iREL at SemEval-2023 Task 9: Improving understanding of multilingual Tweets using Translation-Based Augmentation and Domain Adapted Pre-Trained Models

Bhavyajeet Singh, Ankita Maity, Pavan Kandru, Aditya Hari and **Vasudeva Varma**
IIIT Hyderabad
{bhavyajeet.singh, ankita.maity, siri.venkata, aditya.hari}@research.iiit.ac.in
vv@iiit.ac.in

Abstract

This paper describes our system (iREL) for Tweet intimacy analysis shared task of the SemEval 2023 workshop at ACL 2023. Our system achieved an overall Pearson's r score of 0.5924 and ranked 10th on the overall leaderboard. For the unseen languages, we ranked third on the leaderboard and achieved a Pearson's r score of 0.485. We used a single multilingual model for all languages, as discussed in this paper. We provide a detailed description of our pipeline along with multiple ablation experiments to further analyse each component of the pipeline. We demonstrate how translation-based augmentation, domain-specific features, and domain-adapted pre-trained models improve the understanding of intimacy in tweets. The code can be found at https://github.com/bhavyajeet/Multilingual-tweet-intimacy

1 Introduction

Social media platforms have become a ubiquitous source of communication, with millions of users generating and sharing content daily. Among these, Twitter has emerged as a popular platform for sharing short messages or tweets, which can express a wide range of emotions, opinions, and sentiments.(Liew and Turtle, 2016) (Cislaru, 2015) Such tweets can also provide valuable insights into various social and political issues, making them a valuable resource for researchers and policymakers alike.

Intimacy in language refers to the degree of emotional closeness or familiarity between individuals, which is often reflected in the choice of words, tone, and context of communication. (Wynne and Wynne, 1986). The degree of intimacy can vary based on the relationship between individuals, and it can significantly impact on the results of social interactions.

In recent years, the analysis of intimacy in social media data has gained some attention (Pei and Jurgens, 2020a), as it can provide valuable insights into various aspects of human behavior, such as the formation of social networks, the spread of information, and the dynamics of online communities. However, the analysis of intimacy in social media data is challenging, as it requires sophisticated methods for processing and interpreting large volumes of unstructured text data. Multilingual tweet intimacy analysis is even more challenging, as it involves the complexity of multiple languages and cross-cultural communication.

The task (Pei et al., 2022a) aims at quantifying the intimacy of tweets from a total of 10 different languages in the form of a score between 1 to 5. This paper describes our system used for analysing the intimacy of tweets belonging to 10 different languages. We utilise domain-specific features and domain-adapted pre-trained models in order to improve the understanding of intimacy in tweets. We further utilise a translation-based data augmentation pipeline which proves effective in significantly improving the scores for unseen languages. This paper also discusses multiple ablations applied to our pipeline in order to better understand the contribution of the various components used. Our system achieved third rank for unseen languages with a Pearson's r score of 0.485 and tenth rank overall with a Pearson's r score of 0.592.

2 Related Work

Intimacy can be defined as a particular form of "closeness" to another person founded on self-disclosure, with the quality of self-disclosure characterized by knowledge and understanding of inner selves (Jamieson, 2007; Wynne and Wynne, 1986). Several vital attributes are related to this social phenomenon. It occupies an important position in the hierarchy of human needs as one of the more basic needs (Maslow, 1943). As such, several critical social attributes and functions are associated with it, such as in development (Harlow and Zimmermann,

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2052–2057

1958) and well-being (Sneed et al., 2011).
As this role of prominence would suggest, it plays an integral part in the discourse. This needs to be accounted for in computation methods of textual analysis. Generally, social factors are challenging to interpret for NLP systems, resulting in errors. Methods incorporating social factors have been shown to result in improvements in the performance of such models (Hovy and Yang, 2021). Intimacy specifically can indicate close relations between the participants (Pei and Jurgens, 2020b) or be used to reduce the social distance between the participants (Keshavarz, 2001) Few works have investigated computational methods of dealing with intimacy. (Pei and Jurgens, 2020b) establish intimacy as an impactful social dimension of language by analyzing a dataset collected from Reddit, an online forum.

Recent works have also discussed the idea of using translation based augmentation (Kim et al., 2019) (Chen et al., 2019). These works argue that such augmentation techniques can help the multilingual deep learning models better transfer learnings across languages.

Our work utilises the understandings and inferences from previous research works of this domain to construct a novel pipeline for analysing intimacy in multilingual tweets.

