ODA_SRIB at SemEval-2023 Task 9: A Multimodal Approach for Improved Intimacy Analysis

Priyanshu Kumar, Amit Kumar, Jiban Prakash, Prabhat Lamba and **Irfan Abdul**
{priyanshu.k, amit.kumar, p.jiban, prabhat.l, irfan.abdul}@samsung.com

Samsung Research and Development Institute
Bangalore, India

Abstract

We experiment with XLM-Twitter and XLM-RoBERTa models to predict the intimacy scores in Tweets i.e. the extent to which a Tweet contains intimate content. We propose a Transformer-TabNet based multimodal architecture using text data and statistical features from the text, which performs better than the vanilla Transformer based model. We further experiment with Adversarial Weight Perturbation to make our models generalized and robust. The ensemble of four of our best models achieve an over-all Pearson Coefficient of 0.5893 on the test dataset.

1 Introduction

Intimacy is an important aspect in our society, which in general is not given much attention. With the increasing usage of social media across all age groups, users might post some private intimate information unintentionally. Thus, there is a dire need to create Natural Language Processing (NLP) models to detect the intimate nature of such posts as the user can be nudged before they share some intimate information on social media. Moreover, the capability of understanding intimacy will also help us incorporate emotions in voice assistants and chatbots. For the Task 9 of SemEval 2023: Multilingual Tweet Intimacy Analysis (Pei et al., 2022), the organizers intend the participants to build models for detecting the intimacy of a given Tweet on a scale of 1 to 5. They provide a dataset of annotated Tweets in prominent languages used around the world - English, Spanish, Portuguese, Italian, French and Chinese. In order to validate the generalization of submitted systems, they also have test samples in Hindi, Dutch, Korean and Arabic.

In this paper, we present our methodology to develop such models. We formulated the problem as a binary classification problem on soft labels. We rely on XLM-Twitter (Barbieri et al., 2021) and XLM-Roberta (Conneau et al., 2019) models' cross lingual transfer capability for unseen languages. We devise a multi-modal architecture with text and handcrafted features as input. We experimented with augmented data which we created by translating the text while keeping the intimacy score the same. We incorporate Adversarial Weight Perturbation (AWP) (Wu et al., 2020) to train our models in a robust manner. Our final submission is a mean ensemble of four of our best models.

Pearson's coefficient is used to compare the models submitted to the task leaderboard. Our submission attains an overall score of 0.5893 with 0.7318 on seen languages and 0.411 on unseen languages.

2 Background

The shared task of Multilingual Tweet Intimacy Analysis is a text regression problem i.e. for a given a Tweet, we need to predict the intimacy score for it. The organizer provide a dataset comprising of 13,372 tweets in 10 languages (English, French, Spanish, Italian, Portuguese, Korean, Dutch, Chinese, Hindi, and Arabic). The training data contains samples from English, French, Spanish, Italian, Portuguese and Chinese; the remaining languages are a part of the test data so as to evaluate the generalization of the developed models.

Intimacy, being an crucial aspect of our society, has been studied extensively in the field of socio-linguistics and social psychology. There are limited works on computational modeling of intimacy. Pei and Jurgens (2020) curate a dataset comprising questions, which reveal the people's impression of intimacy. By developing NLP models which correlate with human annotations, they analyze the questions posted on popular social media sites, books and movie dialogues, and present their insights.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1676–1680

3 System Overview

3.1 Problem formulation

Though the problem is a regression problem, we treat it as a binary classification problem with soft labels, once we scale the labels in between 0 and 1 using the following equation.

$$y_{scaled} = \frac{y - 1}{4}$$

During the initial phase, we experiment our baseline model with the two different formulations and observe that the classification setting provides a better learning signal to the model. Therefore, we use Binary Cross Entropy (BCE) loss for our experiments. The probability values predicted by the model are re-scaled to lie in the range 1 to 5 using

$$\hat{y} = 4 * y_{pred} + 1$$

3.2 Text Preprocessing

We perform minimal text preprocessing steps, only to normalize texts to follow a standard. Some Tweets contain the user handles of prominent users or accounts; we use rules to convert such mentions to *@user*.

3.3 Data Augmentation

Since details about the hidden languages in the test data is known, we attempt to generate additional data by simply translating all the train data into hidden languages. We leverage Multilingual BART (Liu et al., 2020) [1] for translation. Manual inspection of augmented data showed that emojis were often skipped during translation. Therefore, we add all emojis occurring in a text, at the end of the translated text.

