I2C-Huelva at SemEval-2023 Task 9: Analysis of Intimacy in Multilingual Tweets Using Resampling Methods and Transformers

Abel Pichardo Estévez, Jacinto Mata Vázquez, Victoria Pachón Álvarez, Nordin El Balima Cordero
Escuela Técnica Superior de Ingeniería. Universidad de Huelva (Spain)
abel.pichardo107@alu.uhu.es, mata@uhu.es, vpachon@uhu.es, nordin.elbalima531@alu.uhu.es

Abstract

Nowadays, intimacy is a fundamental aspect of how we relate to other people in social settings. The most frequent way in which we can determine a high level of intimacy is in the use of certain emoticons, curse words, verbs, etc. This paper presents the approach developed to solve *SemEval 2023 task 9: Multiligual Tweet Intimacy Analysis*. To address the task, a transfer-learning approach was conducted by fine-tuning various pre-trained language models. Since the dataset supplied by the organizer was highly imbalanced, our main strategy to obtain high prediction values was the implementation of round-trip translation for oversampling and a random approach for undersampling on the training set. Our final submission achieved an overall Pearson's r of 0.497.

1 Introduction

Intimacy is defined as the set of thoughts and feelings that human beings keep private. There are many ways of expressing a thought and not all are equally intimate, so it is important to know what determines the degree of intimacy of a particular Tweet. To detect this type of language, SemEval 2023 proposed the *task 9: Multiligual Tweet Intimacy Analysis* (Pei et al., 2023).

The aim of the proposed task is to quantify the intimacy expressed in a set of tweets written in six languages: Chinese, English, Spanish, Portuguese, French and Italian.

We addressed the problem using regression (Specht et al., 1991) and neural networks (Zhou et al., 2015). We utilized different pre-trained models such as RoBERTa (Liu et al., 2019) or BERT (Devlin et al., 2018), available on the *Hugging Face* platform. Given the strong imbalance in the data, we have focused on applying different oversampling (Liu et al., 2007) and undersampling (Bach et al., 2019) strategies.

The main contributions addressed in this work are as follows:

- Analysis of sampling methods to mitigate bias in data.
- Benchmarking of different pre-trained multilingual and monolingual models.

The rest of this paper is organized as follows: in Section 2 we explain the dataset, the meaning of the labels and data distribution. Therefore, we refer to other studies that contribute to our approach. Section 3 discusses the approaches applied to achieve better model performance. In Section 4, we explain the libraries used. In Section 5, the results obtained in test phase are shown. Finally, conclusions and future works are described in Section 6.

2 Background

The training dataset consists of 9491 tweets written in six languages. The distribution of each of the languages is balanced. Figure 1 shows an example of a dataset row.

Text	Label	Language
@user was that YOU dancing in ur recent story???!!! 😯	4.0	English

Figure 1. Dataset example

The main fields are:

- Text. Represents the tweet to be analyzed.
- Label. Determines the level of intimacy assigned to the text with a range [1-5], where 1 represents the lowest level and 5 represents the highest level of intimacy.
- Language. It's the language of the text.

In Figure 2, the dataset's distribution is depicted for each language, along with the distribution of intimacy levels across languages. Four intervals have been considered to study the distribution and focus on the problem. We can see how the number

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 758–762

of tweets between languages is balanced. However, at lower levels of intimacy we have more examples than at the higher levels. This distribution is repeated across all the languages of the dataset. Therefore, we decided to apply the same techniques over the whole dataset without distinction between languages.

Language	Total	1-2	2-3	3-4	4-5
English	1587	945	404	182	56
Spanish	1592	703	496	307	86
French	1588	774	524	223	67
Italian	1532	846	458	190	38
Portuguese	1596	715	536	279	66
Chinese	1596	647	552	294	103

Figure 2. Distribution of languages according to labels

There are not many published studies in the field of intimacy analysis. However, there are many studies related to the classification of levels of aggression and violence in texts and sentiment analysis that can help to better understand our problem and as a starting point to address it. In (Arias and Fabian, 2022) a study on the automatic detection of levels of psychological violence against women in written virtual expressions was conducted. (Zhao et al., 2017) proposed a framework called Weakly-supervised Deep Embedding (WDE), which employs review ratings to train a sentiment classifier.

