WADER at SemEval-2023 Task 9: A Weak-labelling framework for Data augmentation in tExt Regression Tasks

Manan Suri[1]*, **Aaryak Garg**[1]*, **Divya Chaudhary**[2],
Ian Gorton[2], **Bijendra Kumar**[1]
Netaji Subhas University of Technology, New Delhi[1]
Northeastern University, Seattle [2]
{manan.suri.ug20,aaryak.ug20}@nsut.ac.in

Abstract

Intimacy is an essential element of human relationships and language is a crucial means of conveying it. Textual intimacy analysis can reveal social norms in different contexts and serve as a benchmark for testing computational models' ability to understand social information. In this paper, we propose a novel weak-labeling strategy for data augmentation in text regression tasks called WADER. WADER uses data augmentation to address the problems of data imbalance and data scarcity and provides a method for data augmentation in cross-lingual, zero-shot tasks. We benchmark the performance of State-of-the-Art pre-trained multilingual language models using WADER and analyze the use of sampling techniques to mitigate bias in data and optimally select augmentation candidates. Our results show that WADER outperforms the baseline model and provides a direction for mitigating data imbalance and scarcity in text regression tasks.

1 Introduction

Intimacy is considered a fundamental element of human relationships, as recognized by several scholars (Sullivan, 1993; Maslow, 1981; Prager, 1995). Research indicates that intimacy can be modeled computationally and that textual intimacy is a crucial aspect of language (Pei and Jurgens, 2020). Analyzing textual intimacy can reveal social norms in various contexts and serve as a benchmark to test computational models' ability to understand social information (Pei and Jurgens, 2020; Hovy and Yang, 2021). Moreover, intimacy plays a critical role in human development and well-being (Harlow and Zimmermann, 1959; Sneed et al., 2011), and language is an essential means of conveying it in a social context. Individuals negotiate intimacy in language to fulfill fundamental and strategic needs while respecting social norms.

Task 9 of SemEval 2023 (Pei et al., 2022) aims to quantify intimacy in a multilingual context, with evaluation on tweets from 10 languages. The training corpus for the task consists of tweets in English, Spanish, Italian, Portuguese, French, and Chinese. The testing corpus additionally contains tweets from Hindi, Arabic, Dutch and Korean.

The novelty of our strategy, WADER (Weak-labeling strategy for Data augmentation in tExt Regression Tasks) is the use of data augmentation to A) solve the problem of an imbalance distribution of data, B) augment data for a cross-lingual zero-shot set-up. WADER uses the distribution to selectively sample texts with lower representation in the label distribution, uses translation to augment sentences and validates the augmentations against a baseline model, using a distribution based sampling approach. We finetune State-of-the-Art pre-trained language models including XLM RoBERTa (Conneau et al., 2019) and XLNET (Yang et al., 2019). Real world datasets are plagued by the problems of data imbalance and data scarcity, and WADER provides a direction for mitigating these problems for text regression tasks. WADER ranks 32nd overall across languages, 34th on seen languages and 29th on unseen languages. Our code has been released on GitHub. [1]

The main contributions of this paper are as follows:

1. Provide a data augmentation framework specific to text regression.
2. Provide a method for data augmentation in cross-lingual, zero-shot tasks.
3. Benchmark performance of pre-trained language models.
4. Analysis of use of sampling techniques to mitigate bias in data and optimally select augmentation candidates.

*Equal contribution.

[1] https://github.com/Darthfire/wader

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1945–1952

The paper is organized as follows: Section 2 provides background information on the research, including a review of relevant literature, details about the task at hand, and information on the data used. Section 3 presents an overview of our approach, followed by a discussion of the experimental set-up in Section 4. The results of our study are analyzed in Section 5, and the paper concludes with a summary of findings and future directions for research in Section 6.

2 Background

2.1 Past Work

Data imbalance and scarcity are problems that are rampant in real world datasets. The high cost of obtaining large amounts of data, and expert annotations, a wealth of research has been done to support limited data settings. Data augmentation for text is broadly done in two ways, conditional data augmentation which involves data augmentation conditioned by the target label, and unconditional data augmentation which involves working with the corpus features only (Bayer et al., 2021; Liu et al., 2020). Conditional data augmentation is done usually by deep generative models and pretrained language models such as BART (Lewis et al., 2019), CBERT (Wu et al., 2018), GPT2 (Radford et al., 2019). Common ways to perform unconditional data augmentation are lexical substitution and back translation. (Wei and Zou, 2019) introduce several lexical techniques to augment textual data, including synonym replacement, random insertion, random swap and random deletion. However, these methods suffer from lack of sufficient diversity and often produce sentences that are not coherent. Back-translation especially has received widespread attention, because progress in machine translation has made back-translation an efficient way to generate diverse sentences in the dataset without compromise in coherence and semantic quality. Common translation tools used are seq2seq based models, NMT and transformers. Different techniques exist for text classification and NER tasks, but to the best of our knowledge our work is unique in the text regression domain.

