Multilingual Tweet Intimacy Analysis

Sugam Garg
sgarg6@ucsc.edu

Pranjali Basmatkar
pbasmatk@ucsc.edu

Priyesh Vakharia
pvakhari@ucsc.edu

Abstract

This paper describes our system for the Semeval 2023 Task 9 - Multilingual Tweet Intimacy Analysis. We highlight the impact of various tweet cleaning strategies and augmenting the training data with translations. We improve the pearson's score provided by the authors of the task by atleast 15% on all languages, including 4 languages evaluted in zero-shot setting.

1 Introduction

Multilingual Tweet Intimacy Analysis is the task 9 (Pei et al., 2022) of the SemEval 2023. The goal of the task is to predict the intimacy level of tweets in 10 languages. The training data for 6 of these 10 languages - English, Spanish, Italian, Portuguese, French, and Chinese have been provided in the task release. The other 4 languages - Hindi, Dutch, Korean and Arabic will be evaluated under zero-shot setting. The authors provide results on 5 multilingual models - XLM-T (Barbieri et al., 2021), XLM-R (Conneau et al., 2019), MiniLM (Wang et al., 2020), MBert (Devlin et al., 2018), and DistillBERT(Sanh et al., 2019). These results are duplicated in Table 1.

We conduct our experiments with XLM-T and XLM-R as our foundational models. XLM-R (Conneau et al., 2019) has been proven to be a powerful multilingual language model across various tasks as compared to other models such MiniLM. Even in this task's description paper, XLM-T and XLM-R outperformed other models in almost all languages as can be seen from Table 1. In our approach, we reproduce the results described in the paper, conduct an error analysis of the two model's results and experiment with various strategies such as data cleaning, data augmentation and increasing model parameters.

From our experiments, we found the data augmentation approach to yield the most significant improvements in the model performance. Further, through this task, we were also able to explore the impact of this augmentation on zero-shot languages described above. Finally, with this approach, we got an improvement of greater than **0.15** Pearson coefficient score on all the languages with even greater performance gain on zero-shot languages.

2 Background

Intimacy is a fundamental aspect of how humans relate to one another in social settings. Despite its long-term presence, there has been little work to quantify the expression of intimacy in the textual data. Intimacy detection may be used to improve various aspects of conversational AI that often offer monotonous behaviour (Hovy and Yang, 2021). Existing studies (Pei and Jurgens, 2020) suggest that intimacy is an essential component of language and can be modeled computationally. (Danescu-Niculescu-Mizil et al., 2013) quantify the expression of intimacy in textual data using lexical cues and linguistic strategies like accommodation to express the perceived status of text in relation to others. This is the first task to study textual intimacy in a multilingual scenario.

While recent methods (Le et al., 2019), (Ralethe, 2020) have modeled individual languages using BERT-based masked language modeling objective, training individual language models have large data and computational requirement. To overcome this challenge, a multilingual language model is trained on large data from multiple languages with the assumption that low-resource languages may benefit from high-resource languages due to

Language	XLM-T	XLM-R	BERT	MiniLM	DistillBERT
Spanish	0.70	0.63	0.58	0.60	0.54
Italian	0.72	0.63	0.69	0.63	0.60
English	0.67	0.62	0.58	0.53	0.53
Chinese	0.69	0.65	0.55	0.59	0.55
Portuguese	0.70	0.62	0.56	0.59	0.55
French	0.69	0.72	0.65	0.64	0.62

Table 1: Performance of the baselines from task release paper.

shared vocabulary, genetic relatedness (Nguyen and Chiang, 2017), or contact relatedness.

This task aims to set a multilingual benchmark for detecting intimacy in textual data and use it to test the ability of computational models to understand social information across various low and high-resource languages.