3 System Overview

This section describes the development of our approach and outlines the steps taken to achieve the results.

Since only one training dataset was provided, for the experimental environment we decided to split the original dataset into a training/validation dataset (80%) and a test dataset (20%).

3.1 Text pre-processing

Several techniques were employed to simplify the text and remove any potential noise. The main techniques that were applied include:

- Converting all characters to lower case
- Removing url and links
- Removing "@user"

For this study, we considered that both hashtags and emojis were relevant to the analysis of intimacy, so they were not removed from the tweets. Regarding emojis, we consider their usage to be quite prevalent in highly intimate texts and believe that they should be taken into consideration. It was decided to expand the vocabulary of the Transformer models so that they would recognize emojis as one more token without treating them as unknown tokens. The result of a tweet before and after applying the text pre-processing can be seen in Figure 3.

Original tweet	Processed tweet
@user Furthermore, harassment is ILLEGAL in any form!	furthermore, harassment is illegal in any form!

Figure 3. Text-preprocessing example

3.2 Data distribution

As previously stated, one of the primary concerns with the training dataset is the significant imbalance in the data. Most of examples across all languages had an intimacy level below 2.5, and as the level of intimacy increased, the number of examples reduced significantly. Different versions of resampling were implemented on the training dataset to improve the value of the Pearson's r on our test dataset.

The same pre-trained model was fine-tuned to test the performance of the different resampling approaches. Specifically, the multilingual *xlm-roberta-base* [1] (Conneau et al., 2019) was the selected model.

3.2.1 Oversampling

Oversampling consists of modifying the distribution of the data by increasing the number of cases of the minority class.

For this task, the round-trip translation (Xie et al., 2019) was used to perform oversampling. In this way, it is possible to have tweets with the same meaning and level of intimacy but with a different structure. An example of paraphrase is shown in Figure 4.

Original tweet	who should i draw on my live to entertain the horny mfs i know are gonna show up
Paraphrase tweet	who should draw in my life to entertain the horny mfs that I know will appear

Figure 4. Oversampling example

1 https://huggingface.co/xlm-roberta-base

The proposed approaches for implementing oversampling were:

- O1. Oversampling on tweets with label values > 2.5.
- O2. Oversampling on tweets with label values > 3.
- O3. Oversampling on tweets with label values > 4.
- O4. Oversampling with different ranges. Tweets whose labels belongs to the value range [1 - 2.5] were not oversampled. Tweets labeled in the value range (2.5 - 4] were duplicated. Tweets whose labels belong to the value range (4 - 5] were repeated three times.

For approaches O1, O2, and O3, tweets meeting the specified criterion were duplicated. Table 1 shows the results achieved for all oversampling techniques using the pre-trained *xlm-roberta-base* model.

Technique	Pearson's r
O1	0.77
O2	0.74
O3	0.72
O4	0.79

Table 1. Results obtained with oversampling techniques.

3.2.2 Undersampling

Undersampling is used to reduce the number of instances belonging to the majority class. Random undersampling (RUS) was proposed in (Prusa et al. 2015) to randomly reduce the examples of the majority class. To apply some of the undersampling techniques, the balanced datasets with the oversampling techniques were used. The approaches for undersampling were:

- U1. Undersampling on tweets with label values < 2.5 from the original dataset.
- U2. Undersampling on tweets with label values < 2.5 on O1 dataset
- U3: Undersampling on tweets with label values < 2.5 on O1 dataset.
- U4. Undersampling on tweets with label values < 2.5 on O4 dataset.