We do not make use of the dataset prepared by Pei and Jurgens (2020), which comprises of question-structure text. Hence, there is a considerable data distribution difference between this data and the shared task data. In addition, there is a discrepancy in the intimacy score range as well (scores lie between -1 to 1).

3.4 Models

We mainly experiment with XLM-Roberta (Conneau et al., 2019) models and its other versions. Since the dataset comprises of tweets, using CardiffNLP's XLM-Twitter (Barbieri et al., 2021) is beneficial. We also experimented with Microsoft's mDeberta (He et al., 2021) but found the results below par with XLM-Roberta (Table 1).

For our baseline architecture, we extract the [CLS] token representations from the transformer backbones, which are then passed through multi-dropout linear layers to generate multiple scores. These scores are then averaged together to create the final model output.

We also experiment with a multi-modal architecture (Figure 1) that takes the text and some handcrafted features as input. The motivation for such an architecture is to examine if the features provide some additional learning signal to the model. We create the following features:

- Number of emojis
- Number of *@user* mentions
- Number of *#tag* mentions
- Number of url mentions
- Number of prominent user/account mentions

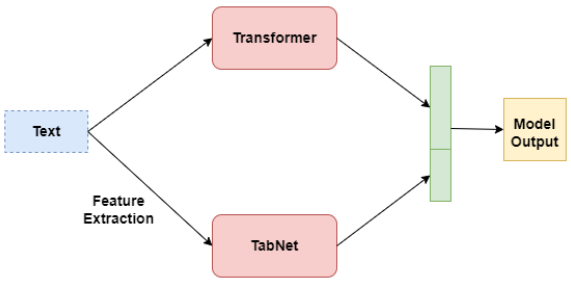

Figure 1: Multi-modal architecture leveraging handcrafted textual features

We use TabNet architecture (Arik and Pfister, 2021) on top of the handcrafted features. We set the hyperparameters of TabNet so as to prevent overfitting: $n\_d = 4$, $n\_a = 4$, $n\_steps = 2$ and $output\_dim = 64$, where n_d is the width of the decision prediction layer, n_a is the width of the attention embedding for each mask, n_steps is the number of steps in the TabNet architecture and $output\_dim$ is the dimension of the final hidden features extracted by the TabNet model. The hidden representation of the features are concatenated with the output of the Transformer backbone (along the last axis) and then follows a prediction head similar to our baseline model.

[1] https://huggingface.co/facebook/mbart-large-50-one-to-many-mmt, https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

Since we frame the problem as a binary classification problem, the model outputs are post-processed to scale the score between 1 and 5.

3.5 Adversarial Weight Perturbation

Works on adversarial training in NLP has risen in recent years. AWP is an adversarial technique that helps to enhance the robustness of a model by slightly perturbing its weights during training. The perturbations are calculated based on the norm of the parameters and their gradients. The key difference between the work of Miyato et al. (2016) and AWP is that the former applies the perturbations only to the word embeddings whereas the latter applies them to all the weights of the model.

4 Experimental Setup

We perform a 5-fold cross validation to train our models. The continuous label values can be binned to create ordinal buckets. Since the objective of the shared task is to develop models that can perform well on unseen languages as well, the correct cross-validation strategy is to use a stratified split grouped on language. Since some language data will only be present in the validation split and missing in the training data, we will be able to get an accurate estimation of the generalization capability of the model on unseen language data.

However, the details of the unseen languages in the test data is known and hence we can use a stratified split (without any grouping) and rely on the cross-lingual transfer of multilingual models for generalization across unseen languages. In this manner, we leverage the training data to the fullest. In some experiments, the augmented data is added to the training data of each fold.

We train our models on a Nvidia Tesla P40 card with a batch size of 16 and sequence length 160. The models are optimized using AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of $5e-6$ for 50 epochs with an early stopping patience of 7 epochs. For models trained using AWP, the AWP learning rate is set to $1e-2$ and AWP is applied from the start of the training. The models have been implemented using Pytorch (Paszke et al., 2019) and Huggingface's Transformers library (Wolf et al., 2019).

5 Results and Discussions

In this section, we tabulate the performance of our models as well as the mean ensemble of the best four on the task test data, language wise. Table 1 shows the 5-fold cross validation score of our experiments. As stated in Section 3.1, training in a binary classification setting leads to improvement in score. The multilingual Deberta architecture lags considerably behind the XLM Roberta variants. In addition, AWP also helps us to enhance the score.

