NLP-LISAC at SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis via a Transformer-based approach and Data Augmentation

Abdessamad Benlahbib 1 , Hamza Alami 2 , Achraf Boumhidi 1 , Omar Benslimane 3

¹ LISAC Laboratory, Faculty of Sciences Dhar EL Mehraz, USMBA, Fez, Morocco
² School of Computer Sciences, Mohammed VI Polytechnic University, Ben Guerir, Morocco
³ LAMA Laboratory, Faculty of Sciences Dhar EL Mehraz, USMBA, Fez, Morocco
abdessamad.benlahbib@usmba.ac.ma, hamza.alami@um6p.ma
achraf.boumhidi@usmba.ac.ma, omar.benslimane@usmba.ac.ma

Abstract

This paper presents our system and findings for SemEval 2023 Task 9 Tweet Intimacy Analysis. The main objective of this task was to predict the intimacy of tweets in 10 languages. Our submitted model (ranked 28/45) consists of a transformer-based approach with data augmentation via machine translation.

1 Introduction

According to the Cambridge Dictionary¹, intimacy is things that are said or done only by people who have a close relationship with each other. Tweet Intimacy is a term that is used to describe the relationship between Twitter users and the way they interact with one another through the social media platform. It refers to the level of personal connection and familiarity that exists between individuals who engage with each other on Twitter. This can be determined by various factors, such as the frequency of interaction, the types of content shared, and the use of personal language or inside jokes. A high level of tweet intimacy suggests that the individuals have a strong, personal relationship, while a low level of tweet intimacy indicates more casual or impersonal interactions.

In this paper, we present our findings on SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis (Pei et al., 2023). Our method consists of a transformer-based approach (Vaswani et al., 2017) along with data augmentation.

The rest of the paper is structured in the following manner: Section 2 provides the main objective of the Task. Section 3 describes our system. Section 4 details the experiments. And finally, Section 5 concludes this paper.

2 Task Description

This task aims at using a collection of tweets in six languages (English, Spanish, Italian, Portuguese, French, and Chinese), which have been marked with scores ranging from 1 to 5 based on their level of intimacy to predict the intimacy of tweets in 10 different languages. The model's effectiveness will be assessed on the test set consisting of the aforementioned six languages, as well as an additional test set that contains four languages not present in the training data (Hindi, Arabic, Dutch, and Korean).

3 System Description

To tackle the SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis, we have fine-tuned a transformer-based model on an augmented training dataset. The different steps of our system are described as follows:

- We have translated the English tweets in the training dataset to Hindi, Arabic, Dutch, and Korean using the Google translate API for Python: Googletrans² by assigning the same intimacy score in the English tweets to the translated tweets, then, we concatenated them with the training dataset of the current task.
- We have used the question intimacy dataset (Pei and Jurgens, 2020) which contains 2247 English questions from Reddit as well as another 150 questions from Books, Movies, and Twitter. The intimacy scores in this dataset range from -1 to 1, so we applied the linear function y = 2x + 3 to map the intimacy scores to the 1-5 range, then, we concatenated them with the training dataset.

¹https://dictionary.cambridge.org/ dictionary/english/intimacy

²https://py-googletrans.readthedocs. io/en/latest/

• We have fine-tuned the Twitter-XLM-Robertabase (Barbieri et al., 2022) model on the augmented training dataset.

4 Experimental Results

We experimented our model on the SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis dataset. The experiment has been conducted in Google Colab environment³, The following libraries: Transformers - Hugging Face⁴ (Wolf et al., 2020), and Keras⁵ were used to train and to assess the performance of our model.

4.1 Datasets

The training set contains 1596 Portuguese tweets, 1596 Chinese tweets, 1592 Spanish tweets, 1588 French tweets, 1587 English tweets, and 1532 Italian tweets. The test set contains 1410 Korean tweets, 1396 English tweets, 1396 Spanish tweets, 1390 Portuguese tweets, 1389 Dutch tweets, 1382 French tweets, 1368 Arabic tweets, 1354 Chinese tweets, 1352 Italian tweets, and 1260 Hindi tweets. Figures 1 and 2 depict the language distribution of the tweets in the training set before and after data augmentation, while Figure 3 shows the language distribution of the tweets in the test set.

