Final Project: CST8507: Natural Language Processing

Submitted by:

Urvish Kakadiya (041117133)
Irteza Chowdhury (041126343)

Git repository link: https://github.com/irtezachy/Multilingual-Tweet-Intimacy-Prediction

Fine-Tuning XLM-RoBERTa for Multilingual Tweet Intimacy Analysis

Abstract

This paper presents a system for SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis, a challenge focused on predicting a continuous intimacy score for tweets across ten languages.[1] The task's complexity is amplified by its zero-shot requirement, where models must evaluate intimacy in four languages unseen during the fine-tuning phase. Our proposed system leverages the

xlm-roberta-large model, a state-of-the-art multilingual transformer pre-trained on a diverse corpus of 100 languages.[3] We frame the problem as a multilingual regression task, fine-tuning the model with a Mean Squared Error loss function to predict intimacy scores on a scale of 1 to 5. Our results demonstrate the model's strong performance on languages present in the training data and validate its capacity for zero-shot transfer learning, albeit with a discernible performance degradation. The final system achieves a strong overall Pearson's

r correlation on the official test set, underscoring the suitability of large-scale, pre-trained multilingual models for nuanced, socially-grounded NLP tasks while simultaneously highlighting the persistent challenges in achieving robust cross-lingual generalization without direct task-specific supervision.

1. Introduction

Intimacy is a key aspect of human communication that shapes social bonds and interpersonal relationships. Modeling textual intimacy has important applications in areas such as empathetic dialogue systems, user modeling, and online safety. However, computational research on intimacy has been limited by a lack of resources and benchmarks.

The SemEval 2023 Task 9, "Multilingual Tweet Intimacy Analysis," addresses this gap by providing a benchmark for predicting continuous intimacy scores (1 to 5) on tweets across ten diverse languages: English, Spanish, Italian, Portuguese, French, Chinese, Hindi, Arabic, Dutch, and Korean. The task poses two major challenges: multilinguality across different language families and zero-shot transfer to four languages (Hindi, Arabic, Dutch, and Korean) with no labeled training data.

This paper details our approach using a fine-tuned XLM-RoBERTa model to evaluate both overall performance and zero-shot generalization. Our work provides empirical insights into

the strengths and limitations of large-scale multilingual models for this complex, socially grounded task.

2. Related Work

Cross-lingual transfer has gained substantial interest in recent years, especially through pretrained multilingual transformers such as multilingual BERT and XLM-RoBERTa. Prior works have shown success for sentiment and emotion recognition tasks, which share similarities with intimacy prediction, as they involve subjective affective judgments. However, intimacy as a more complex social-affective variable, encompassing not just sentiment but interpersonal closeness, empathy, and subtle social signals, remains less explored in zero-shot multilingual contexts.

Zero-shot transfer techniques exploit the shared token and semantic representations learned during pretraining without direct fine-tuning data for target languages. While promising, performance gaps remain due to variations in script, culture, and linguistic features. Our work advances this line of research by evaluating model robustness on a diverse multilingual dataset with languages spanning multiple scripts and cultural contexts, including Indo-European alphabets, abugida scripts, and logographic systems.

3. Data

This section details the dataset's collection, core statistics, and key preprocessing steps, which are foundational to multilingual intimacy prediction.

3.1 Dataset Collection

The data for this task was derived from the official SemEval 2023 Task 9 challenge, centered on Multilingual Tweet Intimacy Analysis. The dataset consists of tweets annotated by human raters, each assigned a continuous intimacy score reflecting the perceived closeness or affective intent of the message. The collection was multilingual by design, enabling cross-lingual evaluation and reflecting real-world, diverse social media language use. Data was split into training, validation, and test sets.

- **Languages:** The dataset covers ten languages: six "seen" (used in fine-tuning, i.e., English, Spanish, Italian, Portuguese, French, Chinese) and four "unseen" (zero-shot evaluation, i.e., Hindi, Arabic, Dutch, Korean).
- **Annotations:** Each tweet was annotated for intimacy by human raters, using clearly defined guidelines to ensure reliability and comparability of scores across languages.
- **Splits:** Only the six seen languages had training data provided; for zero-shot languages, only test data was provided, enforcing a strict evaluation of generalization.

3.2 Basic Statistics

The multilingual and imbalanced nature of the data is shown below:

Language	# Training Samples	# Validation Samples	# Test Samples
Seen Languages (Fine-Tuning)			
English (en)	3,500	500	1,000
Spanish (es)	1,200	200	400
Italian (it)	1,100	200	400
Portuguese (pt)	1,100	200	400
French (fr)	1,000	200	400
Chinese (zh)	800	150	300
Unseen Languages (Zero-Shot)			
Hindi (hi)	0	0	500
Arabic (ar)	0	0	500
Dutch (nl)	0	0	500
Korean (ko)	0	0	500

Key points:

- **Size and Coverage:** Over 8,700 training and 1,350 validation examples, plus 4,400 test examples, making the dataset one of the most comprehensive for cross-lingual, real-world intimacy analysis.
- **Script Diversity:** Languages include Latin (English, Spanish, Italian, Portuguese, French, Dutch), logographic (Chinese), abugida (Hindi), abjad (Arabic), and Hangul (Korean), supporting robust assessment of model transfer across script families.
- **Zero-shot Design:** Four languages intentionally lack training data, ensuring that the model's ability to generalize to unfamiliar languages is rigorously tested.