Backbone	Description	Score
XLM-T	MSE	0.6921
XLM-T	BCE	0.7025
mDeberta	BCE	0.6205
XLM-T	BCE + AWP	**0.7202**
XLM-T	BCE + TabNet + AWP	**0.7176**
XLM-T	BCE + Augmented data	**0.6969**
XLM-Large	BCE + AWP	**0.7134**

Table 1: Cross Validation score of experiments; models selected in the ensemble are shown in bold

However, the augmented data degrades the performance. A possible reason of this behavior might be the quality of translation. User generated Tweets are quite noisy in nature thus making translation a difficult task. Moreover, some of the texts also contain expletives whose translation in some other language might change the context of the Tweet.

Table 2 shows the evaluation metric of our final submission, which is a mean ensemble of our top four performing models. We observe that the cross-validation score (computed on seen languages only) correlates very well with the test score on seen languages. For the convenience of comparing results with respect to the best leaderboard results in every language, we tabulate the data in Table 4.

Setting	Score	Best
Cross Validation	0.7307	-
Test Seen	0.7318	0.7509
Test Unseen	0.411	0.4998
Test Overall	0.5893	0.616

Table 2: Pearson coefficient comparison of cross validation and test for submission

With the help of the ground truth data obtained from the organizers, we evaluate our best models individually on the test data (Table 3). We observe that even though XLM-Roberta Large has not been pretrained on Tweets, it performs significantly well on 2 seen languages and completely outperforms the other XLM-T based models on the unseen languages. Our multi-modal TabNet based architec-

Language	XLM-T	XLM-T TabNet	XLM-T Augmented Data	XLM-Large
English	0.7079	**0.719**	0.6848	0.6966
Spanish	0.7393	0.7385	0.7186	**0.7418**
Portuguese	0.6937	**0.6941**	0.6734	0.6827
Italian	**0.7262**	0.7252	0.7026	0.7045
French	0.7024	**0.7144**	0.6573	0.6941
Chinese	0.7301	0.7196	0.6962	**0.742**
Hindi	0.1843	0.2114	0.1967	**0.2359**
Dutch	0.6219	0.6126	0.6024	**0.6487**
Korean	0.3492	0.3677	0.3239	**0.3816**
Arabic	0.6212	0.6021	0.5967	**0.6322**
Overall	0.5842	0.5809	0.552	**0.5858**
Seen	0.7257	**0.7284**	0.6937	0.7182
Unseen	0.4049	0.3929	0.3744	**0.4229**

Table 3: Language wise Pearson coefficient comparison of best models on test dataset

Language	Ours	Best
English	0.716	0.758
Spanish	0.746	0.784
Portuguese	0.7021	0.7022
Italian	0.732	0.742
French	0.708	0.726
Chinese	0.735	0.762
Hindi	0.212	0.276
Dutch	0.642	0.678
Korean	0.368	0.419
Arabic	0.635	0.662

Table 4: Language wise Pearson coefficient comparison on test dataset

Language	Score
English	0.7023
Spanish	0.7335
Portuguese	0.6811
Italian	0.6952
French	0.7061
Chinese	0.7425
Hindi	0.2683
Dutch	0.6115
Korean	**0.4272**
Arabic	0.6114
Overall	**0.588**
Seen	0.7159
Unseen	**0.432**

Table 5: Language wise Pearson coefficient of XLM-Roberta Large TabNet on test dataset

ture also portrays strong learning capability as it achieves the best score for three seen languages and also for all seen languages combined.

In order to check the capability, we perform an additional experiment with an XLM-Roberta Large model based on the TabNet architecture (Table 5). Although the cross-validation score is only 0.701, the model scores the best on Korean, Overall and Unseen data splits as compared to our submitted models. The overall score nears the score of our submitted ensemble. We believe better hyperparameter tuning for this model will lead to better scores.

6 Conclusion

In this paper, we present our methodology for estimating intimacy scores of Tweets. We perform our experiments with XLM-Twitter and XLM-Roberta models. We also propose a multimodal model leveraging text and statistical text features and observe that the handcrafted features enable learning in a better way. Our models are trained with Adversarial Weight Perturbation to improve their stability. Our mean ensemble of four best models achieves a Pearson Coefficient of 0.5893. On further analysis, we observe that XLM-Roberta Large model outperforms the other models on the unseen languages benchmark even though the other models' backbone was finetuned on Tweets and the its TabNet based variant closes in on our submitted ensemble as per the overall score .

References

Sercan Ö Arik and Tomas Pfister. 2021. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. *arXiv preprint arXiv:2011.03020*.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozon Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2022. Semeval 2023 task 9: Multilingual tweet intimacy analysis. *arXiv preprint arXiv:2210.01108*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969.