Weak supervision of text labeling during data augmentation is an example of Semi-Supervised Learning (SSL) methods. The main idea of these methods is to regularize the learning process by training a network with the given data, using the network to label unlabelled data and finally use both the true-labeled and weak-labeled data points to train the final model.

2.2 Task Description

SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis (Pei et al., 2022) is a task that deals with detecting intimacy in 10 languages. This task is co-organized by University of Michigan and Snap Inc. Intimacy is a fundamental aspect of human relationships, and studying intimacy in a textual context has many potential applications in the field of computational linguistics. The training data is available in 6 languages: English, Spanish, Italian, Portuguese, French, and Chinese. The evaluation is done on the given training languages, as well as 6 unseen languages: Hindi, Arabic, Dutch and Korean.

The metric of evaluation for the task is Pearsons R. Pearsons R, r is expressed as follows for two variables x and y:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (1)$$

The correlation coefficient r ranges from -1 to 1, with an absolute value of 1 indicating a perfect linear relationship between the two variables. In such a case, all data points lie on a line that can be represented by a linear equation. The sign of the correlation coefficient is determined by the regression slope, with a value of +1 indicating that all data points lie on a line where y increases as x increases, and a value of -1 indicating the opposite. A correlation coefficient of 0 implies that there is no linear dependency between the two variables.

2.3 Data Description

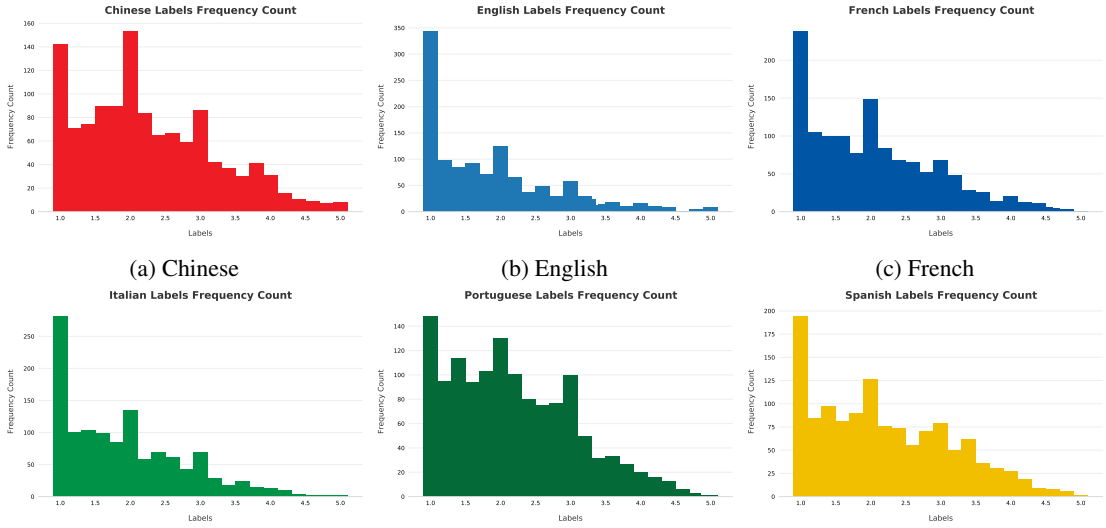
The dataset for the task is the MINT- Multilingual INTimacy analysis (Pei et al., 2022) dataset. The training set contains sentences in 6 languages: Chinese, English, French, Portuguese, Spanish and Italian. The dataset has 9491 tweets. Distribution of sentences in different languages is given in Table 1.

Intimacy is annotated using a 5-point Likert scale where 1 indicates Not intimate at all and 5 indicates Very intimate. (Cite) have described the annotation process in detail.

