Team #28 - Tweet Trackers
Multilingual Tweet Intimacy Analysis

Sehajdeep Singh; 40294241; sehajdeep.singh@mail.concordia.ca
and
Gurleen Pannu; 40289625; gu_pannu@live.concordia.ca

December 12, 2024

Abstract. Through this project, we aimed to build a model that can accurately predict the intimacy scores (on a scale of 1 to 5) for textual data across five languages: English, French, Spanish, Italian, and Portuguese. Recognizing the importance of contextual elements such as emojis, especially in tweets, our methodology emphasized on retaining these features during pre-processing as well as ensuring language-based pre-processing of the text. We explored a range of regression models, beginning with a traditional Random Forest Regressor that combined Word2Vec and Emoji2Vec embeddings as the baseline model, and extending to advanced transformer-based models such as XLM-R, XLM-T, and mBERT. A significant challenge that we faced was the dataset's imbalance, with lower intimacy scores being far more common than the higher ones. To address this, we implemented an oversampling technique based on SMOTE (which is normally used for classification tasks) using K-Nearest Neighbors (KNN) for our Random Forest Regression model and assigned appropriate class weights for the transformer-based models. We further fine-tuned the transformer-based models by modifying the weights of the last layers, extracting and pooling the contextual embeddings by using mean pooling, and adding a FFNN to predict the intimacy scores. The models were then evaluated using the Pearson correlation coefficient (Pearson-r), as outlined by the Semeval 2023 task organizers. Our results demonstrated that mBERT surpassed the baseline scores defined by the organizers across all the five languages. XLM-T also performed well, exceeding baselines for English, Spanish, and Portuguese, while nearing them for French and Italian. The code and dataset for our project are available for reference. [6]

Contents

1 Goal of the project

The goal of this project was to build a model to predict the intimacy score of any text in 5 languages - English, French, Spanish, Italian, and Portuguese. We emphasized on meticulous data pre-processing to retain essential features like emojis, which are often used in tweets. We trained various models, starting with a traditional Random Forest Regressor using combined Word2Vec and Emoji2Vec embeddings as a baseline, and transformer-based models such as XLM-R, XLM-T, and mBERT. We then evaluated the models on test data using the Pearson correlation coefficient (Pearson-r), a standard metric used by the task organizers. Given that this task was originally part of Semeval 2023 [7], our objective was to surpass the baseline scores established by the task organizers for XLM-T and mBERT models. To add on, there were implementations of this task available online since it's a past task and we referred to one of them [2] to understand the problem, but we have implemented our own unique solution.

2 Methodology

The step by step methodology followed is given below:

- **Dataset Selection and split**: A labeled dataset containing tweets in 5 languages: English, French, Spanish, Italian and Portuguese, is used for the project [5] [3]. The dataset contains 3 columns that include - text (tweet), label (an intimacy score ranging between 1.0 - 5.0 (both inclusive, in regressive fashion)), and language of the tweet. The dataset is split into 80:20 ratio for training and testing purposes.

- **Preprocessing**: Stopwords were removed from the text based on the language. Moreover, blank spaces, irrelevant symbols etc. were also removed. But, after careful analysis of the data, we decided to keep a few symbols (+-=) and digits and removed words like - http(s), @user as they were present in almost all the tweets. Since emojis do contribute to the intimacy of the tweets, they were retained. This could be better visualised through figure 1.

- **Exploratory Data Analysis**: To better understand the given data, histograms were plotted by using the technique of score binning. We grouped intimacy scores into bins for easier comparison across languages as well as to have a better picture of irregularities across the groups. The generated graphs are shown in figure 2.

- **Embedding Creation**: For this task, 2 types of embeddings - emoji embeddings and word embeddings were generated and combined into one.

  1. **Emoji Embeddings**: To create embeddings of emojis, emoji2vec was used [1]. Emoji2Vec is a pre-trained embedding for all Unicode emoji which are learned from their description in the Unicode emoji standard. It maps emoji symbols into the same space as the 300-dimensional word2vec embeddings.
  2. **Word Embeddings**: The word embeddings are created from the pre-processed text. Word2Vec [4] embeddings have the same embedding size as that of Emoji2Vec embeddings. It tokenizes the input text into words and retrieves embeddings of each word from the Word2Vec vocabulary and if a word is not present in the vocabulary, it returns a zero vector.