3 Data

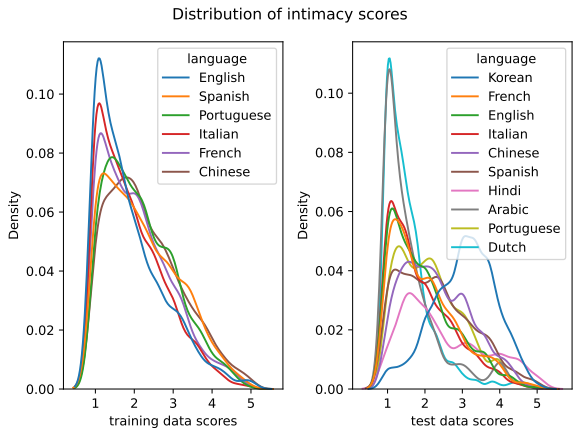

Figure 1: The distribution of intimacy scores of test and training data

We used the Multilingual intimacy analysis (MINT) dataset(Pei et al., 2022b) for this task. The training dataset contains a total of 9491 tweets spread over 6 different languages: English, Spanish, Portuguese, Italian, French and Chinese, and the annotated test data contained a total of 13697 tweets from 10 different languages, which included the 6 languages from the training set and an additional 4 unseen languages : Hindi, Korean, Dutch, and Arabic. The dataset contains three fields, the language of the tweet, tweet text, and the intimacy score. The distribution for intimacy scores for the test and train dataset can be seen in Figure 1. As it can be seen, the distribution of training and test dataset are similar, however the test data contains a greater proportion of tweets with very low (close to one) intimacy scores for some languages, like Dutch and Arabic.

As it can be seen, the distribution of intimacy scores for Hindi and Korean are very different from the rest of the languages, which can be a potential reason for the relatively worse performance of the finetuned transformer models on these languages.

4 Methodology

Our approach utilises sentence representation from BERT-based models, along with components to address the domain-specific features like emojis and augmentation techniques for better transfer of cross-lingual learning. This section describes the entire pipeline of our system.

Our best-performing model contains the following components.

4.1 Data Cleaning

Since tweets posted online are very unstructured in nature and can have a lot of noise in the text, we preprocess the training and test data using the following heuristics.

- All the links in the twitter text are replaced with the text "http"
- We remove all usernames from the text and replace them with the text "@user" instead
- We further remove noise present in the form of repeated punctuation, extra spaces, etc.

4.2 Domain specific pre-trained transformer models

We experiment with 4 different transformer based models. Our best performing model is built on top of TwHIN-BERT (Zhang et al., 2022). TwHIN-BERT is a multilingual language model trained on 7 billion tweets from 100 different languages. TwHIN-BERT is trained on the objective of text-based self-supervision, and a social objective based

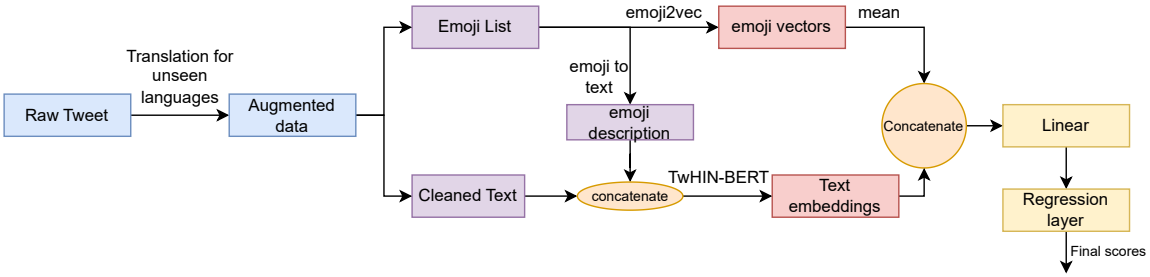

Figure 2: The figure describes the pipeline for the proposed system

on the rich social engagements within a Twitter heterogeneous information network (TwHIN).

4.3 Emoji representation

Our final model uses a two phase process to process emojis. The two components are as follows

- We extract all the emojis from the text. We then use the emoji2vec (Eisner et al., 2016) system in order to get fixed-length vector representations for emojis. The vector representations for all emojis are mean pooled to give a single 300 dimension vector, which is concatenated with the sentence embeddings from the transformer before feeding it to the neural regression head. If there is no emoji present, we forward a vector of all zeros
- We use the python package emoji[1] to convert emoji to text and replace every emoji in text with its textual description provided by the module.

4.4 Translation based data augmentation

Recent works (Kim et al., 2019) (Chen et al., 2019) have shown that training multilingual transformer models on translated data of the target language can lead to better cross-lingual transfer of learning. Hence we augment the training data to contain instances from the unseen languages by translating the instances from other languages. We use the data from Chinese, English, Portuguese, and Spanish to translate into Korean, Hindi, Dutch, and Arabic respectively. We use source data from different languages in order to maintain the semantic diversity of the training data. Since the distribution of the test labels for the unseen languages may not be the same as that of the seen languages, we pick a uniform distribution of intimacy score labels while selecting the samples to be translated. We use the googletrans python module [2] for all our translations.