3 Approach

3.1 Understanding the data

The SemEval 2023 task provides us with a custom dataset of publicly available tweets extracted from Twitter (public media platform) containing multilingual textual data. Various tweets were sampled from 2018 to 2022 belonging to the 10 languages specified. The following preprocessing steps were included to maintain the quality of data:

- Fasttext (Joulin et al., 2016) was used to identify languages with confidence scores larger than 0.8 to assign the language labels.
- Various mentions of users were replaced with a standard '@user' tag to remove randomness.

The dataset was annotated by a diverse set of adult annotators with a 'first language' requirement for quality assurance. A set of 2000 annotated tweets was selected from each of the languages to ensure equal representation. The final intimacy score was calculated as the mean score of all the labels for each tweet. Additionally, a split-half-reliability test conducted on the tweets by randomly splitting labels into two groups and calculating the Pearson correlation between the aggregated scores from the two groups resulted in an average SHR score of 0.68.

The final dataset contains 13,384 tweets, out of which 9790 tweets have been made public for training purposes (containing tweets from the 6 languages), and the remaining data (also containing the additional 4 languages) will be released in January for testing. For the sake of our experiments, we conduct a thorough exploratory data analysis to understand the specifications of the data. We observe the following:

Dataset	Data points
Training samples	7832
Test samples	1958

Table 2: Train/test split of the data samples.

Data Label analysis

The labels of the data lie between a score of 1 and 5. The mean for the label score is 2.0. Of the entire data, 2777(28%) tweets have a score greater than 2.5.

Language distribution analysis

All 6 languages have approximately 1500 data points each, indicating no imbalance in the language data. Further, we observe the following text noises in tweets:

- Presence of mentions - '@user' tags indicating usernames
- Presence of Emoji(s) in tweets
- Presence of trailing 'http' tag
- Presence of Twitter slang such as hashtags, and repetitive chars such as 'Heyyyyyyy' or 'am so Happppppyyyyy'
- Slang alphabets such as 'a':'@', 'i':'!', etc.

We spilt the given data into a train and test sets using an 80-20 split rule with stratification to avoid the effects of class imbalance. The details of the split dataset can be seen in Table 2.

Model	$XLM-R^1$	$XLM-T^1$	$XLM-R^2$	$XLM-T^2$	$XLM-T^2$
Batch Size	128	128	64	64	128
Spanish	0.63	0.72	0.62	**0.726**	0.72
Italian	0.65	**0.69**	0.51	0.587	0.613
English	0.63	0.7	0.55	0.751	**0.758**
Chinese	**0.72**	0.69	0.65	0.657	0.66
Portuguese	0.62	0.67	0.63	**0.715**	0.705
French	0.62	**0.7**	0.59	0.627	0.581

Table 3: Comparing results obtained on our reproduction of $XLM-R$ and $XLM-T$ models with task description paper. $XLM-R^1$ refers to the result posted by original paper on $XLM-R$, $XLM-R^2$ refers to the results obtained by us.

3.2 Problem Formulation

The multilingual tweet intimacy analysis task requires us to quantify the level of intimacy of a tweet in multiple languages. We formulate this task as a regression problem - predicting an intimacy score between 1 and 5, where 1 represents low level of intimacy and 5 represents high levels of intimacy. To evaluate the performance of the multilingual model, we use Pearson's coefficient score. This score measures the linear relationship between two data points and ranges between -1 to 1 where 1 or -1 implies a completely linear relationship while 0 indicates no correlations.

3.3 Modeling Approach

As highlighted in the introduction, XLM-R and XLM-T outperform other models in almost all languages. Thus, we choose these two models as our foundation models and conduct our experiments on these models.

XLM-R : Multilingual RoBERTa model

The RoBERTa base model is a transformer-based masked language model trained on one hundred languages using more than 2TBs of filtered commoncrawl data. It outperforms multilingual BERT (mBERT) on various cross-lingual benchmarks and performs well in low-resource languages such as Urdu and Swahili.

