WKU_NLP at SemEval-2023 Task 9: Translation Augmented Multilingual Tweet Intimacy Analysis

Qinyuan Zheng
Wenzhou-Kean University
zhengqin@kean.edu

Abstract

This paper presents a system designed for the SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis. This system is composed of a pretrained multilingual masked language model as a text encoder and a neural network as a regression model. To enhance model performance in low-resource scenarios, the system employs data augmentation techniques leveraging neural machine translation models. Furthermore, I demonstrate the system can be further improved through the ensemble of top-performing models in each language. This system ranks 4^{th} in languages unseen in the training data and 16^{th} in languages seen in the training data. The code and data are accessible vias the following link: https://github.com/Cloudy0219/Multilingual.

1 Introduction

Intimacy is a fundamental aspect of interpersonal relationships within various social settings. Languages inherently encode rich social information pertaining to intimacy, offering opportunities to explore and analyze intimacy through computational linguistics perspectives. Assessing intimacy in a language not only provides insights into the complexities of human communication but also serves as a test for computational models' ability to comprehend human social interactions and the underlying social norms related to intimacy. Assessing intimacy in a language not only provides insights into the complexities of human communication but also serves as a test for computational models' ability to comprehend human social interactions and the underlying social norms related to intimacy.

Social media platforms, such as Twitter, present a treasure trove of diverse interactions with varying degrees of intimacy, making them an ideal source for investigating linguistic intimacy patterns. Pei and Jurgens (2020) introduced a benchmark to examine intimacy interactions on Twitter, comprising English tweets with human-labeled intimacy scores. Pei et al. (2022) further expanded this benchmark to a multilingual setting, covering ten languages, further increasing the scope and relevance of intimacy analysis across diverse linguistic communities.

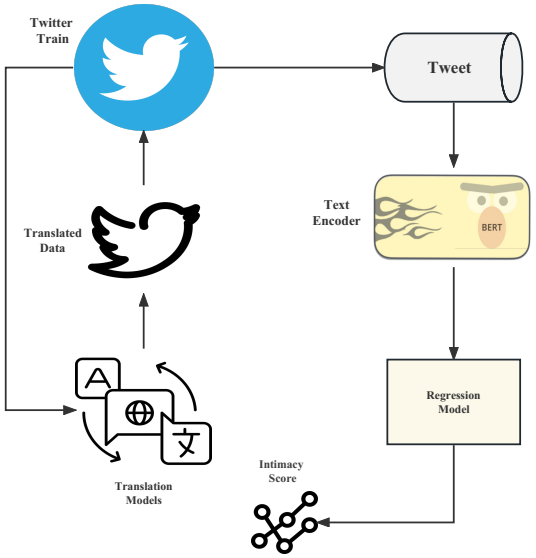

Figure 1: This figure illustrates the architecture of the system. The left part demonstrates the workflow to data augmentation based on the translation model. The right side demonstrates the workflow for intimacy analysis based on the text encoder and regression model.

Despite its significance, intimacy analysis remains a challenging task due to the limited data available and the inherent difficulties for models to interpret and understand human intimacy. This challenge is further intensified in multilingual contexts, as the expression of intimacy varies across languages and cultures. In response, I propose a novel regression model that leverages recent advancements in language modeling and neural networks to capture the semantics of intimacy in multiple languages effectively. To address the issues of limited resources and language differences, I also incorporate data augmentation and model en-

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1525–1530

sembling techniques to enhance my model's performance. By bridging the gap between computational linguistics and the multifaceted nature of intimacy, my approach aims to contribute significantly to understanding human social dynamics across diverse linguistic contexts.

2 System Overview

In this section, I present a comprehensive architecture of my system and elaborate on the implementation details of each component. My system comprises several primary modules: (1) Multilingual Data Augmentation Module; (2) Text Encoder; (3) Regression Model. The system's overall structure is illustrated in Figure 1.

The text encoder transforms tokenized input sentences into high-dimensional representations, embedding the semantic information necessary to determine intimacy scores. As described in Section 2.2, the regression model will predict intimacy scores based on the information contained within these representations.

2.1 Text Encoder

There exist multiple kinds of text encoders, like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), ELMO (Peters et al., 2018), BERT (Devlin et al., 2019), and more. Starting with BERT (Devlin et al., 2019), large pre-trained language models that utilize vast amounts of data have exhibited robust performance across diverse tasks. Models trained using a Masked Language Model (MLM) loss objective (Devlin et al., 2019) have proven to be highly effective encoders, as seen with RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2019). Currently, pre-trained language models trained on extensive data have emerged as one of the top choices for text understanding.