5 Results

Table 1 Shows the results of our experiments and ablation studies performed. All the scores shown use Pearson's r[3] correlation between the ground truth and the predictions as the scoring metric. The first column shows the scores achieved by the submitted system. The following five groups of columns show results achieved in the different verticals of ablations. For each experiment, the entire setup is same except for the modifications in the ablation vertical. For all the experiments, the default transformer model used is xlm-T and the default pipeline setup is the same as described in Section 4

The first group of ablations studied the effect of changing the pre-trained transformer model in the pipeline. The second group focuses on the effects of data preprocessing and cleaning. The third group shows the results with two strategies of processing the emojis, one where emojis are removed, and one where we only convert emojis to text without using the emoji2vec system. The fourth group shows the effects of not augmenting data for unseen languages using translation (no trans), or the effect of translating the unseen languages to English for inference (trans test). Finally, the last column shows the results without the use of emojis and without the translation-based augmentation.

Our submitted system achieved an overall score of 0.592, ranking tenth on the leaderboard, and a score of 0.485 for unseen languages, ranking third on the leaderboard. The submitted system and our best-performing system (tweet-bert with all the components as described in Section 4) only differ in terms of hyperparameters which are explained in detail in section 7

[1] https://pypi.org/project/emoji/

[2] https://pypi.org/project/googletrans/

[3] Pearson's r correlation

Ablation		pretrained model				filtering	emojis		translation		others
	submitted system	distill-bert	mbert	TwHIN-bert	xlm-T	no cleaning	no emoji	emoji-2-text	no trans	trans test	no trans no emoji
English	0.706	0.602	0.636	0.688	0.706	0.704	**0.723**	0.706	0.704	0.702	0.715
Spanish	0.725	0.604	0.622	**0.727**	0.711	0.720	0.705	0.709	0.694	0.680	0.678
Portuguese	0.648	0.514	0.545	0.606	**0.676**	0.671	0.674	0.668	0.645	0.652	0.645
Italian	**0.727**	0.590	0.558	0.710	0.698	0.694	0.690	0.692	0.694	0.709	0.695
French	0.628	0.559	0.580	0.631	0.675	0.681	0.674	0.674	0.681	0.680	**0.692**
Chinese	0.698	0.666	0.664	0.721	0.714	0.717	0.720	**0.729**	0.720	0.677	0.708
Hindi	0.203	0.176	0.174	0.189	**0.217**	0.160	0.184	0.184	0.200	0.235	0.206
Dutch	0.591	0.488	0.487	0.567	**0.630**	0.608	**0.611**	0.603	0.602	0.604	0.592
Korean	0.307	0.277	0.269	**0.404**	0.358	0.306	0.372	0.359	0.322	0.319	0.374
Arabic	**0.644**	0.395	0.365	0.637	0.605	0.572	0.647	0.623	0.653	0.604	0.628
Seen Lang	0.684	0.591	0.612	0.687	0.704	**0.707**	0.699	0.702	0.694	0.684	0.693
Unseen Lang	0.485	0.410	0.384	**0.516**	0.477	0.471	0.484	0.477	0.434	0.367	0.420
Overall	0.592	0.512	0.510	**0.605**	0.602	0.601	0.601	0.600	0.573	0.535	0.570

Table 1: The table shows the results for all the experiments and the ablation studies. The first column highlights our submitted system. All the other columns highlight different ablation experiments where one of the components of our pipeline is modified or removed

6 Ablation Study

This section describes multiple experiments done in order to better understand the contributions from the different components of our pipeline. For each of the experiments, the basic pipeline remains the same, and each ablation study removes or modifies a specific component of the pipeline to understand its contribution and impact. Unless stated otherwise, we use XLM-T model as the underlying decoder for all our experiments.

6.1 Transfer learning with different BERT-based model

The BERT model has displayed impressive results in various natural language understanding tasks, such as sentiment analysis and identifying hate speech. The RoBERTa(Liu et al., 2019) model has demonstrated better performance than BERT in different NLU tasks. Researchers have developed various methods to pre-train BERT-like models from scratch using a vast amount of data specific to a particular domain to learn unique language patterns.

We experiment with two such models, pretrained on large amounts of Twitter data. The TwHIN-BERT (Zhang et al., 2022), which is a multilingual language model based on the BERT architecture, trained on 7 billion tweets from 100 different languages; and XLM-T (Barbieri et al., 2021) which is an XLM-R (Liu et al., 2019) model pre-trained on millions of tweets in over thirty languages.