The dataset is highly imbalanced, with majority of the labels in each language belonging to the lower spectrum of the scale as seen in Fig1. Overall,

Language	Count	Mean	Std. Dev.	25th %ile	50th %ile	75th %ile
English	1587	1.89	0.877273	1.2	1.6	2.4
Chinese	1596	2.27	0.93851	1.5	2	2.8
French	1588	2.06	0.886265	1.34	2	2.6
Italian	1532	1.94	0.835105	1.25	1.8	2.425
Spanish	1592	2.21	0.941339	1.4	2	2.8
Portuguese	1596	2.16	0.872903	1.4	2	2.8
Overall	9491	2.09	0.903512	1.4	2	2.67

Table 1: Description of the training set.



(a) Chinese (b) English (c) French

(d) Italian (e) Portuguese (f) Spanish

Figure 1: Frequency plots for different languages in the training set.

75% of the samples in the dataset have a label less than or equal to 2.667.

The testing set additionally contains 4 unseen languages, Hindi, Korean, Arabic and Dutch.

3 Methodology

3.1 Data Augmentation

As noted in section 2.3, the data is highly imbalanced for the given labels. Moreover, since the task has 4 unseen languages, there is an additional need for data augmentation. WADER performs data augmentation using the framework described in Fig 2. The steps followed are described as follows:

Distribution based Sampling: Since the distribution of labels is skewed, and not all labels need augmentation, we perform a distribution based sampling to select candidate tweets for data augmentation. We fix a threshold p, and sample all tweets above the given threshold. We take the value of p as 3.2, and less

Translation: Data is augmented through translation and back translation. The translation scheme is described in Fig 3.

For an unseen language, L_{unseen}, set of sampled sentences $L_i \forall i \in L$ are taken and translated to the target language, L_{unseen}. The translated sentences are appended to the set T_{unseen}.

For a seen language, say L_i, the language is translated to all other languages except the language itself, $L_k \forall k \in L, k \neq i$. The translated tweets are appended to their respective translated sets T_k, and they are translated back to the source language L_i, appended to the translated set T_i.

Our final translated set has 49774 sentences.

Label Validation: We train a baseline model by finetuning a pretrained language model on the gold labelled data. This model is then used to infer on the concatenated translated corpus of seen and unseen languages.

Difference Based Sampling: We take the absolute difference between predicted and pre-assigned values (derived label from before a sentence was translated). We use this as a metric for quality of translations and pick appropriate thresholds to se-

Parameters	count	mean	std	min	25%	50%	75%	max
Value	49774	0.62	0.51	0	0.23	0.47	0.86	3.550781

Table 2: Analysis of the translated sentence set, specifically the difference during validation.

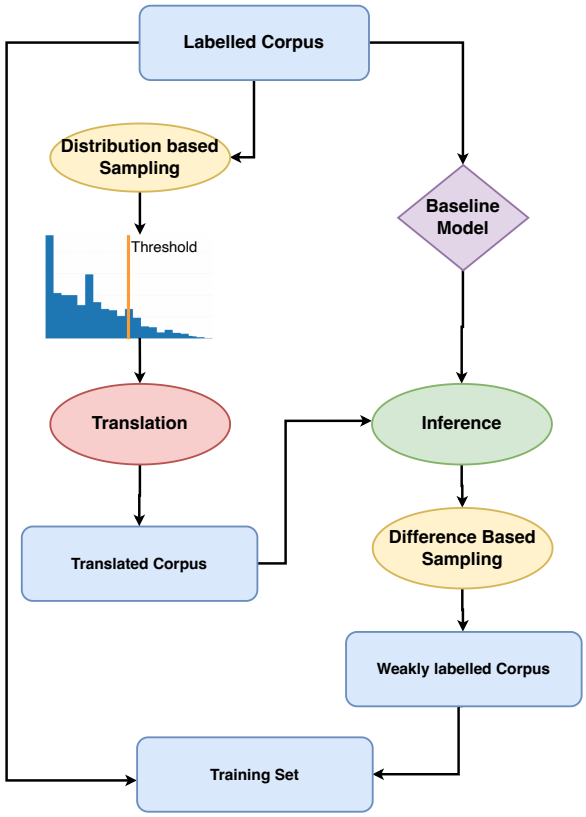


Figure 2: Data Augmentation Flowchart

Difference	Count
<= 0.1	5102
<= 0.2	10581
<= 0.3	16187

Table 3: Count of sentences with chosen absolute difference threshold β after label validation.

lect sentences. Table 2 shows an analysis of the distribution of differences. The mean difference is 0.62, which is a below average translation quality since the resolution of labels is 0.1 in the dataset. However, 75% of sentences have differences $\leq$ 0.86 For our experiments, which means that coarse grain labels (differences of 1) are correctly assigned in most of the cases.

We define β as the parameter which represents the difference threshold. We pick difference values of β as 0.1, 0.2 and 0.3 in our experiments. Table 2 shows the count of sentences in each of these thresholds.

3.2 Finetuning Pre-trained Language Models

Finetuning pretrained language models has become a popular approach for natural language processing tasks in recent years. Transformer based (Vaswani et al., 2017) Pretrained Language Models such as BERT (Devlin et al., 2018), GPT-2(Radford et al., 2019), and RoBERTa(Liu et al., 2019) are trained on massive amounts of text data, which allows them to capture complex linguistic patterns and structures. Finetuning involves taking a pretrained language model and further training it on a specific downstream task, such as sentiment analysis or question answering. This approach has been shown to achieve state-of-the-art performance on a wide range of natural language processing tasks, with significantly less data and computation needed compared to training a model from scratch. Finetuning also allows for the transfer of knowledge learned from a large, diverse set of data to a smaller, more specific task, making it a powerful technique for natural language processing research.

The pre-training models used in our system include:

XLM RoBERTa: XLM-RoBERTa (Conneau et al., 2019) is a variation of the RoBERTa model that has been designed to handle multilingual natural language processing tasks. This model is pre-trained on a massive dataset of 2.5 terabytes of CommonCrawl data filtered for 100 different languages. By training on such a large and diverse dataset, XLM-RoBERTa is able to capture the linguistic nuances and patterns that are unique to different languages. The architecture of XLM-RoBERTa is based on the highly successful BERT model, but with key modifications to hyperparameters such as larger mini-batches and learning rates, allowing it to handle the additional complexity of multilingual data. XLM-RoBERTa has shown impressive results across a range of multilingual natural language processing tasks, demonstrating the power of pre-training on large, diverse datasets for building highly effective models. We use implementation of the XLMR model from HuggingFace .

XLNET: XLNet (Yang et al., 2019) is a state-of-the-art natural language processing model that

