tmn at SemEval-2023 Task 9: Multilingual Tweet Intimacy Detection using XLM-T, Google Translate, and Ensemble Learning

Anna Glazkova
University of Tyumen
`a.v.glazkova@utmn.ru`

Abstract

The paper describes a transformer-based system designed for SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis. The purpose of the task was to predict the intimacy of tweets in a range from 1 (not intimate at all) to 5 (very intimate). The official training set for the competition consisted of tweets in six languages (English, Spanish, Italian, Portuguese, French, and Chinese). The test set included the given six languages as well as external data with four languages not presented in the training set (Hindi, Arabic, Dutch, and Korean). We presented a solution based on an ensemble of XLM-T, a multilingual RoBERTa model adapted to the Twitter domain. To improve the performance of unseen languages, each tweet was supplemented by its English translation. We explored the effectiveness of translated data for the languages seen in fine-tuning compared to unseen languages and estimated strategies for using translated data in transformer-based models. Our solution ranked 4th on the leaderboard while achieving an overall Pearson's r of 0.599 over the test set. The proposed system improves up to 0.088 Pearson's r over a score averaged across all 45 submissions.

1 Introduction

Intimacy is a significant social aspect of language, which helps to explore existing social norms in various contexts (Pei and Jurgens, 2020). The concept of intimacy describes how an individual relates to his addressee in their perceived interdependence, warmth, and willingness to personally share (Perlman and Fehr, 1987). An automatic evaluation of intimacy in language provides us with a clearer picture of social interactions and different linguistic strategies.

This paper describes a system developed for the SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis (Pei et al., 2023). The purpose of the task was to predict the intimacy of tweets. The
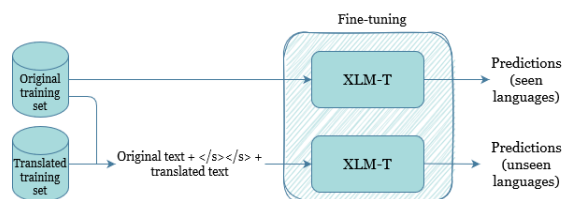

Figure 1: High-level overview of our approach.

test set contained tweets in the languages presented (seen) and not presented (unseen) in the training set. The main challenge of the task was the zero-shot intimacy prediction performance for unseen languages.

Inspired by the recent success of transformer-based models in multilingual tasks (Malmasi et al., 2022; Tayyar Madabushi et al., 2022), we evaluate the performance of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) and its modifications. In our experiments, we investigated the impact of translation to improve the zero-shot performance. We have found that the joint use of original and translated texts increases the model performance for unseen languages in comparison with the separate use of original or translated data.

The paper is organized as follows. Section 2 presents the task and the dataset provided by the task organizers. Our methods are presented in Section 3. In Section 4, we describe the experimental setups that we used during the development phase of the competition. We discuss our results in Section 5. Finally, Section 6 concludes this paper.

2 Task Description

The task aims to predict the intimacy of tweets in ten languages. The organizers presented a training set containing tweets in six languages (English, Spanish, Italian, Portuguese, French, and Chinese) annotated with intimacy scores ranging between 1 and 5 where 1 indicates "not intimate at all" and 5

Language	Train	Test
Amount (avg number of tokens)		
English	1,587 (23.99)	396 (22.79)
Spanish	1,592 (20.68)	399 (20.05)
Italian	1,532 (21.48)	384 (20.71)
Portuguese	1,596 (19.66)	398 (20.69)
French	1,588 (19.36)	393 (20.43)
Chinese	1,596 (29.97)	400 (28.19)
Hindi	-	280 (21.22)
Arabic	-	407 (21.34)
Dutch	-	413 (19.6)
Korean	-	411 (19.37)