In my approach, I leverage the **TwHIN-BERT** (Barbieri et al., 2020) model as the encoder. This multilingual tweet language model is trained on 7 billion tweets from over 100 distinct languages. Similar to BERT (Devlin et al., 2019), the training objective of this model includes an MLM loss. Moreover, the model introduces an additional loss term in the training objective to better comprehend and integrate the rich social engagements present within a Twitter Heterogeneous Information Network (El-Kishky et al., 2022).

2.2 Regression Model

The regression model ingests the high-dimensional representation generated by the text encoder described in Section 2.1 and predicts an intimacy score based on the information encoded in the representation vector.

This module leverages two layers of the fully-connected neural network to project the representation to a single intimacy score. Given the limited training data, it is crucial to prevent overfitting in the model. To this end, I incorporate a Dropout (Srivastava et al., 2014) layer and a non-linear activation layer ReLU (Nair and Hinton, 2010) to the regressor model.

2.3 Data Augmentation Module

To enhance model performance under low-resource scenarios, I employ a machine translation module to generate extra silver data. The data augmentation module comprises multiple neural machine translation models based on the transformer (Vaswani et al., 2017) architecture. For convenience, I utilize off-the-shelf models [1] trained in OPUS (Tiedemann, 2012) corpus described in Section 3.1.

Given a sample t in the source language, l_s and a target language set $L_t = l_{t1}, l_{t2}, ..., l_{tn}$, where n represents the total number of languages in the test set, the data augmentation process is as follows. First, I enumerate each target language to check if a translation model in the off-the-shelf model library exists for converting from the source language l_s to the target language l_{ti}. If a translation model is available, I translate the sample text into the target language and add it to the training corpus.

2.4 Model Ensemble

Models trained during various phases exhibit distinct proficiencies across different languages. I assess model performance for each language at every epoch and select the model with the highest score for evaluating a specific language. For the unseen model, I choose the checkpoint with the highest average Pearson scores across all languages.

3 Experiment Setup

This section describes the dataset used for training my system and the machine translation model used for data augmentation. I also discuss the details of data processing, model parameters, and evaluation.

[1] https://huggingface.co/Helsinki-NLP

3.1 Dataset

Twitter Data The training data consists of 12,000 tweets in six languages: {English, Chinese, French, Spanish, Italian, Portuguese}. The test set consists of 2000 tweets for four languages not included in the training data: {Korean, Dutch, Hindi, Arabic}.

Parallel Corpus The neural machine translation models are trained on OPUS (Tiedemann, 2012) corpus.

3.2 Data Processing

Due to the nature of the informal language, described in Section 3.1, the data contains noise and unique Twitter formatting. I apply various data processing techniques to remove this noise, as explained in the following paragraphs.

Emojis Twitter data is abundant with emojis. Although language models pre-trained for fine-tuning on Twitter data may have an improved ability to represent tweets with emojis, these emoticons are not always encoded effectively. In order to better express emoji information, I translate every emotion icon using Demoji [2] library into a description of the emoji in English.

Mention Tweets contain mentions of Twitter users by an @ symbol followed by a username. These mentions are not helpful in predicting intimacy scores. What's worse, they can potentially introduce spurious correlations (Tu et al., 2020; Puli et al., 2022) to the distribution of training data. For example, the username may contain semantically relevant information pertaining to intimacy. To remove this noise, I employ string-based matching to filter out and remove these mentions.

Url Links Tweets frequently contain URL links, which are unrelated to intimacy scores. I eliminate URL text based on string matching.

3.3 Model Parameters

The random seed is set to 0, the learning rate is set to 10^{-6} and the batch size is set to 16. For model architecture parameters, the output dimension of the first fully-connected later is set to 768 and the second fully-connected layer to 1024. The max input sentence token number after tokenization is 128.

[2] https://github.com/bsolomon1124/demoji

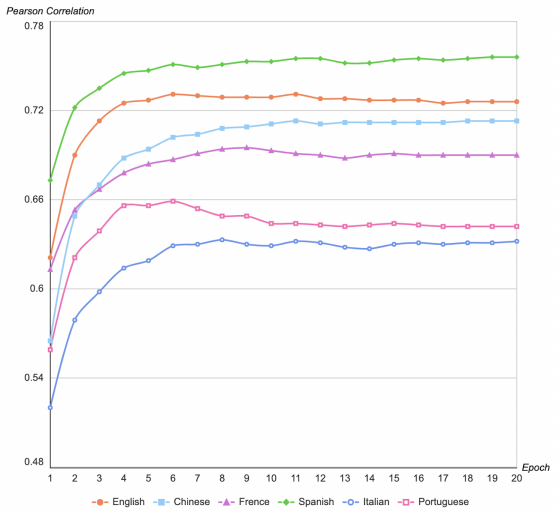

Figure 2: Model performance evaluated on each language individually at each epoch.