XLM-T: Multilingual RoBERTa model finetuned over 200M tweets

Drawing on the generalizing capabilities of the RoBERTa base model, the XLM-T model finetunes the Twitter data. The model has been trained on millions of tweets in over 30 languages. The paper suggests that domain-specific, fine-tuned models can provide consistent performance as compared to large general-domain models.

We attempt to reproduce the performance of these two models on the training data provided. We compare the performance of these two models provided by the task description paper and our training in Table 3. The task description provided only two hyper-parameters for training the models - learning rate of 0.001 and batch size of 128. On a learning rate of 0.001, we obtained a model with zero pearson's score and thus we tuned our own learning rate and obtained best results with a learning rate of 0.0001. Please note that the results are different as the test split was not provided by the authors.

3.4 Error analysis

We conduct an error analysis on the results obtained from the XLM-T model trained on batch size of 64 as this was our best performing model. We generated an error value based on the absolute difference between the predicted score and the ground truth label. We sampled the instances($\sim 13\%$) where this value was greater than 1 for further analysis. After our analysis, we grouped the errors into the following two categories:

- **Wrong labels** This category of error focuses on cases where the annotated value for the tweet intimacy score might be wrong. Examples of such cases can be found in Table 4. In this table, we observe that an unreasonably high value of intimacy score is given to these tweets which have low intimacy.
- **Twitter Language Specific errors** This category focuses on cases where the presence of Twitter-specific slang in the tweets affected the performance of the model. Examples of such tweets can be seen in Table 5. We observed that the additional noise in the tweet negatively impacted the performance of the model.

Original Tweet	Language	English Translation	Target
mon frère ça fait longtemps jte voyait plus	French	My brother, it's been a long time I haven't seen you.	4.2
You are an optimist and I sure hope you are right!	English	N/A	3.25
ottima risposta HAHAHA	Italian	Great answer HAHAHA	3.2

Table 4: Error Analysis : Wrong label

Original Tweet	Language	Target	Predicted
i'll be yours through all the years, till the end of time. #MewSuppasit #MyCandyHeroxMew	English	4	2.4
People are stupid apparently and that's not fromsofts fault	English	2.75	1.64

Table 5: Error Analysis: Twitter language-specific errors

4 Experiments

As mentioned in the previous section, we classify erroneous results as either Wrong Label or Twitter Language Specifics. In the latter category of tweets, we discovered slang, colloquials and other pieces of texts such as emoticons, hashtags and alphanumeric words. We experimented with different cleaning actions to handle each of these language irregularities. The cleaning actions are explained in the following subsections.

4.1 Removing mentions and trailing http

Tweets are filled with many pieces of text that provide little to no information about intimacy. Username mentions of the form '@name' is one such example. From a human perspective, a mention generally should not provide any information about the intimacy of the tweet. While there might be a correlation due to data leakage, this is something we should prevent. Hence, this cleaning action consists of removing all mentions from every tweet. Following the reasoning above, we decided to remove all 'http' trailing text from every tweet.

4.2 Cleaning alphanumeric text

Some tweets contain numeric characters which indicate either a misspelling or slang. Either way, we remove numeric characters from the tweets and study their impact.

4.3 Handling emoticons

Emoticons in their native form are of no use to the model. In unicode, every emoji is encoded as a unique alphanumeric key. We experiment with two actions to handle emoticons. In the first, we simply extract emoticons based on the unicode ranges and remove them. In the second approach, we use the Demoji[1] library that translates every emoticon into a description of the emoji in English Language.

4.4 Data Augmentation

We have $\sim$ 1500 tweets for a given language to train and evaluate the model. To counter the lack of data, we augment the data provided by translating a tweet in all other languages using the m2m100 model (Fan et al., 2020). As noted in the data exploration, training data for only 6 languages was provided. Thus, we use the translations of the tweets in the 6 languages to create an evaluation set for the 4 languages to be evaluated in a zero-shot setting.