The ratio selected to remove the examples was 50% for U1, U2 and U4 approaches, and 70% for U3 approach. Table 2 shows a comparison of the results achieved using the different undersampling approaches with the pre-trained *xlm-roberta-base* model.

Technique	Pearson's r
U1	0.59
U2	0.78
U3	0.71
U4	0.77

Table 2. Results obtained with oversampling techniques.

3.3 A model for each language

In the previous section, the experimentation and results achieved using different resampling approaches have been described. The experiments were conducted using a single multilingual model to assess the impact of the balancing techniques on the original dataset. Multilingual models generally perform satisfactorily for all languages. However, monolingual models tend to perform better in their own languages because they have been specifically trained in those languages. For this experiment, we decided to test the performance with the oversampling O1 dataset for each of the languages using a pre-trained monolingual model.

The pre-trained monolingual models used for this experiment were:

- English: *bert_base_uncased*[2] (Lee et al. 2018).
- Spanish: *mrm8488/bert-spanish-cased-finetuned-ner*[3]
- Chinese: *bert-base-chinese*[4].
- Italian: *dbmdz/bert-base-italian-xxl-uncased*[5]
- Portuguese: *neuralmind/bert-large-portuguese-cased*[6] (Souza et al. 2020).
- French: *Jean-Baptiste/camembert-ner*[7]

Language	Pearson's r
English	0.82
Spanish	0.83
Chinese	0.70
Italian	0.70
Portuguese	0.67
French	0.67

Table 3. Results obtained using a monolingual model for each language.

2 https://huggingface.co/bert-base-uncased
3 https://huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner
4 https://huggingface.co/bert-base-chinese
5 https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased
6 https://huggingface.co/neuralmind/bert-large-portuguese-cased
7 https://huggingface.co/Jean-Baptiste/camembert-ner

Table 3 shows the results obtained for each of the languages using the O1 training dataset.

3.4 Multilingual and monolingual

Table 3 shows that the monolingual models performed better in English and Spanish. However, the monolingual models performed worse compared to the multilingual model for the other languages.

For this reason, a multilingual model with the worst performing languages (Chinese, French, Portuguese and Italian) was fine-tuned. The monolingual models which outperformed the multilingual model (English and Spanish) were used. The overall outcome was nearly identical to that of the multilingual model, exhibiting a meager increase of merely 0.04 in Pearson's r compared to the *xlm-roberta-base* model.

4 Experimental setup

Some of the libraries used for the experimentation were: "NumPy" (Harris et al., 2020), a library for advanced natural language processing; "NLTK" (Loper and Bird, 2002), a tool-kit to work with human language information; "Keras", a neural-network library; "scikit-learn", which contains simple and efficient tools for predictive data analysis; and "Pandas" (McKinney, 2010), used for manipulating data. The metric used is Pearson's r from the *scipy* library as requested by the competition.

For fine-tuning the models, the hyperparameters used were: batch size of 16, learning rate of 5e-5, max length of 128 and weight decay of 0.01, 10 epochs using early stopping with a patience of 3.

5 Results

As can be seen in the previous sections, oversampling O4 was the best performing and most computationally efficient technique. Therefore, we decided to try out other pre-trained multilingual models: *xlm-roberta-large*[8] (Conneau et al., 2019), *cardiffnlp/twitter-xlm-roberta-base-sentiment* [9], *bert-base-multilingual-uncased* [10] (Devlin et al., 2018) and *Twitter/twhin-bert-base*[11] (Zhang et al., 2022). Table 4 shows the results achieved by the different multilingual models.

Model	Pearson's r
xlm-roberta-base	0.793
xlm-roberta-large	0.804
cardiffnlp/twitter-xlm-roberta-base-sentiment	0.810
bert-base-multilingual-uncased	0.795
Twitter/twhin-bert-base	**0.827**

Table 4. Pearson's r achieved by pre-trained models.

In the test phase, four new languages were included: Arabic, Korean, Hindi and Dutch. To predict the level of intimacy of these tweets with our models, they were translated into English.

Finally, the predictions obtained by *Twitter/twhin-bert-base* model and the O4 technique were submitted. The Pearson's r achieved for the official test dataset was 0.497. We were ranked 36th in the overall intimacy score. In Table 5, the results and overall position obtained for each language are shown.

Language	Score	Ranking
English	0.623	36
Spanish	0.673	35
Portuguese	0.619	37
Italian	0.631	35
French	0.578	30
Chinese	0.659	36
Hindi	0.206	36
Dutch	0.450	36
Korean	0.253	28
Arabic	0.405	39
Seen	0.643	36
Unseen	0.305	38
Overall	0.497	36

Table 5. Final results.

6 Conclusions

In this paper, the approach of the I2C Group to address *Task 9: Multiligual Tweet Intimacy Analysis* has been described. Fine-tuning of several pre-trained monolingual and multilingual language models was performed and different resampling techniques were applied.

The best results in the development phase were obtained by using a range oversampling (O4) technique. In this phase, the Pearson's r achieved was 0.827 but, in the evaluation phase, the value of Pearson's r was only 0.497. We think that the reason for the difference between the two values

8 https://huggingface.co/xlm-roberta-large

9 https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

10 https://huggingface.co/bert-base-multilingual-uncased

11 https://huggingface.co/Twitter/twhin-bert-base

could be due to the incorporation of new languages into the test dataset.

For future works, we would like to add additional features to the tweets such as the number and type of emoticons, the use of certain adjectives, etc. We think that the intimacy language vocabulary is very specific, and this could help the prediction models.

Acknowledgments

This paper is part of the I+D+i Project titled "Conspiracy Theories and hate speech online: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people [NON-CONSPIRA-HATE!]", PID2021-123983OB-I00, funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe".

References

Bach, M., Werner, A., & Palt, M. (2019). The proposal of undersampling method for learning from imbalanced datasets. *Procedia Computer Science*, *159*, 125-134.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. *Advances in neural information processing systems*, *32*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DU, Yongping, et al. A novel capsule based hybrid neural network for sentiment classification. *IEEE Access*, 2019, vol. 7, p. 39321-39328.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362.

Pei, J., Silva, V., Bos, M., Liu, Y., Neves, L., Jurgens, D., & Barbieri, F. (2023). SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis. *arXiv preprint arXiv:2210.01108*.

Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., & Napolitano, A. (2015). Using random undersampling to alleviate class imbalance on tweet sentiment data. In *2015 IEEE international conference on information reuse and integration* (pp. 197-202). IEEE.

Lee, J. D. M. C. K., & Toutanova, K. (2018). Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Liu, A., Ghosh, J., & Martin, C. (2007). Generative Oversampling for Mining Imbalanced Datasets. In *DMIN* (pp. 66-72).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, No. 1, pp. 51-56).

Xie, Q., Dai, Z., Hovy, E., Luong, M. T., & Le, Q. V. (2020). Unsupervised data augmentation for consistency training. arXiv. *Learning*, *1*.

Spetch, D. F., et al. (1991). A general regression neural network. *IEEE transactions on neural networks*, vol. 2, no 6, p. 568-576.

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9* (pp. 403-417). Springer International Publishing.

Yallico Arias, T., & Fabian, J. (2022). Automatic detection of levels of intimate partner violence against women with natural language processing using machine learning and deep learning techniques. In *Information Management and Big Data: 8th Annual International Conference, SIMBig* (pp. 189-205).

Zhang, X., Malkov, Y., Florez, O., Park, S., McWilliams, B., Han, J., & El-Kishky, A. (2022). TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations. *arXiv preprint arXiv:2209.07562*.

Zhao, W., Guan, Z., Chen, L., He, X., Cai, D., Wang, B., & Wang, Q. (2017). Weakly-supervised deep embedding for product review sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *30*(1), 185-197.

Zhou, C., Sun, C., Liu, Z., & Lau, F. (2015). A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630*.