3.4 Evaluation

The Pearson Correlation Coefficient is adopted as the evaluation metric. To more accurately assess the performance of the models, I evaluate them separately for each language and conduct specific analyses concerning their generalization capabilities on languages not present in the training set.

Pearson Correlation Coefficient Pearson Correlation coefficient describes the linear relationship and correlation direction between two distributions. Given two distributions, x and y. Pearson correlation coefficient is defined as the quotient of covariance and standard deviation between two distributions as formulated by the following formula:

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

The value of the Pearson correlation coefficient is interpreted as the linear relationship between x and y. As the absolute value of the coefficient increase, the linear relationship between them gets stronger.

3.5 Training Details

For each epoch, I evaluate the Pearson Correlation Coefficient on each language and pick models based on the evaluation results. The performance in each epoch can be found in Figure 2.

Based on the performance curve shown in Figure 2, all models converge after about 10 epochs, and the performance score stabilizes after that. However, performance in Portuguese and Italian

Subset	Language	Score	Rank
Seen	English	0.731	05
	Chinese	0.737	06
	French	0.666	27
	Spanish	0.738	12
	Italian	0.701	19
	Portuguese	0.628	28
Unseen	Hindi	0.234	08
	Dutch	0.551	33
	Korean	0.359	19
	Arabic	0.491	34
Combined	Seen	0.711	16
	Unseen	0.485	04
	Overall	0.555	26

Table 1: Pearson's Correlations (scores) and ranks in the test set reported in the leaderboard.

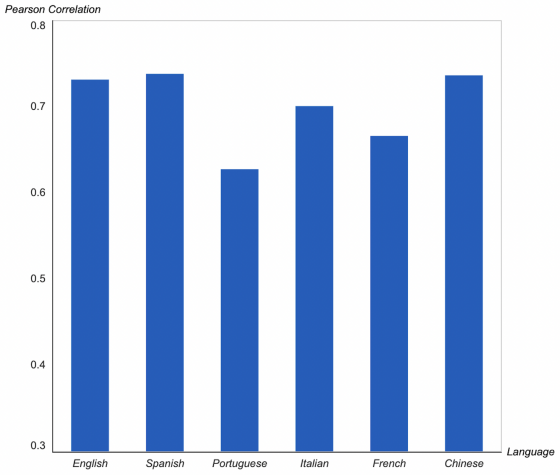

Figure 3: Pearson Correlation score on languages that appears in the training set.

will decrease as training goes on. This phenomenon could be caused by the difference in languages and the habit of expressing emotion and intimacy in these languages. To make sure, I can leverage the best model for each language, I save the best model on each language and perform model ensemble as described in Section 2.4.

4 Results

Model performance is analyzed in three aspects: (1) On languages that appear in the training data; (2) Generalization on languages that are unseen in the training data; (3) Effectiveness of translation data augmentation.

4.1 Leaderboard Scores

In this section, the scores of the test set reported in the leaderboard are shown in Table 1.

4.2 In-Domain Performance

As depicted in Figure 3, the model achieves its best performance in Spanish and its worst in Italian. There are three potential reasons for the varying performance across languages.

First, the amount of training data for each language differs, resulting in an inherent imbalance in the text encoder's ability to process texts. This may subsequently affect the performance of downstream tasks.

Second, intimacy is closely tied to the cultural context embedded in a language. Differences in expressing emotions and intimacy may present varying levels of challenges for intimacy analysis.

Third, the data and intimacy labels themselves could be subject to distribution shifts caused by the subjective factors associated with annotators.

4.3 Generalization to Unseen Languages

Models are evaluated against performance on language unseen in the training data, as illustrated in Figure 4. This evaluation examines the model's generalization capabilities with respect to new languages.

As indicated in Figure 4, model performance is not as good as on languages seen in the training data. However, the model shows great generalization ability in Dutch, which is close to performance in Portuguese, which has been seen in the training data.

The model's weakest performance is in Hindi. There are two possible reasons for this outcome. First, Hindi is a relatively low-resource language in the encoder's training corpus compared to high-resource languages like English and Chinese. Second, Hindi has fewer connections with the languages on which the model is trained. For instance, Korean and Chinese share some cultural connections, which may enhance the generalization performance of Korean. Conversely, Hindi is more distantly related to the languages in the training data than the other unseen languages.