  After calculating the embeddings of both words and emojis, the two vectors are stacked upon each other (i.e. concatenated). The reason for concatenating the embeddings instead of adding them is to preserve both the information in their original dimensions which helps the model to learn both the embeddings independently, allowing it to understand the relationship between them more efficiently.

- **Random Forest**: Since it's a regression problem, the random forest strategy is considered a good start, because it does not assume a linear relationship between features and the target variables as well as creates multiple decision trees and averages the result which helps to reduce overfitting and handle noise. However, the problem was having an imbalanced dataset, as the random forest can be biased towards the majority class and might not sufficiently explore the patterns in the underrepresented target regions (region 4-5 in figure 2). To handle this imbalance, KNN - based oversampling technique was used.

  KNN-based oversampling technique: This method uses KNN to create synthetic samples by adding a small amount of noise to the existing data points and predicting their corresponding target values in the underrepresented regions. By implementing this, we are ensuring that all the regions have similar representation which helps to balance the dataset eventually making it possible to apply random forest. The random forest algorithm was not able to produce expected results, so the approach was changed to use Bert-based models.

- **Bert - based Models**: Since Bert-based models are pre-trained on large amounts of multilingual texts, it is a wise choice to use them instead of using a vanilla BERT model as these models would have weights and align better to this problem. For this task, 3 different Bert-based models were used:

  1. **mBert**: Multilingual BERT is a variant of the original BERT model that has been trained on text data from 104 different languages.
  2. **XLM-R** : XLM-RoBERTa (XLM-R) is a powerful multilingual language model designed for cross-lingual understanding and transfer learning across multiple languages. It is trained on a massive dataset of 2.5TB of filtered CommonCrawl data across 100 languages.
  3. **XLM-T**: XLM-T is built on the XLM-RoBERTa (XLM-R) architecture, pre-trained on a large corpus of Twitter data in multiple languages (so it is expected to perform even better for this task).

  Approach: The approach that was adopted for using these models was to use them as it is as well as by fine-tuning them. Fine-tuning of the models was done incrementally. We started by applying a feed-forward neural network (FFNN) ahead of the transformer. This FFNN maps the representations generated by the tranformer to a final output that is predicting a score between 1-5. After that, a mean pooling layer was added between a transformer and FFNN which helps by averaging the hidden states of all non-padding tokens, providing a more holistic representation of the entire input sequence, instead of relying on a single token's embedding. At last, the transformer's last layer was fine-tuned wherein the weights of all the layers was frozen, except the last layer which gets updated during the training process.

3 Evaluation

3.1 Results

We evaluated all our models using the Pearson correlation coefficient (aka the Pearson-r Score), which in our project measured the relation between the predicted and true intimacy scores (the closer the score to 1, the better the model performance), as this was used as the standard metric for evaluation by the task organizers. The results of the Transformer-based models (XLM-R, XLM-T and mBERT) before fine-tuning was observed as seen in Table 1. We observed the results as shown in Table 2 for all our models including the Random Forest model and the fine-tuned Transformer-based models.

3.2 Analysis

Based on our results we could observe that XLM-T proved to be the best performing model amongst all the models tested, owing to its pre-training on multilingual Twitter data, which makes it adept at understanding social media specific nuances, and informal contexts. It achieved the highest average Pearson-r score of 0.6897, and significantly outperformed the mBERT (0.5732) and XLM-R (0.6096) models. It is also notable to observe how far behind the traditional Random Forest Regressor (0.2153) lies even after our enhancements with the concatenated emoji2vec and word2vec embeddings and custom oversampling technique, highlighting the limitations of traditional models in capturing the nuanced and multilingual nature of the data and underscoring the transformative impact of fine-tuned transformer-based models.

Our XLM-T Model's performance exceeded the baseline scores for English, Spanish and Portuguese demonstrating its strong proficiency in these languages. However, its performance for Italian and French fell just slightly below the baseline. This shows that while XLM-T is effective in most cases, there is still room for improvement in handling a few languages. Upon comparison between XLM-T and XLM-R's performance, we can see the importance of domain specific pre-trained transformer-based models. Another key observation would be the impact of fine-tuning of the pre-trained Transformer-based models, as we can see considerable improvement in the performance of the models after they were further fine-tuned.