We finetune multilingual BERT, XLM-RoBERTa, TwHIN-BERT, and XLM-T on our training data. We use a linear layer for the regression task over sentence embeddings that we get from these models. Experiment results show that TwHIN-BERT performed the best for overall and unseen languages; however XLM-T performed the best for seen languages. We also note that while the performances of XLM-T and TwHIN-BERT are similar, they both significantly outperform the other two models. While our submitted and best performing models utilise TwHIN-BERT, for our further experiments in the ablation, we used XLM-T since that is considered to be more robust in general.

6.2 Effects of data cleaning and pre-processing

Table 1 shows that removing data cleaning does not have a massive impact on the overall performance of the model, however, it can be noted that using cleaned data leads to minor gains in the overall performance and the performance over unseen languages.

6.3 Using emoji representations

We experiment with two ablations on our pipeline concerning the handling of emojis. For one of the ablations, we do not use emojis at all and remove them from the text. This results in a significant drop in the scores for seen languages and a minor drop in the overall score. In a second ablation, we use only the textual descriptions of the emojis, and not their vector representations. This also leads to worse performances as compared to using both.

6.4 Translation based ablations

As described in 4, we used translated data in order to improve the intimacy understanding over the unseen languages. For our ablations, we try two approaches. The first approach does not make use of any translated data during training and uses the text from unseen languages without any modifications during inference. This approach results in a significant drop in the performance for unseen languages and also for the overall scores, proving that translation of training data is helpful for this task.

For the second ablation, we do not translate the training data; instead, we translate the data from the unseen languages to English during inference, considering that English is one of the most resource rich languages, and the model performed well on English. This resulted in a greater drop in the overall scores, and those for unseen languages. A possible reason for this could be the propagation of errors from the translation pipeline, since translations are not perfect, or the differences in the style and distribution of actual English tweets and the translated ones.

6.5 Without translation and emoji processing

The final experiment shows the model's performance without the use of translations on training text and without any specific pipeline for handling emojis. As it can be seen this performs worse than the complete pipeline as well as the other two ablations of emojis and translations. Further strengthening the fact that both emoji processing and translations are helpful to our system.

7 Implementation Details

All the models are trained on a single Nvidia GTX 1080 Ti GPUs. For all our experiments, we use a learning rate of 1e-5 and a dropout of 0.3. We optimize our models on the mean squared error (MSE) loss using AdamW (Kingma and Ba, 2014) optimiser.

7.1 Submitted system

For the submitted system, we used a batch size of 6, with a training of 3 epochs. The maximum token length for each sample was 512 and the samples were padded to the maximum length.

7.2 All other systems

For our best performing model and all the other reported experiments, we used a batch size of 18, with a training of 5 epochs. The maximum token length for each sample was 256 and the samples were padded to the maximum length.

8 Conclusion

This work shows how carefully designed data augmentation techniques can help in better cross lingual transfer of learning and improved scores over unseen languages. The work also highlights the importance of using efficient encoding strategies to include domain specific features like emojis for an improved understanding of the text. The results also show that, though efficient, the transformer based deep learning models are prone to variance, and just utilising the right set of hyperparameters can result in significant gains. Finally, we also see that domain-adapted pre-trained transformers can capture nuances of in-domain text, and when used with simple deep learning and machine learning models, could give competitive results.

References

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

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Georgeta Cislaru. 2015. Emotions in tweets: From instantaneity to preconstruction. *Social Science Information*, 54(4):455–469.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description.

Harry F. Harlow and Robert R. Zimmermann. 1958. The development of affectional responses in infant monkeys. *Proceedings of the American Philosophical Society*, 102(5):501–509.

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

Technologies, pages 588–602, Online. Association for Computational Linguistics.

Lynn Jamieson. 2007. *Intimacy*. John Wiley Sons, Ltd.

Mohammad Keshavarz. 2001. The role of social context, intimacy, and distance in the choice of forms of address. *International Journal of The Sociology of Language*, 2001:5–18.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Jasy Suet Yan Liew and Howard R. Turtle. 2016. Exploring fine-grained emotion detection in tweets. In *Proceedings of the NAACL Student Research Workshop*, pages 73–80, San Diego, California. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

A. H. Maslow. 1943. A theory of human motivation. *Psychological Review*, 50:370–396.

Jiaxin Pei and David Jurgens. 2020a. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online. Association for Computational Linguistics.

Jiaxin Pei and David Jurgens. 2020b. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online. Association for Computational Linguistics.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozon Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2022a. Semeval 2023 task 9: Multilingual tweet intimacy analysis.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozon Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2022b. Semeval 2023 task 9: Multilingual tweet intimacy analysis.

Joel Sneed, Susan Whitbourne, Seth Schwartz, and Shi Huang. 2011. The relationship between identity, intimacy, and midlife well-being: Findings from the rochester adult longitudinal study. *Psychology and aging*, 27:318–23.

Lyman C. Wynne and Adele R. Wynne. 1986. The quest for intimacy*. *Journal of Marital and Family Therapy*, 12(4):383–394.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations.