jelenasteam at SemEval-2023 Task 9: Quantification of Intimacy in Multilingual Tweets using Machine Learning Algorithms: A Comparative Study on the MINT Dataset

Jelena Lazić
PhD student
School of Electrical Engineering
University of Belgrade
lazic.jelenaa@gmail.com

Sanja Vujnović
Assistant Professor
School of Electrical Engineering
University of Belgrade
svujnovic@etf.bg.ac.rs

Abstract

Intimacy is one of the fundamental aspects of our social life. It relates to intimate interactions with others, often including verbal self-disclosure. In this paper, we researched machine learning algorithms for quantification of the intimacy in the tweets. A new multilingual textual intimacy dataset named MINT was used. It contains tweets in 10 languages, including English, Spanish, French, Portuguese, Italian, and Chinese in both training and test datasets, and Dutch, Korean, Hindi, and Arabic in test data only. In the first experiment, linear regression models combine with the features and word embedding, and XLM-T deep learning model were compared. In the second experiment, cross-lingual learning between languages was tested. In the third experiments, data was clustered using K-means. The results indicate that XLM-T pre-trained embedding might be a good choice for an unsupervised learning algorithm for intimacy detection.

1 Introduction

Natural language processing (NLP) models have access to more information than any human speaker during their life. Still, it would be hard for them to reach human-level competence and performance (Hovy and Yang, 2021). NLP community needs to address social factors to get closer to the human-like language understanding. Performances of the NLP models for detection of the social-components in the language like patronizing and condescending language detection, irony and sarcasm understanding, persuasion, humor and emphasis detection, and many others, should be improved. The aim of our work was analysis of intimacy in tweets, as intimacy is one of the fundamental social aspects of language.

In this paper, a new multilingual textual intimacy dataset named MINT annotated by (Pei et al., 2023) was used. The dataset is released along with the SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis, co-organized by University of Michigan and Snap Inc. It contains a total of 13,384 tweets in 10 languages, including English, Spanish, French, Portuguese, Italian, and Chinese in both training and test data, and Dutch, Korean, Hindi, and Arabic in test data only.

During work on this task, a three experiments were conducted:

- **Models Comparision**
  Recently, models based on neural networks have become increasingly popular. While these models achieve very good performance in practice, they tend to be relatively slow both at training and test time, limiting their use on very large datasets (Joulin et al., 2016). Meanwhile, traditional models are often considered as strong baselines. In the first experiment comparative study of feature-based, embedding-based and deep learning (DL) models was conducted.

- **Cross-lingual evaluation**
  Zero-shot learning in NLP is still challenging. In their work (Mozafari et al., 2022) show that only a few shots could lead to learning improvement for low-source languages. In the second experiment cross-lingual learning was researched, training models in one language and evaluating it on another.

- **Clustering**
  In the third experiment, clustering was researched. The unsupervised learning method was used to discover hidden structure in a data. The assumption was that embedding with more information would lead to more separable clusters. Therefore, clustering was used as an initial feature inspection. The aim was to research weather tweets will group by language, intimacy score, or type of intimacy it relates to (self-disclosure, love-partners intimacy, etc.).

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 638–643

The results indicate that NLP models could predict intimacy score in tweets. The DL model gained the best results, while embedding-based model gained slightly worse results, but with much lower complexity. It is possible to transfer knowledge from one to another language, and with appropriate training-test pair, it is possible to have good predictions even for the zero-shoot languages. Clustering of data using embedding is probably based on the intimacy score rather than on language.

This paper is organized as follows: In Section 2, an overview of related literature was given. In Section 3, explanation of the models used for experiments was given, as well as key details about experimental setup. Section 4, presents the results of the experiments. In the end, in Section 5, there is a conclusion followed by the references.

2 Background

Intimacy is a fundamental aspect of how we relate to others in social settings (Pei and Jurgens, 2020). We express it by offering emotional support, reassurance of love and care, or a deeper understanding of self and others. Intimate contact has a high impact on personal growth, feeling secure, achievement of our potential in friendships and romantic relationships (Prager et al., 2013). There are multiple conceptions of intimacy. Some define intimacy as a kind of intimate interactions, which are most frequently associated with verbal self-disclosure.

In their work (Umar et al., 2019) proposed a method for detection and analysis of self-disclosure in online news commentaries. Their model used detection of sentence form, active or passive, based on the subject-verb-object relation in the sentence. They tested model over different categories based on the type of personal info disclosed: age, race, sexuality, location or affiliation, relationship status, money, interests, opinions and feelings. Their model was used by (Blose et al., 2020) in a study of self-disclosure during a coronavirus pandemic.

In their work on automatic classification of sexism in social networks, (Rodríguez-Sánchez et al., 2020) compared BERT model with BiLSTM and traditional classifiers based on combined features with word embedding and tweet and user meta-data. The results show that a combination of additional information about the tweet can increase classification accuracy. Authors pointed out that some tweets could sound lovely or friendly, but also could be considered sexist.

The authors (Shen and Rudzicz, 2017) investigated binary anxiety detection on a small English Reddit dataset. They used posts without meta-data and preform classification with word2vec, doc2vec, the Latent Dirichlet Allocation (LDA), N-grams combined with logistic regression (LR), Support Vector Machine (SVM) and artificial neural networks (ANN), optimized over features and classifier. The results indicated the differences in the lexics of these two classes, in the most frequent unigrams, bigrams, and trigrams.

In their research (Pei and Jurgens, 2020) conducted several studies of expressions of intimacy in language, and showed, using a fine-tuned RoBERTa, that the intimacy of language is not only a personal choice, where people may use different linguistic strategies for the expressions of intimacy, but reflects constraints of social norms, including gender and social distance. They reproofed stranger on a train phenomenon.

The data set used in this research was multilingual, which rises questions about cultural differences expressed through different language groups on Twitter. With the language chosen, a social identity is created. Speakers may make use of specific words or stylish elements to represent themselves in a certain way (Nguyen et al., 2015). People from different countries have their own ways of redrawing the boundaries of what may be said, written and done within a given discourse community (Kramsch, 1998). Some expressions of intimacy that are acceptable in one culture, could be rude and impolite or even absurd in another one.

The authors (Chang et al., 2022) researched how multilingual language models maintain a shared multilingual representation space, while encoding language sensitive information in each language. They used XLM-R model and compared space between 88 languages. After performing dimension reduction on embedding from different layers in the network, the authors concluded that there are a language-sensitive axis, and a language-neutral axis. The limitation of their work was that the most of languages were European. The multilingual pre-trained BERT embedding and LDA topic model were used by (Xie et al., 2020), for analysis of topic evolution in monolingual and multilingual topic similarity settings. For each topic, they multiply its LDA probability value by the averaged tensor similarity of BERT embedding to explore the evolution of the topic in scientific publications.

The explosion of digital trace data with the availability of computational methods that are faster, cheaper and easier to use, has ushered in a new scientific approach to measuring culture, bringing the new challenges to overcome. Biased NLP algorithms, caused by inappropriate data collection, can cause instant negative effects on society by discriminating against certain social groups and shaping the biased associations of individuals through the media they are exposed to. In the end, it is important to mention that social media languages have their own specifications, like more colloquial expressions and more linguistic variation, such as the use of slang and dialects, platform-specific factors such as misspellings, vulgarisms, emoji and multi-modality.

3 System Overview

The task of intimacy analysis was conducted through three experiments, where each one concentrates on specific aspect of this problem. In the first experiment, comparative study of intimacy detection with different models was conducted. We have noticed that in the last few years on SemEvals, most participants researched deep learning models, which was expected considering they are state-of-the-art in many NLP tasks. However, we think that traditional models could be useful, not just because of their low complexity, but also because of their non-black-box nature.

Two linear regression (LR) models with L2 regularization separately trained on either bag of words (BoW) features and term frequency-inverse document frequency (tf-idf) features were included. Features were extracted using CountVectorizer and TfidfVectorizer from the sklearn library (Pedregosa et al., 2011), with unigrams, bigrams and trigrams, and jieba tokenizer. Regularization parameter was determined using cross-validation. Tweets from the training dataset, for each seen language, were randomly split into 5 folds, stratified by the intimacy score. These folds were used for training and validation. Based on the cross-validation results, the regularization parameter was determined. The model trained with the optimal regularization parameter was tested on train and test data.

Apart from features, word embedding-based models were compared. Embedding were generated using word2vec from gensim library (Rehurek and Sojka, 2011) and the pre-trained XLM-T model introduced by (Barbieri et al., 2021). The word2vec based model was optimized over the vector size, starting from 100 and increasing the size for 50, until model performance stopped improving. The optimal size was 200. For XLM-T embedding, a pre-defined size of 768 was used. Both embedding were used with the LR models, with parameters set in the same way as for a feature-based model.

In the end, XLM-T model was used, as it represents one of the DL models that achieved the best performance in the task proposal paper (Pei et al., 2023). The data was split randomly stratified to language, into a train and validation set, with a ratio of $80 : 20$. Batch size was 64, and the model was optimized over 0.01, 0.001 and 0.0001 learning rate. Better results were gained using 0.001 and 0.0001 learning rates and 0.001 was chosen. The model was trained with 15 epochs. We evaluated the model after each 100 steps, and loaded the best model in the end.

In the second experiment, cross-lingual analysis was conducted. Models were trained on data in one language from train dataset and tested on all languages in test dataset. XLM-T model fine-tuned in the same way as in the first experiment was used. The aim was to research what is the best choice for zero-shot for each language, and to compare similarities for intimacy expressions of languages got through cross-lingual training, with the known cultural similarities.

In the third experiment we clustered the data, in an attempt to see whether the tweets would group by language or intimacy score, thus verifying the validity of the extracted features. Based on the results of the first experiment, we decided to use XLM-T embedding as a feature for clustering. K-means was implemented using sklearn library (Pedregosa et al., 2011). During experiment, we were changing the number of clusters and observing the statistics of clusters. The inspection involved examining the number of unique languages in each cluster and the average intimacy score for a fixed number of clusters. All clustering were performed with the random initialization. After clustering training data, the test data clusters were predicted.

4 Results

4.1 The First Experiment - Models comparision

The results of the models comparison are shown in Table 1. The best results were gained with a DL model, for both seen and unseen languages.

This was expected considering the large number of parameters of this model as well as the large pre-training corpus. BoW gave worse performance results than Tf-Idf in overall comparison, but this improvement with Tf-Idf was gained with seen languages, while the BoW model gave much better results for unseen languages. During the training we noticed that if we let the BoW model to overfit, it would gain results similar to Tf-Idf. This indicates that penalizing words that often appear in documents and reduction of the count of these words, could lead to worse performances of the model for quantifying intimacy. This agrees with the existing finding, (Pei and Jurgens, 2020) shows that individuals use some words more frequently in intimate questions, for example people use some swearing in intimate settings, likely as a way of indicating closeness and hedge when asking intimate questions to decrease risk.

Even though embedding-based models take into account the placement of words in a document, unlike features based-models, the word2vec based model performed the worst. We assume that this is happening because the train dataset is still very small, and the word2vec model was not able to capture word relationships in the embedding space. The model based on XLM-T pre-trained embedding and LR, was almost as good as the fine-tuned XLM-T model. The time of the cross-validation and training of the LR + XLM-T model was less than a minute, while the training time of XLM-T model was around 40 minutes, using Nvidia K80/T4 GPU and Tensorflow framework. Results indicate that the power of the XLM-T model is probably mostly based on a very large training corpus, rather than on the large number of learning parameters.