Table 1: Data statistics. The number of tokens is obtained using the tokenizer of XLM-T (Barbieri et al., 2022).
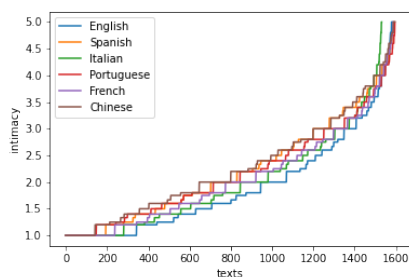

Figure 2: The distribution of intimacy scores for each language (training set).

indicates "very intimate". The model performance was evaluated on the test set consisting of tweets in the given six languages and the four unseen languages that were not included in the training data (Hindi, Arabic, Dutch, and Korean). The total amount of entries in the training set is 9,491. The size of the test set is 3,881. Pearson's r was utilized as an evaluation metric. The dataset statistics are presented in Table 1. A sample of annotated tweets for English is shown in Appendix A. Figure 2 demonstrates the intimacy distribution for the training set.

3 System Overview

3.1 Translated Tweets

Previous studies have shown the effectiveness of the use of translated datasets for zero-shot learning (Schwenk and Douze, 2017; Eriguchi et al., 2018; Ebrahimi et al., 2022). We used the public translation engine Google Translate[1] and the deep-translator Python tool[2] to translate the training and test sets to English. We compared three strategies for representing input data: a) original

dataset, i.e. texts are fed into the model in the same form as they are presented in the original dataset; b) translated dataset, i.e. all texts in all languages are replaced by their translations into English; c) joint dataset, each entry of which contains an original and translated text connected by a sequence of characters *</s></s>*. For English tweets, the original and translated tweets match. The examples of input texts are given in Table 2.

3.2 Models

We compared several pre-trained language models on the tweet intimacy prediction task:

- BERT[3] (Devlin et al., 2019): BERT multilingual base model (cased), a multilingual version of BERT pre-trained on 104 languages.

- XLM-R[4] (Conneau et al., 2019): XLM-RoBERTa base, a model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages.

- XLM-T[5] (Barbieri et al., 2022): Twitter-XLM-Roberta base, a multilingual RoBERTa model trained over 200M tweets.

- RoBERTa[6] (Liu et al., 2019): RoBERTa base model, a modification of BERT that is pre-trained using dynamic masking. Since RoBERTa is a monolingual model, we utilized it only using data translated into English.

All models were implemented using the Simple Transformers[7] library based on Transformers[8] (Wolf et al., 2020) by HuggingFace. We fine-tuned each model for 3 epochs using a batch size of 8, a maximum sequence length of 128, and a learning rate of 4e-5.

The code we used to fine-tune models as well as the translations we made are presented on Github[9].

---

[1] https://translate.google.com/
[2] https://github.com/nidhaloff/deep-translator

[3] https://huggingface.co/bert-base-multilingual-cased
[4] https://huggingface.co/xlm-roberta-base
[5] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base
[6] https://huggingface.co/roberta-base
[7] https://github.com/ThilinaRajapakse/simpletransformers
[8] https://github.com/huggingface/transformers
[9] https://github.com/oldaandozerskaya/intimacy_tmn_semeval23_task9

Original	Translated	Original + translated
j'ai plus aucune force	I have no more strength	j'ai plus aucune force </s></s> I have no more strength
La mia prima stagione e, forse per questo, la mia favorita buon compleanno #Reggina107 http	My first season and, perhaps for this reason, my favorite happy birthday #Reggina107 http	La mia prima stagione e, forse per questo, la mia favorita buon compleanno #Reggina107 http </s></s> My first season and, perhaps for this reason, my favorite happy birthday #Reggina107 http
@user Es normal cuando no se sale bien	@user It's normal when it doesn't work out	@user Es normal cuando no se sale bien </s></s> @user It's normal when it doesn't work out
开学了会更新的慢一点了快万粉了你们有什么想看的吗 http	The school has started, the update will be a bit slower, I am almost 10,000 fans, do you have anything you want to see http	开学了会更新的慢一点了快万粉了你们有什么想看的吗 http </s></s> The school has started, the update will be a bit slower, I am almost 10,000 fans, do you have anything you want to see http

Table 2: Strategies for input representation.

