ROZAM at SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis

Mohammadmostafa Rostamkhani, Ghazal Zamaninejad, Sauleh Eetemadi
Iran University of Science and Technology
{mo_rostamkhani97, gh_zamaninejad@comp.iust.ac.ir, sauleh@iust.ac.ir

Abstract

We build a model using large multilingual pre-trained language model XLM-T for regression task and fine-tune it on the MINT (Multilingual INTmacy) analysis dataset which covers 6 languages for training and 4 languages for testing zero-shot performance of the model. The dataset was annotated and the annotations are intimacy scores. We experiment with several deep learning architectures to predict intimacy score. To achieve optimal performance we modify several model settings including loss function, number and type of layers. In total, we ran 16 end-to-end experiments. Our best system achieved a Pearson Correlation score of 0.52.

1 Introduction

Intimacy has long been considered as the main dimension of human relationships (Maslow, 1981; Sullivan, 2013; Prager, 1995).

Although intimacy has a significant role in language, there are few datasets in this field. The first textual intimacy dataset (Pei et al., 2023) contains 2,397 English questions collected through social media. However, models trained on this dataset may not generalize well to other situations.

MINT dataset (Pei et al., 2022) is a textual intimacy dataset covering 10 languages which is collected from people's tweets and annotated with an intimacy score of 1 to 5. The task is to predict intimacy scores of tweets for 10 languages. Six languages are seen during training including English, Spanish, Portuguese, Italian, French, Chinese and 4 other languages (Hindi, Korean, Dutch, and Arabic) are not included in training data. These four languages are used for evaluating zero-shot performance of proposed models. By doing this task, we can extract some social information inside the sentence and hence calculate how intimate it is.

XLM-RoBERTa (Conneau et al., 2019) is a multilingual pre-trained language model, pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages (it is a multilingual version of RoBERTa) and it was pre-trained with the Masked Language Modeling (MLM) objective. The main strategy is to use XLM-T (Barbieri et al., 2021) which is a XLM-Roberta-base model trained on about 198M multilingual tweets.

Our model contains XLM-T for extracting embeddings of sentences and also some dense and dropout layers for predicting (approximating) the intimacy score of a sentence. We fine-tune the last few layers of XLM-T and train the regressor part of our model using the training dataset. Since our dataset was relatively small, we also used some data augmentation methods. Eventually, teams are ranked based on Pearson Correlation metric calculated on overall score for seen and unseen languages. The proposed method was ranked 33 out of 45. The details of our implementations are available through our github repository.

2 Background

The inputs to our model were tweets collected from twitter and we expect our model to produce a number between 1 to 5 as an intimacy score as final output. A schematic of our model is shown below:

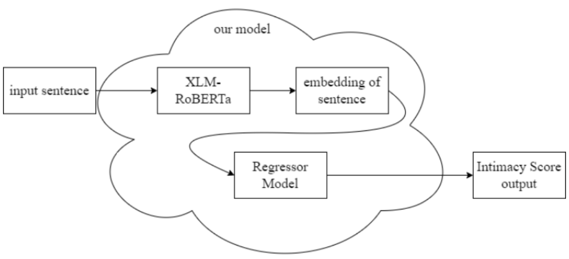

Figure 1: System overview

Output of XLM-T which is an embedding of the input sentence is a tensor of shape (768*1) is fed to the regressor model to predict a number as an intimacy score. Furthermore, some augmentation methods were added for zero-shot languages in order to improve the model performance.

3 System Overview

3.1 Preprocessing Data

We use MINT as our main dataset. Since we have no training data for zero-shot languages, we translate tweets from one language to zero-shot languages using Google Translate API to train our model on them for achieving better model performance. However, we faced some issues. For example, we don't want the '@user' token (which are mentioning non-verified-users) to be translated. So we decided to remove all the mentions in the translated sentences. We choose English tweets for translating to zero-shot languages for preparing our data. We assume translated sentences have the same intimacy scores. Furthermore, we translate all tweets of seen languages to English and add them to our dataset. After augmentation, we store all data in a csv file and feed it as input to our model for training and validation.

3.2 Creating Dataset

For creating the dataset we store tokenized sentences and intimacy scores. We use XLM-T Tokenizer for tokenizing sentences. We save tokenized sentences for feeding to the model. It also has special tokens for emojis which are used frequently in tweets.