Table 1: Pearson's coefficient of correlation for predictions of models trained on all provided train data. Overall column is measured on all test data, seen column is measured on test samples in languages present in train data, unseen column is measured on test samples in languages not included in training data.

Model	Overall	Seen	Unseen
LR + BoW	0.39167	0.48347	0.31170
LR + Tf-Idf	0.40211	0.53094	0.25805
LR + Word2vec	0.26256	0.30460	0.25919
LR + XLM-T	0.52388	0.64786	0.40373
XLM- T	0.58047	0.70567	0.42646

4.2 The Second Experiment - Cross-lingual evaluation

Results of the second experiment are in Table 2. An interesting observation is for Spanish→Italian case in which zero-shot model outperformed its mono-lingual counterpart. In the majority of languages, the model trained on all training data outperformed models trained on mono-lingual datasets. This was expected since the complete dataset has six times more samples than any mono-lingual dataset. However, despite the six-fold increase in the size of the training data, the Pearson's coefficient only increased up to 0.08. Also, English→Dutch, Portuguese→Portuguese and Spanish→Korean outperformed models trained on multilingual data. English→Dutch model was the best model for Dutch, both languages are from the same group of languages. Hindi and Korean languages, are from completely separated groups from languages in the training dataset, and results on these languages were worse than for the Dutch language. The Arabic language is also not related to any language from the training dataset, but it is known that caused by historical interaction of Arabic countries with Europeans, in some dialects of the Arabic language there are Italian, Spanish and Portuguese words. This could be an explanation why results for the Arabic language are better than for Hindi and Korean.

We didn't get the same optimal cross-lingual pairs as (Barbieri et al., 2021). For Hindi the best pair for their model was Italian→Hindi and for Arabic it was English→Arabic, while for our model the best combinations were Spanish→Hindi and Italian→Arabic. This indicates that the subject matter of corpus could be important when choosing optimal cross-lingual pair, which can be explained as the different cultures can be similar in some subjects, and completely different in others. Barbieri et al. identified trending topics for Arabic language, such as iPhone or vegetarianism, and the Portuguese dataset was dominated by comments on TV shows (Barbieri et al., 2021), while we used data specifically collected for intimacy analysis.

4.3 The Third Experiment - Clustering

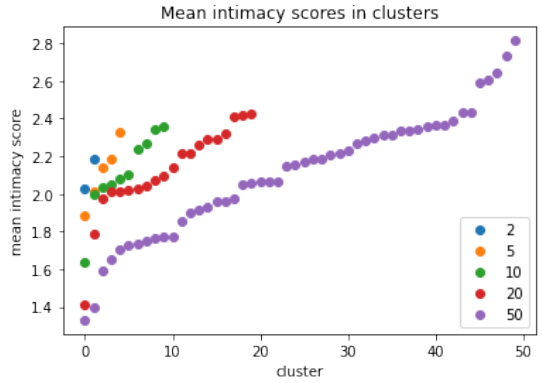
Results of the third experiments indicate that clustering data using XLM-T pre-trained embedding and K-means was more likely to be based on intimacy score than on the language of tweet. For the larger number of clusters, this effect is easier to

Table 2: Cross-lingual learning, Rows represent the language used for training and columns the language used for testing the model

	English	Spanish	Portuguese	Italian	French	Chinese	Hindi	Dutch	Korean	Arabic
English	0.68539	0.67927	0.61743	0.62033	0.55580	0.62802	0.16141	0.60715	0.28245	0.62557
Spanish	0.66484	0.71799	0.58425	0.67350	0.64219	0.56302	0.18252	0.53620	0.35988	0.57469
Portuguese	0.64256	0.65262	0.67576	0.59144	0.62305	0.58065	0.17705	0.57559	0.29469	0.60594
Italian	0.59816	0.66287	0.50119	0.65408	0.59649	0.54914	0.09593	0.54175	0.28236	0.63647
French	0.61330	0.67938	0.61435	0.65295	0.65731	0.60850	0.16394	0.55685	0.23085	0.60548
Chinese	0.62076	0.65002	0.44520	0.50465	0.56180	0.68010	0.07766	0.49669	0.34399	0.56764
All	0.70092	0.70955	0.66738	0.70949	0.66943	0.71029	0.23202	0.59981	0.31401	0.64795