4 Experimental Setup

In this section, we describe the experimental setup utilized during the development phase of the competition. In this phase, the organizers provided participants with a training set containing 9,491 entries. We shuffled the provided data with a fixed random seed and divided it into training and validation subsets in a ratio of 70:30. Therefore, the training and validation subsets consisted of 6,643 and 2,848 entries respectively. To estimate the performance for unseen languages, we consistently excluded one language from the training set. Therefore, the model was fine-tuned on the five remaining languages and evaluated on the unseen language.

To evaluate the results during the development phase, we utilized two metrics. The first one was Pearson's r which was an official evaluation metric. The second metric we used was the mean squared error (MSE) which estimates the mean squared distances between actual and predicted values.

5 Results

5.1 Development Phase

We compared the performance of several pre-trained transformer-based models fine-tuned on the training subset formed during the development phase. The models were evaluated on the validation subset that did not participate in fine-tuning. All languages presented in the validation subset were also presented in the training subset. Therefore, the validation subset did not contain unseen languages.

The results of the comparison are presented in Table 3. The best results in terms of MSE and Pearson's r were achieved by the XLM-T fine-tuned on the original dataset. Similar to the scores obtained by Pei et al. (2023), our results showed the superi-

| Model	MSE	r
Original data		
BERT	0.562	0.18
XLM-R	0.499	0.517
XLM-T	**0.437**	**0.572**
Translated data		
BERT	0.531	0.322
RoBERTa	0.495	0.431
XLM-R	0.488	0.422
XLM-T	0.494	0.534
Original + translated		
BERT	0.545	0.353
XLM-R	0.504	0.467
XLM-T	0.455	0.554

Table 3: Development phase. Evaluation on the validation subset.

ority of the XLM-T over the XLM-R and BERT for the task of multilingual tweet intimacy prediction. Fine-tuning on the translated dataset improves the results of BERT; however, the scores for XLM-R and XLM-T were worse in comparison with the results achieved while fine-tuning on the original dataset. The joint use of original and translated data leads to the highest Pearson's r score among the BERT models. In general, joint fine-tuning outperformed fine-tuning on translated data only but did not surpass the performance of the models fine-tuned on the original texts. The results in terms of MSE scores do not always correspond to the values of Pearson's r. In particular, the XLM-R demonstrated its best MSE and worst Pearson's r using translated dataset.

Further, we evaluated the effectiveness of different input representations for zero-shot tweet intimacy prediction. For this purpose, we excluded languages from the original training set one by one. The models were fine-tuned on the remaining languages and then evaluated on the excluded
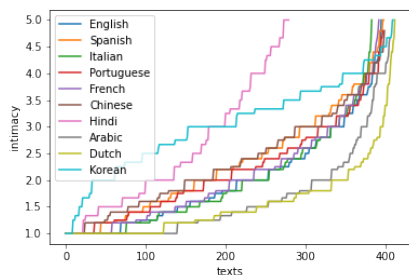
a) gold labels
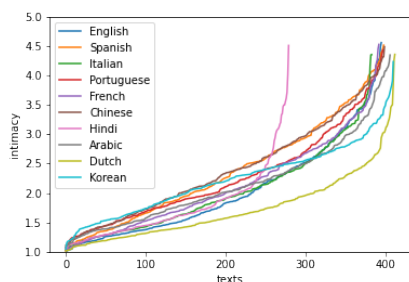
b) predictions

Figure 3: The distribution of intimacy scores for each language (test set).

language. We did not exclude English since we utilized English as an auxiliary language during the formation of input representations. The results are demonstrated in Table 4. The first column presents the types of input representation for training and validation subsets. For example, the first row shows the results obtained for the models fine-tuned and evaluated on the original data.

