SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis

Mohammadmostafa Rostamkhani - 401722235

Ghazal ZamaniNejad - 401722244

Introduction:

In this task, we are going to predict **textual intimacy of multilingual tweets**.

Dataset:

This dataset is **multilingual** text data which was collected from **twitter**. The tweets are sampled from **2018 to 2022**.

The dataset is called **MINT** (Multilingual INTimacy dataset) which covers tweets in **6 languages** as the **training** data, including **English**, **Spanish**, **French**, **Portuguese**, **Italian**, and **Chinese**, covering major languages used in The Americas, Europe, and Asia.

A total of **12,000 tweets** are annotated for **6 languages**. For testing the **generalization** of model under **zero-shot** setting, they also annotated small **test sets** for **Dutch**, **Korean**, **Hindi**, and **Arabic** (**500 tweets for each** of those).

For collecting this dataset, they use the ***lang_id*** in the tweet object to select **English** and **Chinese**. For **other** languages they use **fastText for language identification** and assign labels when the **model confidence** is **larger than 0.8**. All the mentions of **unverified users** are replaced with a special token "**@user**" during pre-processing to remove noise from random and very infrequent usernames.

For annotating tweets they recruited some **annotators**. They set a "**first language**" requirement during annotator pre-screening.

Intimacy is annotated using a **5-point Likert Scale** where **1** indicates "**not intimate at all**" and **5** indicates "**very intimate**". For each language in the training set, they collected annotations for 2,000 tweets. **Each tweet was annotated by 7 annotators and each**

annotator was shown 50 tweets. Also the annotators were balanced by sex and they were from 73 unique counties and regions.

For post-processing they remove annotations from users who failed the attention test. They also remove potential noise in the crowdsourcing setting, similar to trimmed mean. They **removed one highest score and one lowest score for tweets with at least five labels**. Then they **kept** the tweets with **at least two valid scores**.

For **external test languages** (i.e. Dutch, Hindi, Korean, and Arabic), they only kept tweets with a **relatively low label diversity** (i.e. **standard deviation lower than 1**) to ensure a **good golden test set for the zero-shot setting**. The **final intimacy score** is calculated as the **mean score of all the remaining labels** for each tweet.

The final dataset includes **13,384 tweets** annotated with the textual intimacy score.

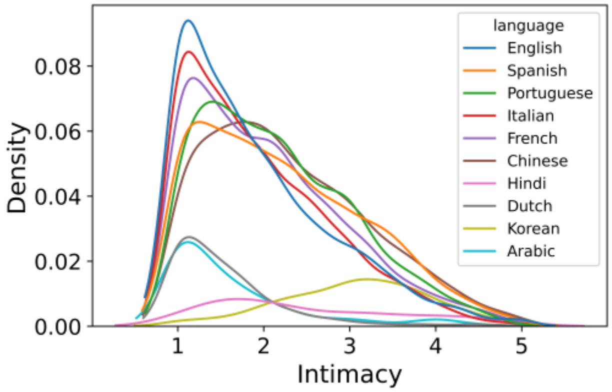
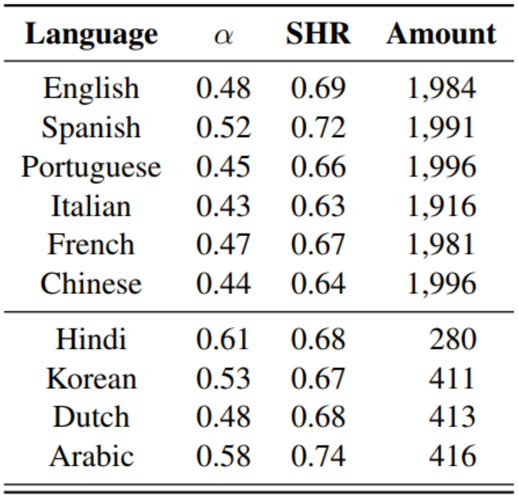


Figure 1: The distribution of intimacy scores for each language

To **verify the quality** of the annotations, they conduct a **split-half-reliability test (SHR)** which randomly splits labels into **two groups** and calculates the **Pearson correlation** between the aggregated scores from the two groups. All the **SHR scores** are **above 0.63 with an average of 0.68**. Suggesting that the final aggregated scores are **reliable**. Then split all dataset (13,384 tweets) into **training**, **validation**, and **test** sets following a ratio of **7:1:2**.



Language	α	SHR	Amount
English	0.48	0.69	1,984
Spanish	0.52	0.72	1,991
Portuguese	0.45	0.66	1,996
Italian	0.43	0.63	1,916
French	0.47	0.67	1,981
Chinese	0.44	0.64	1,996
Hindi	0.61	0.68	280
Korean	0.53	0.67	411
Dutch	0.48	0.68	413
Arabic	0.58	0.74	416

Table 2: Statistics for the annotated dataset

Since the dataset is collected from Twitter, it is **conversational** and **informal**.

Since recognizing the intimacy of text can serve as an important benchmark to test the ability of computational models to understand social information, we are interested in this project.

Since the dataset was **ready**, we **don't use any API** for collecting it, we just download it and use it.

Since annotating textual intimacy is **challenging** because of the **subjective nature of intimacy perception** and **potential individual rating bias**, we have limitations for collecting more data.

Examples for analysis:

Work work work,1.0,English:

- We agree with this example. Because it's not intimate at all.

@user hmm 🙃,1.3333333333333333,English

- We also agree with this example. It's a bit more intimate with respect to the previous example.

he wants to have a partyyy 😭😭😭😭,2.0,English

- We agree with this. It’s a bit more intimate with respect to the previous example.

@user @user Come and join! @user @user @user,3.0,English:

- We disagree. It’s not so intimate.

i would love to be dead rn,4.0,English:

- We disagree. It’s not very intimate.

You'll miss me when I'm gone,3.333333333333333,English

- We agree. It’s quite intimate.

@user @user Love you 💗💗,4.0,English:

- We agree. It’s very intimate.

I loved u,4.2,English

- We agree. It’s very intimate.

need a kiss 😔,4.75,English:

- We agree. It’s very intimate.

@user I think I fell in love with you,4.8,English

- We agree.