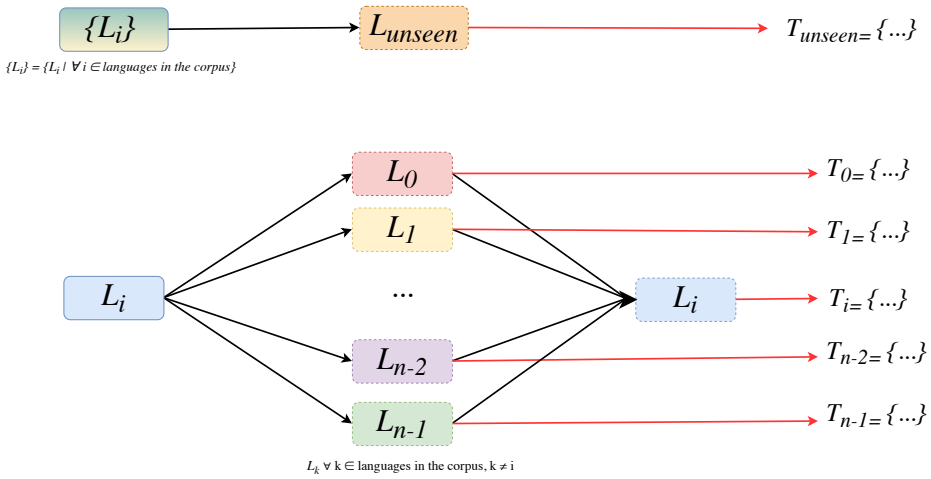


Figure 3: Translation scheme

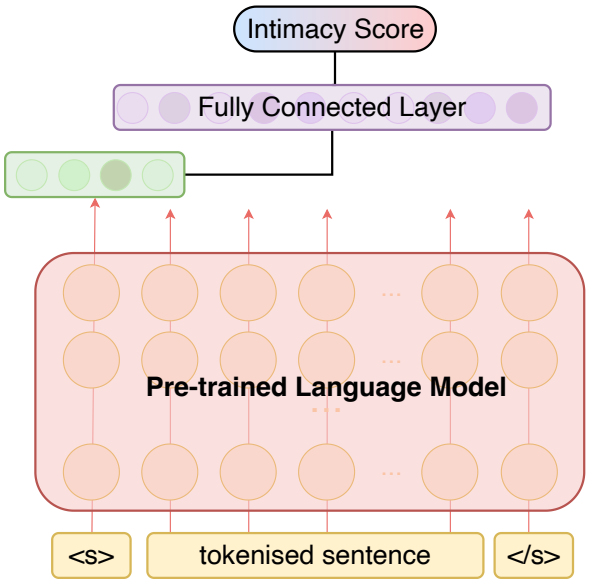


Figure 4: Fine-tuning Architecture

extends the Transformer-XL architecture and uses an innovative pre-training method. Unlike BERT, which corrupts input with masks and neglects dependencies between masked positions, XLNet is able to learn bidirectional contexts by maximizing the expected likelihood over all possible permutations of the factorization order. This allows XLNet to capture complex linguistic patterns and dependencies in the input sequence. XLNet also integrates ideas from Transformer-XL, which is currently the most advanced autoregressive model in use. With its autoregressive formulation, XLNet is able to overcome the limitations of BERT and achieve even better performance on a range of natural language processing tasks.

For finetuning the pre-trained models, we add a single linear layer on top of the embeddings of the classification token `<s>` for XLM-RoBERTa and `[cls]`. Since this is a text regression task and the scores are in a limited, we apply a clamp function as the final activation function which clamps the scores in the range $[1, 5]$. Fig 4 is a representation of our finetuning procedure.

3.3 Ensembling

We evaluate results on the test set using 6 models, XLM RoBERTa and XLNET trained on augmented sets with difference sampling parameter $\beta = 0.1, 0.2, 0.3$.

We choose 6 ensembles. The configurations of ensembles are defined in Table 4.

Ensemble	XLM RoBERTa			XLNET		
	0.1	0.2	0.3	0.1	0.2	0.3
Ensemble 1						
Ensemble 2						
Ensemble 3						
Ensemble 4						
Ensemble 5						
Ensemble 6						

Table 4: The configurations of the different chosen ensembles that we experimented with. The different choices are motivated by A) Model choice, B) Threshold of difference sampling β.

Ensembling is done by taking the mean prediction of all the ensembled models.

System	Overall	Seen Langs.	Unseen Langs.	English	Spanish	Portuguese	Italian	French	Chinese	Hindi	Dutch	Korean	Arabic
Baseline-XLM RoBERTa	0.52	0.65	0.35	0.60	**0.69**	0.60	0.64	0.60	0.70	0.19	0.59	0.37	0.42
0.1-XLM RoBERTa	0.52	0.66	0.34	0.61	0.66	0.60	0.67	0.63	0.72	0.19	0.59	0.35	0.48
0.2-XLM RoBERTa	0.52	0.67	0.33	0.63	0.66	**0.61**	0.67	**0.64**	**0.72**	0.19	0.60	**0.38**	0.49
0.3-XLM RoBERTa	0.53	0.66	0.35	0.63	0.67	0.60	0.67	**0.64**	**0.72**	**0.20**	**0.61**	0.43	0.50
Baseline-XLNET	0.38	0.51	0.22	0.62	0.61	0.42	0.47	0.47	0.24	-0.08	0.37	-0.03	0.05
0.1-XLNET	0.41	0.52	0.26	**0.64**	0.61	0.47	0.49	0.49	0.20	-0.08	0.41	-0.03	0.14
0.2-XLNET	0.41	0.51	0.29	0.61	0.58	0.43	0.50	0.51	0.24	-0.05	0.45	0.08	0.22
0.3-XLNET	0.42	0.52	0.29	0.61	0.63	0.46	0.53	0.50	0.19	-0.06	0.44	0.16	0.19
Ensemble-1	0.53	0.67	0.34	0.63	0.68	**0.61**	**0.68**	**0.64**	0.72	0.20	0.61	0.40	**0.49**
Ensemble-2	0.43	0.53	0.30	0.63	0.63	0.47	0.52	0.52	0.22	-0.06	0.45	0.08	0.20
Ensemble-3	0.52	0.64	0.36	**0.64**	0.67	0.58	0.63	0.60	0.67	0.11	0.55	0.29	0.45
Ensemble-4	0.52	0.63	0.37	0.63	0.66	0.57	0.62	0.60	0.67	0.11	0.56	0.34	0.48
Ensemble-5	0.52	0.64	**0.38**	**0.64**	**0.69**	0.57	0.64	0.61	0.64	0.11	0.57	0.41	0.47
Ensemble-6	**0.53**	**0.65**	0.37	**0.64**	0.68	0.58	0.64	0.61	0.67	0.11	0.57	0.36	0.48

Table 5: Pearson's R score of different system settings on the test set. $\beta - Model$ represents $Model$ finetuned on Gold labels + β difference set.