We perform various experimental iterations based on the data cleaning techniques described above, the hyper parameters - batch size and learning rate, and the translation augmented data. The details of these iterations are provided below :

- **Itr 1:** XLM-T, Batch Size: 64, Learning Rate: 0.0001
- **Itr 2:** XLM-T, Batch Size: 128, Learning Rate: 0.0001

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

Language	Itr 1	Itr 2	Itr 3	Itr 4	Itr 5	Itr 6	Itr 7	Itr 8	Itr 9
Spanish	0.726	0.72	0.77	0.77	0.707	0.729	0.76	0.71	**0.862**
Italian	0.587	0.613	0.578	0.58	0.595	0.598	0.58	0.6	**0.84**
English	0.751	0.7587	0.655	0.67	0.749	0.74	0.66	0.74	**0.796**
Chinese	0.657	0.66	0.675	0.67	0.235	0.6379	0.67	0.65	**0.689**
Portuguese	0.715	0.705	0.718	0.71	0.696	0.711	0.73	0.7	**0.846**
French	0.627	0.581	0.699	0.68	0.618	0.622	0.69	0.6	**0.725**

Table 6: Results of experiments

Language	Paper XLM-T	Zero-Shot XLM-T	Zero-Shot XLM-T with augmentation
Hindi	0.21	0.568	**0.687**
Korean	0.33	0.489	**0.712**
Dutch	0.59	0.723	**0.814**
Arabic	0.64	0.563	**0.721**

Table 7: Performance of XLM-T model under zero-shot setting. (Please note: the paper results and our results cannot be compared as they are scored on different data. The values provided are just for reference.

- **Itr 3:** XLM-T, Batch Size: 64, Learning Rate: 0.0001, Demoji
- **Itr 4:** XLM-T, Batch Size: 128, Learning Rate: 0.0001, Demoji except for English
- **Itr 5:** XLM-T, Batch Size: 128, Learning Rate:0.0001, Emojis removed, text cleaned
- **Itr 6:** XLM-T, Batch Size: 128, Learning Rate: 0.0001, text cleaned, emojis kept
- **Itr 7:** XLM-T, Batch Size: 64, Learning Rate: 0.0001 with linear layer and RELU activation, Demoji and text cleaned
- **Itr 8:** XLM-T, Batch Size: 64, Learning Rate: 0.0001 with linear layer and RELU activation, Demoji, no text cleaning
- **Itr 9:** XLM-T Batch Size: 64, Learning Rate: 0.0001, with translation augmented data

5 Results

In this section, we present the outcome of the experiments described in the previous section. Table 6 highlights the outcome of various experiments we ran on the XLM-T model. As can be seen from the table, the most significant performance improvement was observed with the model trained on the translation augmented data. With this approach, we observed that there was little improvement in Chinese language. This is primarily due to the poor translation quality observed from other languages to Chinese with the m2m100 model.

Further, the model's performance on the languages to be evaluated in zero-shot setting is provided in Table 7. We can see from this table that the model trained on the augmented training set outperforms the one without by a considerable margin.

6 Conclusion

In this report, we highlight the efficacy of different data cleaning strategies and data augmentation on the Multilingual Tweet Intimacy Analysis task. From our experiments, we gained a performance of over $\sim 15\%$ on all languages using the data augmentation strategy. Further, we highlight the effectiveness of this strategy under zero-shot evaluation in 4 different languages.
We also noted that there is no performance gain in Chinese language using the above strategy. In future, we will be exploring other translation models and APIs which have shown better performance in Chinese language.

7 Team Contribution

All the contributors in this project made an equal contribution to the literature survey, baseline setup, model training, tuning and report generation stages.

References

Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. 2021. A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250*.

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. *arXiv preprint arXiv:1911.02116*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. *arXiv preprint arXiv:2011.03020*.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozon Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2022. Semeval 2023 task 9: Multilingual tweet intimacy analysis.

Sello Ralethe. 2020. Adaptation of deep bidirectional transformers for afrikaans language. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2475–2478.

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

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. *Advances in Neural Information Processing Systems*, 33:5776–5788.