4 Limitations

Several limitations in our project could be addressed in future work:

- The lack of language detection tools, such as langdetect or langid, limits the ability to handle unknown or mixed-language inputs effectively. Our model's pre-processing pipeline would struggle to adapt in such a scenario.

- The absence of language-specific tokenizers, like spaCy or Stanza during pre-processing restricts the model's ability to handle any unique grammar and syntactical structures of the individual languages, affecting the pre-processing quality.
- Our model currently cannot adequately address the different regional dialects, slang, and out-of-vocabulary (OOV) words, which in the context of tweets could have added to a better understanding of the text.
- The model currently lacks cross-lingual transfer learning, and thus cannot be generalized across unknown languages.
- This problem involves evolving language styles, requiring periodic model retraining as new datasets become available. While our current model does not address this, future work could explore continual learning techniques like selective replay or latent replay to adapt to these changes.

5 Difference with your original proposal

- The proposal initially included 6 languages. However, Chinese was excluded due to challenges with stopword removal since it is not supported by NLTK. and other alternatives like Jieba require manual handling of stopwords, which wasn't feasible due to our lack of familiarity with the Chinese language. Hence, we dropped Chinese from the dataset.
- ChatGPT's output is not deterministic and depends on specific prompts and configurations, making the quality of generated data for augmentation unreliable. Therefore, this approach was not pursued, but can be explored in the future.

6 Conclusions

Through our project, we were able to successfully predict the intimacy scores for text across 5 languages, using different regression models, adequate preprocessing, oversampling techniques, and fine-tuning the transformer-based models. We successfully retained contextual features and address class imbalance challenge by concatenating both embeddings and using oversampling with KNN before sending it to our Random Forest model. The performance was far from perfect for random forest and XLM-T proved to be the most effective model, surpassing the baseline scores for English, Spanish, and Portuguese, while also delivering very close results for Italian and French. This result was expected since XLM-T is trained predominantly on tweets, and it thus understands the nuances of tweets better than the other multilingual transformer-based models. The further fine-tuning of our transformer-based models added significantly to their performance across all the languages. The project addresses the Semeval 2023 Task substantially, however, there remain many enhancements that could further improve the performance.

7 References

References

[1] Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., and Riedel, S. emoji2vec: Learning emoji representations from their description, 2016.

[2] HULAT Intimacy Team. Hulat intimacy code. https://github.com/isegura/hulat$_{i}ntimacy.$

[3] Jiaxin Pei and Vítor Silva and Maarten Bos and Yozen Liu and Leonardo Neves and David Jurgens and Francesco Barbieri. Semeval 2023 task 9: Multilingual tweet intimacy analysis. https://arxiv.org/pdf/2210.01108v2(site visited on December 12, 2024).

[4] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space, 2013.

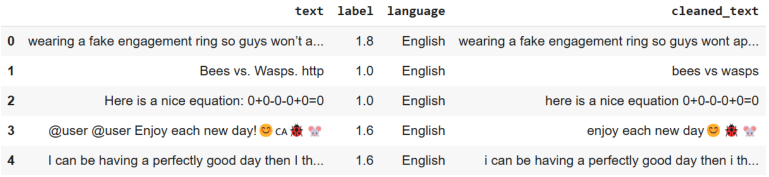
[5] Pei et al. Mint (Multilingual Intimacy Analysis). Tech. rep. https://paperswithcode.com/dataset/mint(site visited on December 12, 2024).

[6] Sehajdeep Singh, Gurleen Pannu. Github repo containing code and dataset. https://github.com/sehaj-deep/Multilingual$_{I}ntimacy_{A}nalysis.$

[7] University of Michigan and Snap Inc. Semeval 2023 task 9: Multilingual tweet intimacy analysis. https://sites.google.com/umich.edu/semeval-2023-tweet-intimacy/home(site visited on December 12, 2024).

A Appendix

A.1 Additional Results and Visualizations



	text	label	language	cleaned_text
0	wearing a fake engagement ring so guys won’t a...	1.8	English	wearing a fake engagement ring so guys wont ap...
1	Bees vs. Wasps. http	1.0	English	bees vs wasps
2	Here is a nice equation: 0+0-0-0+0=0	1.0	English	here is a nice equation 0+0-0-0+0=0
3	@user @user Enjoy each new day! 😊 CA 🐞 🐭	1.6	English	enjoy each new day 😊 🐞 🐭
4	I can be having a perfectly good day then I th...	1.6	English	i can be having a perfectly good day then i th...