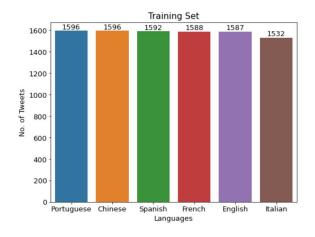


Figure 1: Language distribution of the tweets in the training set before data augmentation

4.2 Evaluation Metric

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between -1 and 1 that measures the strength and direction of the relationship between

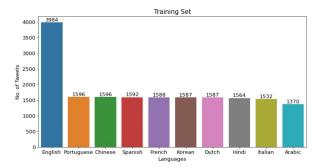


Figure 2: Language distribution of the tweets in the training set after data augmentation

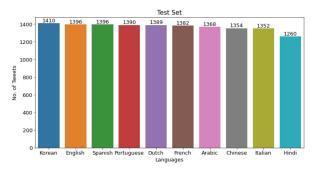


Figure 3: Language distribution of the tweets in the test set

two variables. Pearson's r is used as the evaluation metric where:

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

We denote *n* as the sample size. x_i, y_i are the individual sample points indexed with *i*. $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ (the sample mean); and analogously for \bar{y} .

4.3 Experimental Settings

During the fine-tuning of Twitter-XLM-Robertabase model, we set the hyper-parameters as follows: 10^{-5} as the learning rate, 12 epochs, 200 as the max sequence length, and 32 as batch size. Table 1 summarizes the hyperparameters settings of Twitter-XLM-Roberta-base model.

4.4 System Performance

The organizers benchmarked several baseline models on the tweet intimacy prediction task (Pei et al., 2023). They compared 5 multilingual pre-trained language models: multilingual BERT (Devlin et al., 2019), multilingual RoBERTa: XLM-R (Conneau et al., 2019), multilingual RoBERTa model trained over 200M tweets: XLM-T (Barbieri et al., 2022),

³https://colab.research.google.com/

⁴https://huggingface.co/docs/transformers/index

⁵https://keras.io/

Hyperparameters	Settings	
Learning rate	10^{-5}	
Batch size	32	
Epochs	12	
Max sequence length	200	
Optimizer	Adam	
	(Kingma and Ba, 2015)	
Loss	Mean Squared Error	

Table 1: Hyperparameters settings for the model in the experiments

multilingual distilled BERT (Sanh et al., 2019), and multilingual MiniLM (Wang et al., 2020). Table 2 shows the performance of the baselines as reported by the organizers (Pei et al., 2023). and Table 3 depicts the results of our proposed system on SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis test set. As we can see, our approach achieves similar results to the baselines (Pei et al., 2023). We also noticed that our system slightly outperforms the baseline models for Spanish and Italian tweets, and slightly underperforms the baseline models for the remaining tweets, which shows that the use of machine translation as a data augmentation technique doesn't improve the performance of tweet intimacy analysis.

Language	XLM-T	BERT	XLM-R	DistillBERT	MiniLM
English	0.70	0.58	0.63	0.54	0.60
Spanish	0.72	0.60	0.63	0.60	0.63
Portuguese	0.67	0.58	0.62	0.53	0.53
Italian	0.69	0.55	0.65	0.55	0.59
French	0.70	0.56	0.62	0.55	0.59
Chinese	0.69	0.65	0.72	0.62	0.64
Korean	0.33	0.36	0.32	0.32	0.39
Hindi	0.21	0.04	0.20	0.15	0.16
Dutch	0.59	0.44	0.60	0.42	0.54
Arabic	0.64	0.36	0.49	0.29	0.43

Table 2: Performance of the baselines (Pei et al., 2023). The bottom four rows are tested under the zero-shot setting

5 Conclusion

Tweet intimacy refers to the level of personal connection and emotional closeness that individuals express through their tweets. Research has shown that tweet intimacy can be influenced by factors such as the language used, the content of the tweets, and the relationship between the tweeter and their audience (Pei et al., 2023). Tweet intimacy has become an important area of study for researchers interested in understanding how individuals use

Language	Score	Ranking
English	0.6767195177	27
Spanish	0.7231124024	21
Portuguese	0.6648634542	16
Italian	0.710760164	16
French	0.6709944734	23
Chinese	0.6998464728	27
Korean	0.3285271079	25
Hindi	0.1918843595	33
Dutch	0.5965411825	28
Arabic	0.6251867203	17

Table 3: Results of our proposed approach on SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis test set

social media to form and maintain relationships. By examining tweet intimacy, researchers can gain insights into the ways in which social media is changing the nature of human communication and social interaction.

In this paper, we described our approach for tackling SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis. Our submitted system consisted of a transformer-based approach along with data augmentation and secured the 28th position among a total of 45 teams.

As social media continues to play an increasingly important role in our daily lives, it is likely that research on tweet intimacy will continue to be an important area of investigation.

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 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.

- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.
- 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. *ArXiv*, abs/1910.01108.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep selfattention distillation for task-agnostic compression of pre-trained transformers.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.