4.4 Translation Data Augmentation

As described in Section 2.3, several neural machine translation models are utilized to generate silver data. In order to study the effect of using silver data to enhance, I use the original training model of the same model conducted research data. The original neural training on Twitter data and the

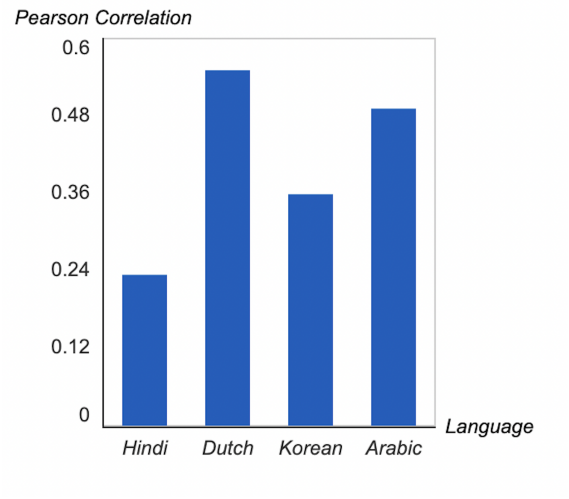

Figure 4: Pearson Correlation evaluation results on unseen languages.

Language	Test		
	Original	+ Aug	Δ
English	0.731	0.719	-0.012
Chinese	0.713	0.724	0.011
French	0.695	0.697	0.002
Spanish	0.756	0.754	-0.002
Italian	0.633	0.645	0.012
Portuguese	0.659	0.659	0

Table 2: Pearson correlation coefficient evaluated on models trained before and after data augmentation using translation module.

use of silver data further enhance the performance of the model, as shown in Figure 2. The results showed that silver data was of great help to the Chinese and Italian, with limited contributions to the French, Spanish, and Portuguese. And silver data led to a drop in English scores.

Translation Model Alternative Various alternative methods can generate silver data using multilingual auto-regression models, such as mT5 (Raffel et al., 2020) and M2M-100 (Fan et al., 2020). Given the volume of data these models are trained on, the resulting models are often more diverse and contain rich semantic information about intimate relationships. There is reason to believe that the system's performance could be further improved using these alternatives. However, due to computational constraints, I have not included them in my study, leaving it as a potential direction for future work.

5 Related Works

The study of language intimacy has garnered significant interest, with a growing focus on understanding and predicting intimacy through computational linguistics methods. Pei and Jurgens (2020) proposed a framework for examining language intimacy, a dataset, and models to assess intimacy levels. Investigating language intimacy necessitates the exploration of social status and power, as the construction and expression of these aspects are based on each individual's unique response to understanding language intimacy while adhering to appropriate social norms (Norona et al., 2013). Danescu-Niculescu-Mizil et al. (2013) and Prabhakaran et al. (2014) discovered that individuals of varying social statuses utilize distinct vocabulary and language strategies to adapt to society, shedding light on relevant research in computational linguistics. Louviere et al. (2015) and Kiritchenko and Mohammad (2016) conducted intimacy tests on 2397 data points using Best-Worst-Scaling.

6 Conclusion

In this paper, I present a system for multilingual intimacy analysis, comprising a pre-trained multilingual masked language model and a regression model designed to project representations into a single intimacy value. To enhance the model's performance in low-resource settings, I employ translation models for data augmentation, which results in improved performance on languages not encountered during training. In order to select the best-performing model, I employ an ensemble of multiple models, each with optimal performance in their respective languages. This system achieves a rank of 4^{th} for unseen languages and 16^{th} for languages present in the training data.

References

Francesco Barbieri, José Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *ArXiv*, abs/2010.12421.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Daniel Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Ahmed El-Kishky, Thomas Markovich, Serim Park, Chetan Kumar Verma, Baekjin Kim, Ramy Eskander, Yury Malkov, Frank Portman, Sofia Samaniego, Ying Xiao, and Aria Haghighi. 2022. Twhin: Embedding the twitter heterogeneous information network for personalized recommendation. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. *ArXiv*, abs/1712.01741.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Jordan J. Louviere, Terry N. Flynn, and Anthony A. J. Marley. 2015. Best-worst scaling: Theory, methods and applications.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*.

Jerika C. Norona, Avril Thorne, Madeleine R. Kerrick, Halley B. Farwood, and Neill Korobov. 2013. Patterns of intimacy and distancing as young women (and men) friends exchange stories of romantic relationships. *Sex Roles*, 68:439–453.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Conference on Empirical Methods in Natural Language Processing*.

Jiaxin Pei, V'itor Silva, Maarten van den Bos, Yozon Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2022. Semeval 2023 task 9: Multilingual tweet intimacy analysis. *ArXiv*, abs/2210.01108.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *North American Chapter of the Association for Computational Linguistics*.

Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. 2014. Gender and power: How gender and gender environment affect manifestations of power. In *Conference on Empirical Methods in Natural Language Processing*.

Aahlad Manas Puli, Nitish Joshi, Hera Y. He, and Rajesh Ranganath. 2022. Nuisances via negativa: Adjusting for spurious correlations via data augmentation. *ArXiv*, abs/2210.01302.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *International Conference on Language Resources and Evaluation*.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.