Figure 1: Dataset before and after cleaning

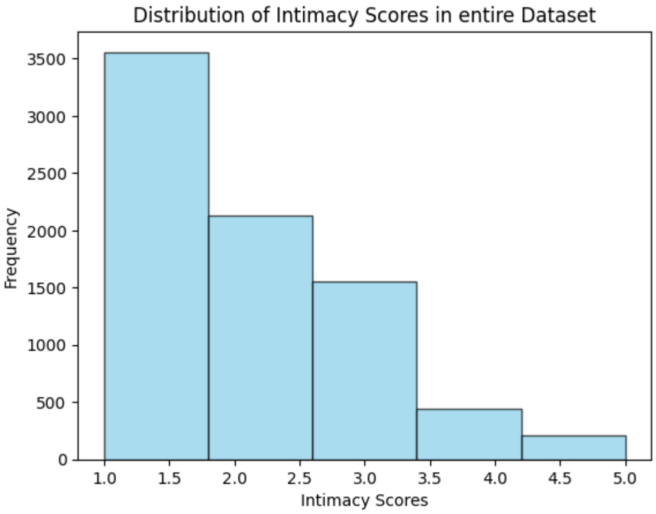


Figure 2: Data imbalance between different intimacy scores

Model	English	Spanish	Portuguese	Italian	French	Average
mBERT	0.6124	0.6289	0.4711	0.5739	0.4812	0.5536
XLM-R	0.6242	0.6156	0.2914	0.5486	0.3693	0.4898
XLM-T	0.6954	0.7116	0.5868	0.6350	0.5942	0.6446

