Arizonans at SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis with XLM-T

Nimet Beyza Bozdag, Tugay Bilgis, Steven Bethard
University of Arizona
{nbbozdag, tbilgis, bethard}@arizona.edu

Abstract

This paper presents the systems and approaches of the Arizonans team for SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis. We finetune the Multilingual RoBERTa model trained with about 200M tweets, XLM-T. Our final model ranked 9th out of 45 overall, 13th in seen languages, and 8th in unseen languages.

1 Introduction

Intimacy is an integral part of human relationships and an essential component of language (Pei and Jurgens, 2020). Pei and Jurgens also propose that intimacy can be modeled with computational methods and that analyzing textual intimacy can potentially reveal information about social norms in different contexts.

With social media taking an increasingly large part in people's lives, intimacy has found itself a new platform for display. In these digital social contexts', Twitter is an important source for analyzing textual intimacy. To further promote computational modeling of textual intimacy, Pei et al. (2023) make use of publicly available Twitter data to create and annotate a textual intimacy dataset called MINT from the following 6 languages: English, Spanish, French, Portuguese, Italian, and Chinese. They also annotate a smaller set of tweets in Dutch, Hindi, Korean, and Arabic to allow for zero-shot prediction experiments. SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis (Pei et al., 2023), asks participants to predict the intimacy of tweets in the aforementioned 10 languages.

In this paper, we explain our approach to intimacy analysis: finetuning XLM-T, the Multilingual RoBERTa model trained with about 200M tweets. To improve performance we experimented with different hyperparameters and changed the training data label distribution by duplicating tweets with certain labels. We discuss the issues we ran into while working on our model, such as the imbalanced range of intimacy scores in the provided training dataset. Our final submission ranked 9th out of 45 overall, 13th in seen languages, and 8th in unseen languages. We release our code at https://github.com/beyzabozdag/tweet-intimacy.

2 Background

Intimacy is an essential part of language, however resources on textual intimacy analysis remain rare (Pei et al., 2023). The first textual intimacy dataset was annotated by Pei and Jurgens (2020) with 2,397 English questions collected mostly from social media posts and fictional dialogues. However, Pei et al. (2023) claim that models trained over phrases in the question structure might not generalize well to text in other forms of languages.

To work on a more generalizable dataset and promote computational modeling of textual intimacy, Pei et al. (2023) proposed their new dataset MINT, which is a Multilingual intimacy analysis dataset covering 13,384 tweets in 10 languages including English, French, Spanish, Italian, Portuguese, Korean, Dutch, Chinese, Hindi, and Arabic. Each tweet in the dataset has a label between 1 and 5 as an intimacy score; 1 indicating "Not intimate at all' and 5 indicating "Very intimate". These scores were generated by taking the mean of all labels given by different annotators.

SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis asks participants to predict the intimacy of tweets in the 10 languages in MINT. They benchmark some of the large multilingual pretrained language models such as XLM-T (Barbieri et al., 2021), XLM-R (Conneau et al., 2019), BERT (Devlin et al., 2018), DistillBERT (Sanh et al., 2019) and MiniLM (Wang et al., 2020), and report that distilled models generally perform worse than other normal models (Pei et al., 2023). The XLM-T model which is a Multilingual RoBERTa model trained with about 200M tweets, is the best-performing model in the task organizers' reports, and the model we experiment with.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1656–1659
July 13-14, 2023 ©2023 Association for Computational Linguistics

Language	Count
Italian	1532
English	1587
French	1588
Spanish	1592
Portuguese	1596
Chinese	1596

Table 1: Distribution of tweets in the training dataset provided by the task organizers.

3 Methodology

We decided that using the best-performing model in the baseline models listed in Pei et al. (2023) should be the starting point and therefore commenced our experiments with XLM-T. We used the `https://github.com/cardiffnlp/xlm-t` quick prototyping Google Colab notebook for fine-tuning.

Our initial goal was to first reach the baseline performance for each language as reported in the task paper (Pei et al., 2023), and then improve from that point. Since there were no data available in the zero-shot languages (Hindi, Korean, Dutch, Arabic), we shifted our focus to training individually on the 6 languages; English, Spanish, Portuguese, Italian, French, and Chinese.