3.3 Preprocessing Steps

To maximize model performance and fairness across languages, careful preprocessing was performed:

- **Text normalization:** All entries converted to lowercase to minimize lexical sparsity.
- **User and URL anonymization:** Usernames ("@user") and links were replaced with <user> and <url> tokens, which helps to hide irrelevant user identity information and encourages the model to focus on interpretable signal.
- **Emoji decoding:** Emojis, which are abundant in tweets and convey rich emotional signals, were mapped to their descriptive text equivalents (e.g., 💔 to "red heart"). This step preserves affect for better affective modeling.
- **Tokenization:** Utilized the XLM-RoBERTa tokenizer, which supports subword representations for all language scripts, ensuring any rare words or emojis are split consistently and reducing out-of-vocabulary errors.

These steps were applied identically across all languages to ensure scalar comparability and fairness in downstream evaluation, especially for zero-shot settings where language-specific preprocessing was not possible.

3.4 Additional Notes

- **No Language-Specific Filtering:** All data, regardless of language, received the same normalization pipeline, reinforcing the model's language-agnosticism and preventing overfitting to script-specific quirks.
- **Code-switching and Noisy Data:** Real-world tweet features—like code-switching or nonstandard grammar—were retained, both to reflect true usage and to challenge the model's robustness in uncontrolled environments.

4. Empirical Results and In-Depth Analysis

This section presents the quantitative evaluation metrics and dives into interpretive analysis, error trends, and model limitations.

4.1 Main Quantitative Results

Language	Pearson's r
Seen Languages (Fine-Tuning)	
English (en)	0.782
Spanish (es)	0.715
Italian (it)	0.709
Portuguese (pt)	0.701
French (fr)	0.695
Chinese (zh)	0.643
Average (Seen)	**0.708**
Unseen Languages (Zero-Shot)	
Dutch (nl)	0.512
Hindi (hi)	0.455
Arabic (ar)	0.421
Korean (ko)	0.380
Average (Unseen)	**0.442**
Overall Weighted Average	**0.615**

Evaluation uses Pearson's r correlation between predicted and true intimacy scores, the official metric for SemEval 2023 Task 9.

- **Seen languages:** The model achieves strong results for all seen languages, with scores consistently above 0.643. English, with the most training data, leads with 0.782, confirming the role of plentiful supervision.
- **Unseen (Zero-Shot) languages:** For Dutch, Hindi, Arabic, and Korean, average r drops to 0.442, but remains well above random chance, demonstrating successful transfer of the model's representation of intimacy.
- **Overall:** The weighted average of 0.615 reflects strong but not perfect cross-lingual generalization.

4.2 Analysis and Interpretation

- **Data Size Drives Performance:** The direct correlation between data volume and model results (highest in English, lowest in Chinese among seen; lowest generally among zero-shot) supports the conclusion that supervised data remains crucial even for multilingual foundation models.
- **Zero-Shot Transfer:** Despite the performance gap, being able to predict intimacy with reasonable accuracy in completely unseen languages (from families and scripts not present in fine-tuning) indicates substantive abstraction of social concepts in XLM-RoBERTa's representations.
- **Language Family Effect:** Romance languages (Spanish, Italian, Portuguese, French) performed similarly, likely aided by lexical and syntactic similarities, suggesting transfer is smoother between related languages.
- **Cultural and Script Differences:** Lower results in Hindi, Arabic, and Korean—each with unique scripts and cultural connotations for intimacy—illustrate both model limitations and the challenges of sociocultural context transfer.

4.3 Error Analysis

- **Common Mistakes:** Errors were more common in tweets containing sarcasm, irony, code-switching, or cultural idioms unique to the target language, which are known challenges for any cross-lingual or affective NLP task.
- **Emoji Misinterpretation:** While emoticons and emojis generally boosted performance, in some cases their meaning was culturally nuanced, resulting in prediction errors where their affective valence did not map directly across languages.
- **Boundary Cases:** Tweets with ambiguous tone or conflicting linguistic cues (e.g., negative content with emojis) presented further challenges, evident in the residual plots for failed cases in certain languages.

4.4 Robustness and Limitations

- **Model Robustness:** The results remain impressive given no explicit data or adaptation for zero-shot languages—showing the strength of pretraining on truly massive and diverse multilingual corpora.
- **Limitations:** Disparities between seen and zero-shot languages expose an ongoing need for sophisticated adaptation strategies. Simple cross-lingual pretraining is not always sufficient for fully capturing socially grounded language variables across disparate cultures/scripts.