3.3 Model

XLM-T is chosen because it was pre-trained on multilingual tweets containing emojis as desired. The structure of the model is as follows: one or two dense layers with ReLU as the activation function followed by a dropout layer and another dense layer for generating one number as output of the regression task. We fine-tune the model with four different settings. First, we freeze all the XLM-T blocks and just train the regressor part. In the second setting, we just fine-tune the last dense block of the model and freeze all other blocks. In the third setting, blocks 0 to 10 were frozen and we fine-tune block 11 and the last dense block. Finally, in the forth setting, we freeze blocks 0 to 9 of XLM-T. Due to hardware limitations, we were unable to experiment with more settings. We use different loss functions such as MSE (mean square error) and negative of Pearson Correlation and also MSE+Pearson Correlation because predictions should be as accurate as possible while maintaining high Pearson Correlation. We experiment with different loss functions to examine which one performs best.

- MSE (mean square error)

$$\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2 \tag{1}$$

- negative of Pearson Correlation: We need to minimize the loss, so instead of maximizing Pearson Correlation, its negative is minimized. We can see Pearson Correlation equation below:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2}$$

  where r is correlation coefficient, x_i is values of the x-variable in a sample, $\bar{x}$ is the mean of the values of the x-variable, y_i is the values of the y-variable in a sample and $\bar{y}$ is the mean of the values of the y-variable.

- MSE-Pearson Correlation: We want predictions to be as accurate as possible and have high Pearson Correlation simultaneously.

AdamW is chosen as the optimizer. We freeze blocks of 0 to 9 of XLM-T.

3.4 Post-processing Data

We use different kinds of post-processing because the output of our model was in different ranges. When the output of our model was in the range of 1 to 5, we didn't apply post-processing. When the output of our model was in the range of -1 to 1, we used shifting and scaling as post-processing. When the output of our model was in the range of 10 to 50, we used scaling as post-processing. We also clipped output of our model when it was less than 1 or greater than 5. The final output must be in the range of 1 to 5.

4 Experimental setup

The data is splitted into three groups with the portion of 0.9, 0.05, 0.05 for training, testing and validation.

Learning rate of 0.0001 is chosen after some trial and error. Choosing a large learning rate causes

Experiment	Dataset size	Number of training epochs	Learning rate	Freezed blocks	Train loss	Validation loss
1	original	5	1e-3	All blocks	0.0344	0.0322
2	original	5	1e-3	0-11	0.0321	0.0305
3	augmented	5	1e-5	0-11	0.0302	0.0281
4	augmented	6	1e-5	0-10	0.0262	0.0235
5	augmented	6	1e-4	0-10	0.0254	0.0231
6	augmented	6	1e-4	0-9	**0.0225**	**0.0227**

Table 1: First set of experiments using MSE loss

Model	**XLM-T**	**BERT**	**XLM-R**	**DistillBERT**	**MiniLM**
English	**0.70**	0.59	0.65	0.55	0.61
Spanish	**0.73**	0.62	0.64	0.61	0.67
Portuguese	**0.65**	0.54	0.61	0.52	0.53
Italian	**0.70**	0.57	0.67	0.58	0.62
French	**0.68**	0.55	0.63	0.54	0.57
Chinese	0.70	0.65	**0.72**	0.67	0.65
Hindi	**0.24**	0.09	0.24	0.17	0.18
Dutch	0.59	0.47	**0.60**	0.44	0.57
Korean	0.35	0.32	0.33	0.26	**0.41**
Arabic	**0.64**	0.35	0.48	0.32	0.38
Overall	**0.58**	0.48	0.53	0.52	0.53

Table 2: Performance of the baselines. The bottom four rows are tested under the zero-shot setting. (Pei et al., 2022)

overshoot from optimum point and choosing small one may cause the optimizer to get stuck in local minima.

For the number of epochs for training, 10 is chosen. Choosing a small number of epochs is because of the large language model that has been used, and was pre-trained on tweets which are the same data as our dataset.

Batch size of 32 is chosen. Layers of 0 to 9 of XLM-T were freezed.

loss function	MSE	MSE(-1,1)	MSE(10,50)	MSE Pearson	MSE Pearson(-1,1)	MSE Pearson(10,50)
English	**0.681**	0.665	0.646	0.674	0.657	0.674
Spanish	0.713	0.686	**0.722**	0.703	0.698	0.708
Portuguese	0.648	0.63	0.638	0.640	0.648	**0.650**
Italian	0.679	0.627	**0.683**	0.673	0.656	0.669
French	0.683	0.655	0.681	**0.685**	0.649	0.675
Chinese	**0.692**	0.674	0.685	0.681	0.675	0.671
Hindi	0.235	0.131	0.181	0.160	0.162	**0.244**
Dutch	0.520	0.502	0.540	0.500	0.524	**0.530**
Korean	**0.348**	0.333	0.318	0.347	0.331	0.296
Arabic	0.525	0.5695	0.544	0.514	0.534	**0.581**
seen	**0.691**	0.668	0.687	0.685	0.675	0.685
unseen	0.289	0.274	0.279	0.282	0.287	**0.322**
Overall	0.514	0.493	0.508	0.508	0.502	**0.520**