The lowest Pearson's r scores for the original dataset were obtained for Chinese (0.153), French (0.158), and Italian (0.162). The result for the Portuguese language was more competitive. For Spanish, we achieved 0.579 which is higher than the score obtained on the mixed language validation set (0.572) in Table 3. Fine-tuning on data translated into English as well as fine-tuning on the original dataset and the subsequent validation on translated data in most cases led to a deterioration in the results. However, a significant growth in Pearson's was obtained for Chinese while using the translated training set (+0.23). Slight improvements were also demonstrated for Italian (+0.02 of Pearson's r) and French (+0.069) for the model fine-tuned on translated texts and validated on the original dataset. The best MSE scores for Italian and Chinese were shown on translated data. Taking into account the relatively low Pearson's scores, it can be assumed that the model predicts close to the

average values for any data. The joint fine-tuning using original and translated data was superior compared to the fine-tuning on the original dataset for three of the five languages (Italian, French, and Chinese) in terms of Pearson's r. Moreover, the average Pearson's r was 0.301 for original data and 0.426 for the joint use of original and translated texts. The growth of the averaged Pearson's r using translated texts as auxiliary data was 0.125 in our experiments.

On the basis of our experiments, we combined the XLM-T fine-tuned on the original dataset provided by the organizers of the competition with the XLM-T fine-tuned in a joint manner simultaneously utilizing original and translated texts. The first type of model was used for making predictions on the seen languages (English, Spanish, Italian, Portuguese, French, and Chinese) while the second type was applied to the unseen languages (Hindi, Dutch, Korean, and Arabic).

5.2 Overall Performance

Similar to our previous work on tweet analysis (Glazkova et al., 2021a,b), we applied ensemble learning for the final submission. We used a two-part approach based on the ensembles of the XLM-T fine-tuned on the original and joint data for the seen and unseen languages respectively (Figure 1). We fine-tuned seven models on each data type and took an average across all the models of the same type.

Table 5 shows the official scores of our team. The proposed model ranked 4th at SemEval-2023 Task 9 and obtained an overall Pearson's r of 0.599 over the test set. This result improved up to 0.088 Pearson's r over a score averaged across all 45 submissions. The proposed model showed above-average scores for all languages from the dataset and entered the top ten teams for both seen and unseen languages. The highest results were obtained for Arabic (2nd), Hindi (5th), Italian (6th), and French (9th).

5.3 Error Analysis

Figure 3 shows the distribution of intimacy scores in the official test set and in the predictions made using our approach. From the chart, it can be seen that the maximum values predicted by our model are lower than in the official set. The maximum result predicted by the model is 4.56 and the minimum result is 0.99 while test data contains values between 1 and 5, inclusive. The scores produced

Model	Unseen language									
	Spanish		Italian		Portuguese		French		Chinese	
	MSE	*r*	*MSE*	*r*	*MSE*	*r*	*MSE*	*r*	*MSE*	*r*
Train: original Validation: original	**0.414**	**0.579**	0.452	0.162	0.469	**0.452**	**0.435**	0.158	0.587	0.153
Train: original Validation: translated	0.513	0.205	0.463	0.142	0.46	0.227	0.51	0.101	0.677	0.113
Train: translated Validation: translated	0.461	0.103	0.451	0.11	**0.455**	0.134	0.483	0.123	**0.541**	0.383
Train: translated Validation: original	0.465	0.444	0.471	0.182	0.532	0.19	0.508	0.227	0.599	0.383
Train, validation: original + translated	0.42	0.565	**0.436**	**0.259**	0.462	0.439	0.459	**0.32**	0.577	**0.547**

Table 4: Development phase. Evaluation on unseen languages.