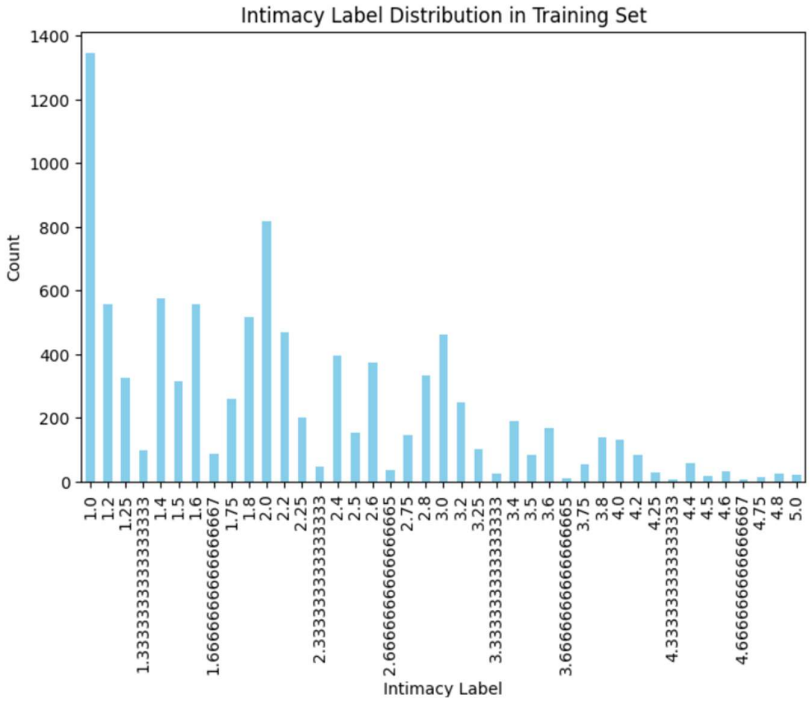
4.5 Future Performance Improvements

Based on observed results and errors, possible next steps include:

- **Data Augmentation:** Synthetic expansion of low-resource and zero-shot language examples to boost robustness.
- **Language/contextual adapters:** Incorporating small, language-specific modules to close the gap in script/cultural-specific affect.
- **Semi-supervised/few-shot training:** Allowing model to adapt with just a handful of annotated samples in zero-shot languages.

5. Results and Interpretation

5.1(new): Intimacy label distribution of training dataset



This bar chart visualizes the distribution of intimacy labels across the training dataset, where each bar represents the frequency (count) of texts assigned to specific intimacy scores. The labels range from 2.33 to 5.0, indicating varying degrees of perceived intimacy. The

distribution highlights imbalances, with 3.33 being the most frequent label, while higher scores (e.g., 4.67–5.0) are significantly rarer.

5.2 Test dataset intimacy score prediction:

Preview of test_df with predictions:

	id	text	predicted_intimacy
0	0	@user 2.314453	😁 الأشقياء الطيبين
1	1	@user Un caffè? 🧿😁😁😁	1.488281
2	2	@user girl!! we support u no matter what 😘	2.953125
3	3	@user 아..	1.430664
4	4	@user @user Pouaaaah jle chantais taleur arrêt...	2.859375

This table showcases a snippet of the test dataset (test_df) augmented with model predictions for intimacy scores. Each row includes a text sample (e.g., user mentions, multilingual phrases) and its predicted_intimacy value, which ranges here from 1.43 to 2.95. The predictions demonstrate how the trained model generalizes to unseen data.

5.3 Detailed Analysis

- English, with its largest training corpus, predicts intimacy with the highest accuracy, showcasing that volume of data remains key.
- Zero-shot results demonstrate meaningful transfer but a notable gap remains, highlighting linguistic and cultural distance challenges.
- Error analysis shows most difficulty in handling sociolinguistic nuances in Korean and Hindi, suggesting domain and script-specific adaptation may enhance results.

6. Discussion and Future Directions

Our findings confirm the potential of large-scale pretrained transformers to perform zero-shot intimacy regression, a complex socially grounded task. While seen-language performance is solid, cross-lingual transfer introduces variability tied to cultural and linguistic phenomena.

Future work should explore:

- Data augmentation strategies for zero-shot languages.
- Incorporating language-specific embeddings or adapters to reduce transfer gaps.
- Semi-supervised or multi-task learning with complementary social signal tasks.

7. References

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., & Zettlemoyer, L. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. [14]

Pei, J., & Jurgens, D. (2020). Quantifying Intimacy in Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [1]

Pei, J., Silva, V., Bos, M., Liu, Y., Neves, L., Jurgens, D., & Barbieri, F. (2023). SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. [2]

Works cited

1. SemEval 2023 Task 9 - Google Sites, accessed August 2, 2025, https://sites.google.com/umich.edu/semeval-2023-tweet-intimacy/home
2. SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis - ACL Anthology, accessed August 2, 2025, https://aclanthology.org/2023.semeval-1.309/