Table 3: Second set of experiments using different loss functions. Numbers show **Pearson Correlation**

	1 hidden 20 neurons	1 hidden 40 neurons	2 hidden 20 neurons	2 hidden 40 neurons
English	0.674	0.651	0.646	**0.680**
Spanish	**0.708**	0.694	0.647	0.696
Portuguese	**0.650**	0.627	0.618	0.620
Italian	**0.669**	0.659	0.614	0.650
French	**0.675**	0.662	0.645	0.674
Chinese	0.671	**0.684**	0.683	0.657
Hindi	**0.244**	0.138	0.214	0.149
Dutch	0.530	**0.547**	0.541	0.515
Korean	**0.296**	0.267	0.267	0.284
Arabic	**0.581**	0.542	0.553	0.542
seen	**0.685**	0.674	0.653	0.671
unseen	0.322	0.301	**0.353**	0.260
Overall	**0.520**	0.507	0.516	0.494

Table 4: Third set of experiments using different Regressor models. Numbers show **Pearson Correlation**

5 Results

At first, we ran 5 experiments with different setups to find the best hyper-parameters. Table 1 shows that using augmented dataset with learning rate 10^{-4} and frozen layers 0-9 got the best results.

In the next step, we used the same model that achieved the best results in the first set of experiments. Table 2 shows the performance of baselines. Table 3 indicates the best results obtained on ***seen languages***, ***unseen languages*** and ***overall*** are respectively related to ***MSE***, ***MSE-Pearson(10,50)*** and ***MSE-Pearson(10,50)***.

Finally, we decided to use *MSE-Pearson(10,50)* loss as it got the best Pearson Correlation in the previous step and run a new set of experiments with different number of hidden layers and neurons. Table 4 shows the best results attained on ***seen languages***, ***unseen languages*** and ***overall*** are respectively related to ***1 hidden + 20 neurons***, ***2 hidden + 20 neurons*** and ***1 hidden + 20 neurons***.

As we can see in Table 5, there are some cases which our model predicts well and some that our model predicts poorly. As an instance, *"Full list of companies currently doing business in Russia: http"* annotated as 1.00 which implies that it is not intimate at all and our model could predict it well. As another example, for *"Know this my Angel: It is impossible to not wanting to kiss you whenever you smiles."*, the true label is 4.25 showing it is very intimate but our model predicts that it is almost not intimate. The reason of failure might be the

Text	Label	Prediction
Dear 2021 could you not have taken Fauci instead of Betty White?	1.40	1.39
Full list of companies currently doing business in Russia: http	1.00	1.00
Working hard for something we don't care about is called stress. Working hard for something we love is called passion.	2.20	2.19
@sapnapalt @GeorgeNootFound delete this. rn	1.60	1.60
@user yes i would say maid of honor but that is my baby sister	2.50	2.49
@user @user You say?	1.75	1.76
Aite. I'm done being single… I changed my ways	3.75	3.72
@user Probably start something new to keep my creativity going with ideas	2.75	1.75
Know this my Angel: It is impossible to not wanting to kiss you whenever you smiles.	4.25	2.02
@user You only JUST started doing this?	1.0	2.93

Table 5: Samples of predictions

small size of our dataset. Furthermore, our model could not fine-tune well. Also, most of the tweets annotated as intimate in the original dataset include profane words and might cause the model to get biased toward profane words. The latter mentioned tweet does not include any profane word, so our model could not predict well.

6 Conclusion

In this paper we present some models with different hyper-parameters for MINT dataset for SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis and try to optimize the model with different loss.

Also we augment the dataset using translating existing tweets to other languages which affects the results. Choosing the right learning rate also has effects on converging faster and the results. Training the last few layers of XLM-T improves results.

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. *CoRR*, abs/1911.02116.

Abraham Harold Maslow. 1981. *Motivation and personality*. Prabhat Prakashan.

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*.

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.

Karen Jean Prager. 1995. *The psychology of intimacy*. Guilford Press.

Harry Stack Sullivan. 2013. *The interpersonal theory of psychiatry*. Routledge.