Score	Language												
	English	Spanish	Portuguese	Italian	French	Chinese	Hindi	Dutch	Korean	Arabic	Seen	Unseen	Overall
r	0.717	0.74	0.684	0.734	0.708	0.721	0.242	0.639	0.361	0.662	0.637	0.357	0.599
Rank	11	10	12	6	9	15	5	9	17	2	9	9	4
avg *r*	0.635	0.664	0.596	0.636	0.611	0.629	0.187	0.539	0.407	0.524	0.727	0.441	0.511

Table 5: Official results.

by the model are closer to the average scores. The average value for the official scores is 2.13 with a standard deviation of 0.97. For our predictions, the average value is 2.11 and the standard deviation is 0.74.

The correlation between the predictions and ground truth test labels is shown in Figure 4. In general, the results for the seen languages are predictably better. The lowest correlation among all languages was obtained for Hindi and Korean.

6 Conclusion

In this work, we have explored an application of pre-trained language models and translated data for the task of multilingual tweet intimacy detection. We showed that the joint use of original texts and texts translated into English improves the zero-shot performance of XLM-T for unseen languages. Further research in zero-shot intimacy prediction could be conducted to estimate the quality of the translations used and the assessment of the proximity of languages from training and test sets.

References

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
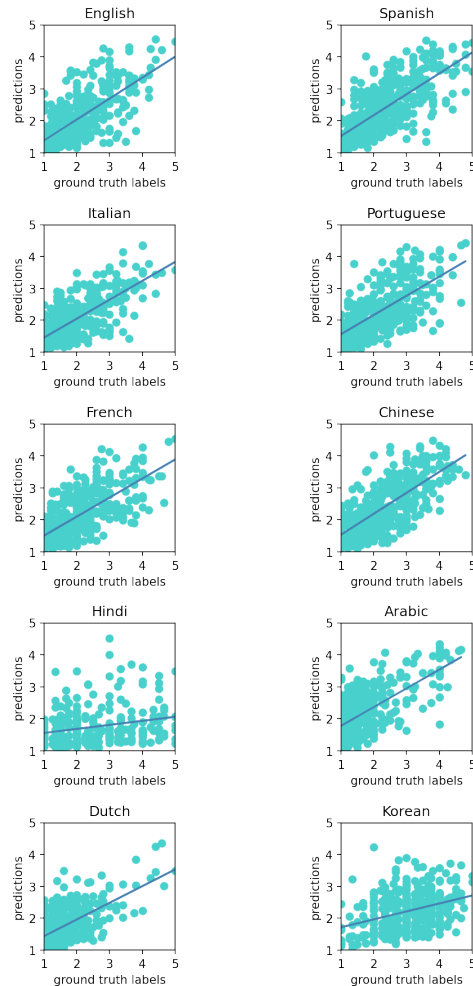
Figure 4: The correlation between predictions and ground truth test labels.

258–266, Marseille, France. European Language Resources Association.

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. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686.pdf*.

Anna Glazkova, Maksim Glazkov, and Timofey Trifonov. 2021a. g2tmn at constraint@ AAAI2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, pages 116–127. Springer.

Anna Glazkova, Michael Kadantsev, and Maksim Glazkov. 2021b. Fine-tuning of pre-trained transformers for hate, offensive, and profane content detection in English and Marathi. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*, pages 52–62. CEUR Workshop Proceedings.

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 preprint arXiv:1907.11692*.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. SemEval-2022 task

11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

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. 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.

Daniel Perlman and Beverley Fehr. 1987. The development of intimate relationships.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 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, Online. Association for Computational Linguistics.

A Appendix

Tweet	Intimacy
Here is a nice equation: 0+0-0-0+0=0	1.0
Try that again Timmy... Ill turn you into bone-meal http	1.4
@user @DoorDash Damn Shame, you tipped a $1.30	2.25
@user @verge I can't believe you don't value family	3.0
i'll be yours through all the years, till the end of time. #MewSuppasit @MSuppasit #My-CandyHeroxMew	4.0
fren: who you smiling about me: nothing, just the love of my life the love of my life: http	4.4

Table 6: A sample of tweets from the dataset.