi	0.26557
Korean	0.33267
Arabic	0.52264
Dutch	0.56298

Table 5: Pearson's r value for 10 languages for XLM-R

Language	Pearson's r
English	0.69827
Spanish	0.73223
Portuguese	0.68071
Italian	0.70291
French	0.68425
Chinese	0.70054
Hindi	0.17540
Korean	0.31668
Arabic	0.64702
Dutch	0.57839

Table 6: Pearson's r value for 10 languages for XLM-T

ter on the training set but relatively low in the four zero-shot languages. This means the models should still need some improvements and can enlarge the pre-training corpus for more information. The XLM-T model performs the best among almost all the 6 training languages except for Chinese, which means the in domain texts can provides more relevant and contextually rich information for the model to learn from. Additionally, the model can be benefited from the multilingual transferring training structure.

For the four zero-shot language tasks, XLM-T works better on Arabic and Dutch, mDeBERTa works the best on Korean and XLM-R performs the best on Hindi. The high performance on zero-shot task may implies a shared linguistic feature among languages in pre-trained corpus.

Our result after stacking ensemble algorithm has a significant improvement on 9 languages except for Hindi. The result shows that each model may capture different aspects of language understanding, and aggregates them in the stacking stage. Also, the combination may mitigate the overfitting phenomenon caused by transformer model.

Language	Pearson's r
English	0.70093
Spanish	0.72084
Portuguese	0.67366
Italian	0.72271
French	0.69328
Chinese	0.73350
Hindi	0.21487
Korean	0.37809
Arabic	0.59555
Dutch	0.60238

Table 7: Pearson's r value for 10 languages after stacking

The ensemble can really produce robust and discover insights from different features among the models. However, the performance of Hindi is not improved, which indicates the corpus of this small language is not enough. We need other methods to analysis it.

7 Conclusion

In this project, I fine-tune four different large language models based on a intimacy regression task, and develop a stacking algorithm to combine the results and get a better result for 10 languages than the baseline methods and individual model. The result is acceptable, but still needs to be improved in some points, such as the further exploration for small languages like Hindi.

8 Other Things We Tried

We tried the basic multilingual BERT model as well. This model has a similar performance with mDistilBERT model, but a relatively low generalizability for zero-shot tasks. This may because that the model is large and leads to some overfitting results. Hence we finally choose mDistilBERT model instead. We also tried BertweetTokenizer to tokenize the tweet texts. However, it seems not support multilingual case for some languages such as Hindi. Then we finally not use it.

9 What I Would Have Done Differently or Next

In this project, I did not adjust the hyperparameters for the LLMs many times. I can choose the hyperparameter and fine-tune the model with different hyperparameters in the future. Additionally, I just use the original data for training, but the corpus

for some small language like Hindi is relatively small so I cannot get a satisfying result in this language. I can apply some data augmentation skills or some sampling skills to enhance the data. Besides, the stacking method can also be explored. I can implement different meta models instead of linear regression, such as random forest and other regression models. The result may be different and I can choose the best one in the future.

References

Harish B, Naveen D, Prem Balasubramanian, and Aarthi S. 2023. CKingCoder at SemEval-2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2009–2013, Toronto, Canada. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Qisheng Cai, Jin Wang, and Xuejie Zhang. 2023. YNU-HPCC at SemEval-2023 task 9: Pretrained language model for multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 733–738, Toronto, Canada. Association for Computational Linguistics.

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.

José Antonio García-Díaz, Ronghao Pan, Salud María Jiménez Zafra, María-Teresa Martn-Valdivia, L. Alfonso Ureña-López, and Rafael Valencia-García. 2023. UMUTeam and SINAI at SemEval-2023 task 9: Multilingual tweet intimacy analysis using multilingual large language models and data augmentation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 293–299, Toronto, Canada. Association for Computational Linguistics.

Pan He and Yanru Zhang. 2023. Zhegu at SemEval-2023 task 9: Exponential penalty mean squared loss for multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 318–323, Toronto, Canada. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Maarten Bos Yozon Liu Leonardo Neves David Jurgens Francesco Barbieri Jiaxin Pei, Vítor Silva. 2023. SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Abel Pichardo Estevez, Jacinto Mata Vázquez, Victoria Pachón Álvarez, and Nordin El Balima Cordero. 2023. I2C-Huelva at SemEval-2023 task 9: Analysis of intimacy in multilingual tweets using resampling methods and transformers. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 758–762, Toronto, Canada. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.