Once we confirmed that we could attain close results to the baseline XLM-T scores for the above-mentioned 6 languages by training the model on each language and testing on that language, we started training the model on all available languages making use of XLM-T's multilingual capacity.

After exploring hyperparameters on the validation data to decide which hyperparameter settings performed the best, we trained a model on the combined training and validation data to make use of all of the provided data.

4 Experimental Setup

The data provided by the task organizers include 9491 annotated tweets in English, Spanish, Portuguese, Italian, French, and Chinese, with almost an equal number of samples from each language, as shown in Table 1. The two languages with the most samples (Chinese and Portuguese) have a total of 1596 tweets, and the language with the least samples (Italian) has a total of 1532 tweets.

Language	Pearson's r of XLM-T	
	Baseline	Finetuned
English	0.70	0.68
French	0.70	0.67
Spanish	0.72	0.71
Italian	0.69	0.64
Portuguese	0.67	0.65
Chinese	0.69	0.69

Table 2: Comparison of XLM-T baseline presented in the task paper and our XLM-T models fine-tuned on each language.

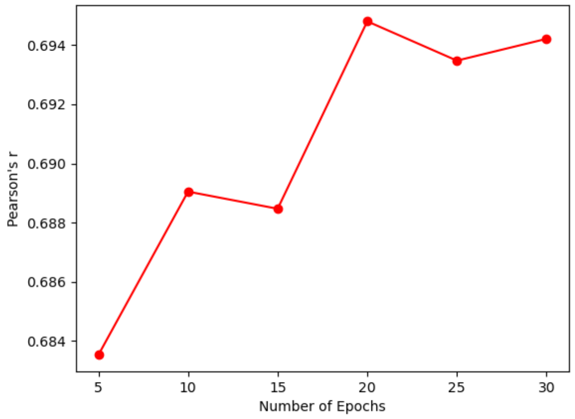

Figure 1: Number of epochs and the Pearson's r achieved

To test each individual language in the dataset, we split a given language into 60% train, 20% validation, and 20% test, setting the random state to 42. This required creating two text files for each split (train, validation, test); one for labels and one for tweets, with the order of labels matching the order of tweets. We set the learning rate to $2e^{-5}$, the number of epochs to 10, batch size to 32, max length to 512, and weight decay to 0.01. We performed all evaluations with `sklearn.feature_selection.r_regression` from the scikit-learn library (version 1.2.1), which computes Pearson's r between the model predictions and the annotated labels.

Table 2 shows that we were able to reach results close to the baseline performance. We then kept the learning rate and the batch size fixed, and observed the effects of different epochs on the accuracy of the model. Figure 1 shows that performance increased until epoch 20, beyond which there were no further gains. We thus picked 20 epochs for tuning our final model for the test set, training on all provided tweets and labels (not just the training

Language	Pearson's r	Rank
English	0.6737381586	28
French	0.7071241905	10
Spanish	0.7352225700	13
Italian	0.7270082535	10
Portuguese	0.6608813413	20
Chinese	0.7107407934	18
Hindi	0.2590025814	4
Dutch	0.6012688745	27
Korean	0.3386983875	21
Arabic	0.6578595143	3

Table 3: Our model's performance on the evaluation dataset.

data we had split out).

5 Results

According to the released evaluation results our submitted model had an accuracy of 0.7129227988 for seen languages, and 0.4441965257 for unseen languages, placing our model 13th and 8th respectively out of 45 submissions, and 9th overall.

For seen languages, our model performed as expected with some improvements for Spanish, Italian, and French relative to the baselines stated by Pei et al. (2023). Table 3 shows that out of all the seen languages, our model performed the worst in Portuguese and English. This is interesting since English and Portuguese are the top two most frequent languages in the 198M tweet corpus which the Twitter-based language model XLM-T is finetuned on (Barbieri et al., 2021)

Another interesting result is the unseen (zero-shot) languages. Our model surpasses the baseline for both Arabic, and Hindi, placing our model on 3rd and 4th for the respective languages.