Table 1: Pearson’s r-score for the Transformer models before fine tuning

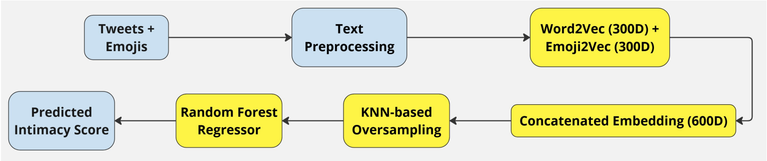


Figure 3: Methodology of random Forest

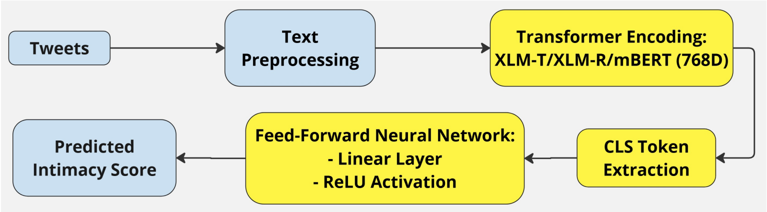


Figure 4: Methodology of transformer

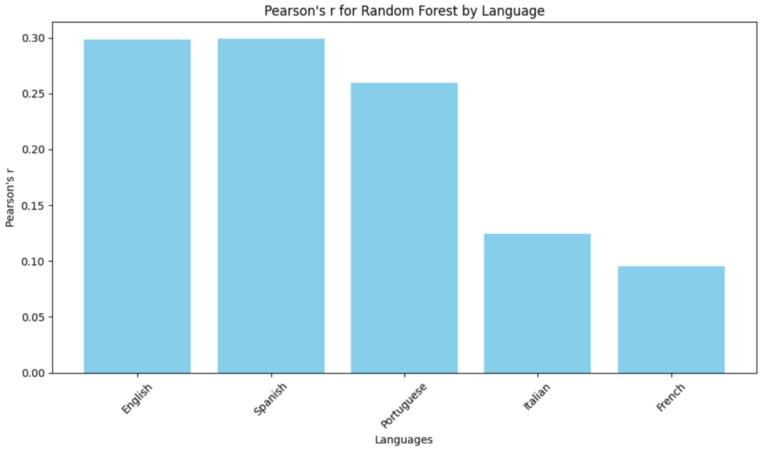


Figure 5: Random Forest Results

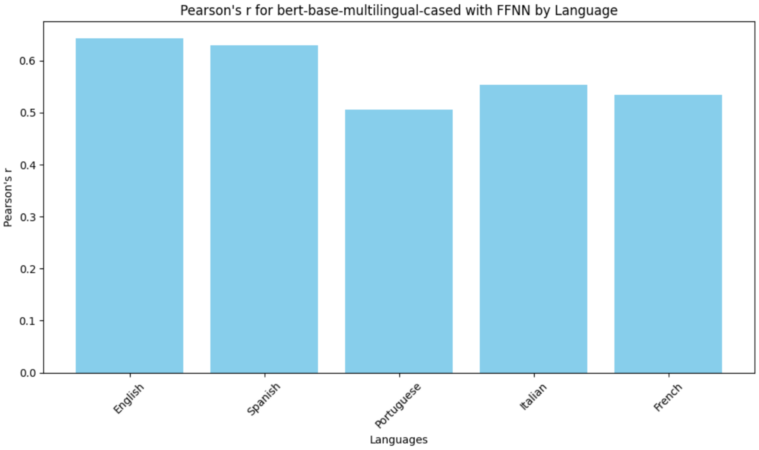


Figure 6: mBERT

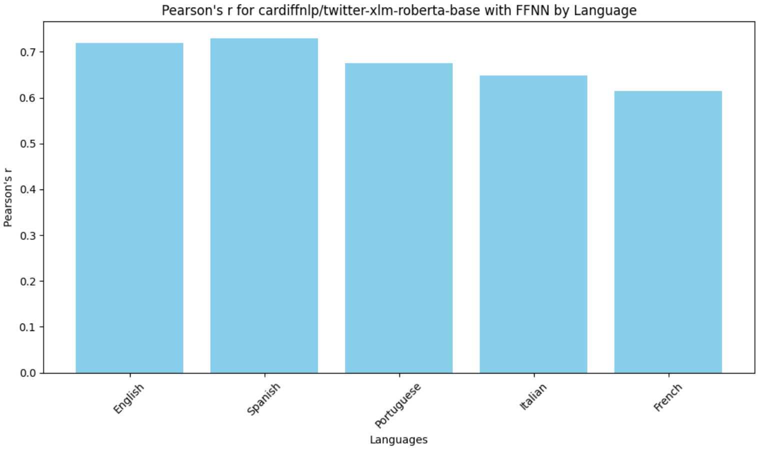


Figure 7: XLM-R

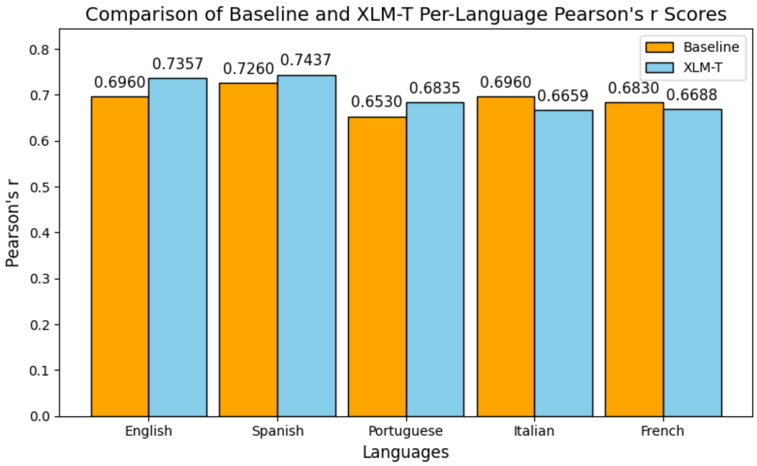


Figure 8: Best Score Achieved vs Baseline

Model	**English**	**Spanish**	**Portuguese**	**Italian**	**French**	**Average**
Random Forest	0.2981	0.2991	0.2593	0.1244	0.1675	0.2153
mBERT	0.6431	0.6298	0.5061	0.5739	0.5344	0.5732
XLM-R	0.6607	0.6469	0.6124	0.5928	0.5354	0.6096
XLM-T	0.7357	0.7437	0.6835	0.6659	0.6688	0.6897

Table 2: Per-Language Pearson's r-score and Average Scores for All Models