Final Project Proposal
SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis

Orion Lowy UHID:1449248

Aisha Farooque UHID:1809992

Jordan Yu UHID:1663303

1 Problem Statement

For our final project, we will work on the Multilingual Tweet Intimacy Analysis task, originally proposed by Jiaxin Pei, Francesco Barbieri, Vítor Silva, Maarten Bos, Yozon Liu, Leonardo Neves, and David Jurgens for the SemEval 2023 workshop. The goal of this task is to train ML models to recognize "intimacy" in text communications, which the authors define as "closeness and interdependence, self-disclosure, and warmth or affection" expressed in the language used to communicate.

We will use two datasets, the Reddit Questions dataset and the Multilingual Tweet Intimacy dataset, compiled by Pei et al., to study whether knowledge about the intimacy level of text communication can be transferred more easily from tweets to questions or vice versa.

2 Datasets

We will use two datasets in this project. The first dataset, Reddit Questions (https://blablablab.si.umich.edu/projects/intimacy/), contains approximately 2000 questions in English, taken from the online discussion website Reddit, and annotations scoring the level of intimacy they express.

The second dataset is the Multilingual Tweet Intimacy dataset (https://arxiv.org/abs/2210.01108), which contains a set of around 9,000 tweets in six languages (English, Spanish, Italian, Portuguese, French, and Chinese) annotated with intimacy scores.

Since the test dataset for this task will be released in 2023, we will split the training data into a slightly smaller training set and evaluation and test sets. We will use Pearson's r as the evaluation metric.

3 Methodology

We will start by downloading a RoBERTa model, pretrained on unlabeled language datasets. Then we will make three copies of it. We will train the three models on the Reddit Questions set, the Multilingual Tweets dataset, and a combination of the two datasets, respectively.

Since the labels are on a scale of -1 to +1 in the Reddit Questions set and from 1 to 5 in the Multilingual Tweets set, we will transform the labels to match a uniform scale across both datasets. We will prepare the data by tokenizing and extracting features, then train each of the first two models on one of the datasets and the third one on a concatenation of both datasets. We will conduct experiments to fine-tune the training parameters, getting feedback using the validation split of the same dataset on which we are training the model.

After training, we will test the models trained on only one dataset on the test split of the same dataset they were trained on. Then we will test them on the other dataset they were not trained on to see if the knowledge they learned from one type of data is transferable to the other type of data without additional specific training.

We will also test the model trained on both datasets on the individual test set for each dataset to see if the combined information helps the model detect intimacy better in either type of text.

4 References

[1] 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. Association for Computational Linguistics.

[2] Jiaxin Pei, Vítor Silva, Maarten Bos, Yozon Liu, Leonardo Neves, David Jurgens and Francesco Barbieri. 2022. SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis In arXiv:2210.01108

[3] Francesco Barbieri, Luis Espinosa Anke, Jose Camacho-Collados. 2021. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In arXiv:2104.12250