5.1 Error Analysis

Using the training dataset, we examined the errors of our model. Table 4 shows some errors, along with the difference between our model's predictions and the human judgments.

Figure 2 shows that our model's errors are largest for the tweets that have actual intimacy scores within the 4-5 range. Figure 3 explains this trend since the dataset provided by the task organizers has far fewer examples of tweets with intimacy labels in the 4-5 range. We tried addressing this problem by augmenting our dataset with duplicates of the tweets with intimacy scores in the 4-5 range, but this resulted in a drop in worse performance in our experiments on our train/validation/test split.

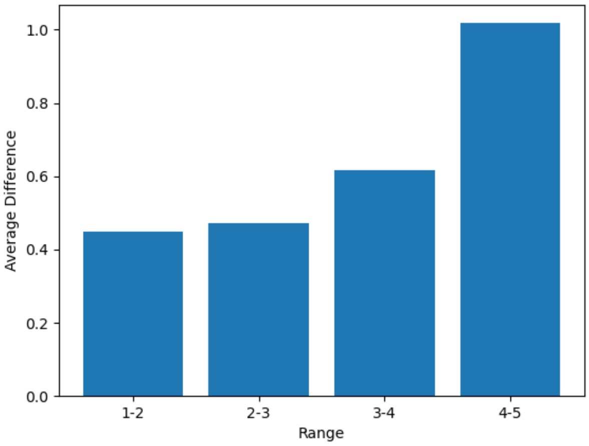

Figure 2: Average difference in predicted and actual intimacy scores by actual intimacy score.

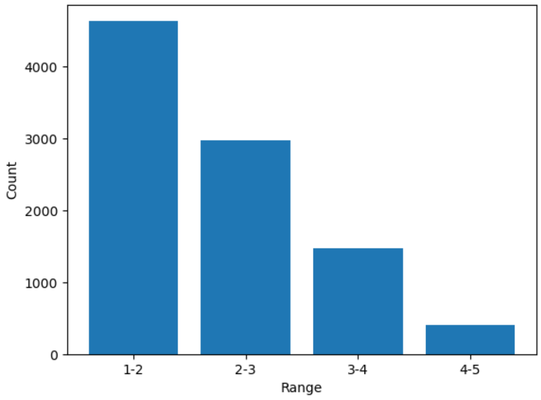

Figure 3: The distribution of intimacy scores in the training dataset provided by the task organizers.

An error analysis we would have liked to perform was to examine the whether our models were worse on the samples where humans had greater disagreement. Intimacy is a subjective matter, so labeling can differ greatly from annotator to annotator, in which case the average of the labels might not accurately reflect annotator judgments of intimacy. However, the task organizers were unable to release the individual annotator judgments to us for this analysis.

6 Conclusion

In our work, we have shown that intimacy prediction on multilingual tweets with the XLM-T model performs well in the languages available in the training dataset, and also the zero-shot languages

Tweet	Actual	Predicted	Difference
@user @user I hope they never google aral sea	3.33	1.119051	2.2142823
@user @user LMFAO. We stood a better chance with the women. That one go cut off your ‚Äútwins‚Äù.	4	1.9799163	2.0200837
@user Ice is bad for you	1.2	3.1258984	1.9258983
@user Aimbot, Cronus, Hacks the works but plenty of positive comments to. Nearly 4k comments man!!	1.5	1.4592953	0.040704727
I feel like I need to delete myself	3	2.954813	0.045186996
@user You admit that me being spoiled is your fault?? I'm keeping this forever and any time you call me a brat I'm going to share it.	3.2	3.2479608	0.04796076

Table 4: Examples of predictions vs actual intimacy scores for English tweets in the training dataset.

Dutch and Arabic. Since the model makes the most errors in predicting higher intimacy scores, future work with balanced training datasets in terms of the number of instances of tweets in different ranges of labels would likely yield more accurate predictions. We also would have liked to incorporate the item-level agreement information into our training process to use as weights, and put more emphasis on labels that the annotators agreed more on. This could help the model make more accurate predictions, and be swayed less by the subjective nature of intimacy perception.

References

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond.

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.

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

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, Online. Association for Computational Linguistics.

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.

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

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.