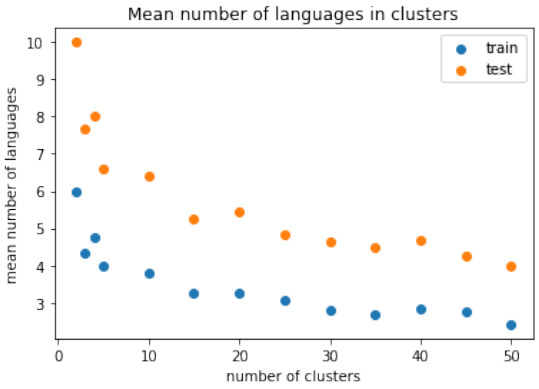
note, and the range of means of intimacy scores in clusters is wider. Larger number of clusters should enable more precisive detection of extremely intimate tweets and extremely non-intimate tweets. Results are shown in Figure 1.

Figure 1: The mean value of intimacy score in clusters for different number of clusters - train dataset



For a smaller number of clusters, all languages were present in all clusters. For the number of clusters above 4, some clusters were mono-lingual. With the increase in the number of clusters, more mono-lingual clusters appeared, but not all languages had their own mono-lingual clusters. With more than 10 clusters, usually, but depending on k-means initialization, all languages have their own clusters, while multi-lingual clusters still exist. In all cases, the first few the largest clusters were multi-lingual, but the size of all clusters was comparable. The mean number of languages in test clusters was on average larger than in the train clusters, because of unseen languages. Results are shown in Figure 2.

Figure 2: The mean value of number of languages in clusters for different number of clusters



5 Conclusion

This paper represents an analysis of intimacy detection in a multilingual data collected and proposed for SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis. The task was to train model to predict the level of intimacy in tweets, for languages present and non-present in train data. In three experiments, we compared feature-based, embedding-based and DL models, cross-lingual learning and clustering. Our study indicates that XLM-T pre-trained embedding might be a good choice for an unsupervised learning algorithm for intimacy detection. An increase in the training data even in different languages could improve performance of the model, and the best choice for training language might be made using prior knowledge about cultural or historical connections of languages. The clustering of data was more closely related to intimacy score than to the language of the Tweet, indicating that embedding generated using pre-trained language model can capture an hidden information meaningful for intimacy analysis. This finding indicates that properly collected and pre-processed data can be utilized for unsupervised learning, reducing the need for extensive data labeling efforts. Intimacy is one of the core components of our relationships and interpersonal interactions. Considering our everyday exposure to social networking platforms, it is important to continue researching on models for social factor detection in these media.

References

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond.

Taylor Blose, Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2020. Privacy in crisis: A study of self-disclosure during the coronavirus pandemic. arXiv.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. arXiv.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.

Claire Kramsch. 1998. Language and culture. England. Oxford University Press; 1st edition.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2022. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2015. Computational sociolinguistics: A survey. arXiv.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. volume 12, pages 2825–2830.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. arXiv.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozon Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2023. Semeval 2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Karen Prager, Forouz Shirvani, Jennifer Garcia, and Minnotis Coles. 2013. Intimacy and positive psychology. In *Positive Psychology of Love, Series in Positive Psychology*. Oxford Academic.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. volume 3.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. volume 8, pages 219563–219576.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety on reddit.

Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2019. Detection and analysis of self-disclosure in online news commentaries. In *The World Wide Web Conference*, WWW '19, page 3272–3278, New York, NY, USA. Association for Computing Machinery.

Qing Xie, Xinyuan Zhang, Ying Ding, and Min Song. 2020. Monolingual and multilingual topic analysis using lda and bert embeddings. volume 14, page 101055.