Team	Overall	Seen Languages	Unseen Languages	English	Spanish	Portuguese	Italian	French	Chinese	Hindi	Dutch	Korean	Arabic
WADER	32	34	29	34	32	36	35	34	34	40	30	15	35

Table 6: Rank achieved by our system in the shared task.

4 Experimental Setup

We use the original test and train set. Further, we take 15% of the train set, sampled randomly from each language as our validation set.

We build our models using open source available implementations of the XLM-RoBERTA and XLNET available on HuggingFace. We use `xlnet-base-cased` `xlm-roberta-base`[2] and [3]. We use Adam (Kingma and Ba, 2014) as our optimiser. The size of the the embeddings are $D \in 768$ and the size of the linear layer is $D/2 \times 1$. The batch size is taken as 8 and the learning rate is 4e-5. We train the models for 2 epochs. Experiments are performed on Google Colab cloud GPU. Google Translate API has been used to perform translations. These hyperparameters are common for all system settings including our two baselines: 1) XLM RoBERTa finetuned on only Gold data, 2) XLNET finetuned on only Gold data.

Final submission is reported on Ensemble 6, configured as per the desciption in Section 3.3.

5 Results and Discussion

Table 5 represents the scores achieved by our system in different experimental settings. The final submission for the competition is denoted by Ensemble 6. Table 6 shows our rank under different categories of the shared task.

As we can observe from Table 5, WADER seems to improve on existing transformer baselines for all categories except one where it ties with an ensemble.

5.1 Comparison of Pre-trained Language Models:

We observe a general trend that XLM RoBERTa performs better than XLNET on multilingual baselines in our experiments. This can be demonstrated by the fact that the XLNET Baseline outperforms XLM RoBERTa only on English. For all other languages, there is a significant margin in between performance of XLNET and XLM RoBERTa. For Hindi, and Korean which have non latin characters, performance of XLNET is even worse with a negative R coefficient. which has This demonstrated the importance of multilingual pretraining.

5.2 Comparison of Difference Sampling Threshold β:

While lower values of $\beta(= 0.1)$ give more accurate labelled sets, we observe that moderate values of $\beta(= 0.1, 0.2)$ outperform them. This is because, moderate values of β allow for larger sized training corpuses, which would positively effect the performance of the models. Moreover, moderate values β include more number of low quality translations, due to a higher difference. We hypothesize that this would have a regularising effect by providing the model with diversity in the training set, and preventing it from overfitting on the training corpus.

5.3 Discussion on Performance

We rank 32 overall, 34 on seen languages and 29 on unseen languages. The lower performance of our model can be understood by the following factors:

[2] https://huggingface.co/xlm-roberta-base

[3] https://huggingface.co/xlnet-base-cased

- **Translation Quality:** The quality of translation is a key driver in WADER's performance. Lower quality translations would produce augmentations with noisy and unreliable labels. Translation quality is often dependant on the pair of languages in question. For languages such with a non latin script such as Hindi, translations are often of a lower quality which is also reflected in the results.

- **Overfitting:** By translating the data, while we increase linguistic diversity, most sentences would still be semantically similar, causing the model to overfit. This can further be seen by the fact that settings like 0.2, 0.3-XLM RoBERTA (where we can expect higher diversity from gold sentences due to higher differences) give the best performance for a lot of languages. Similarly, Ensemble 1 which preserves data quality while also reaping the regularising benefit of ensembling performs quite well in the given setting. Another indication of overfitting is the the better rank of our model on unseen languages.

- **Word Sensitivity:** For a task like Intimacy Detection, specific vocabulary used is key to identify the intimacy level. Translations can lead to replacement of words which do not hold the same degree of influence in accounting for textual intimacy.

6 Conclusion and Future Work

This paper proposes a novel data augmentation framework, WADER, for text regression tasks that use weak-labeling strategies to solve the problems of data imbalance and data scarcity. We also provide a method for data augmentation in cross-lingual, zero-shot tasks. Our approach uses sampling techniques to mitigate bias in data and optimally select augmentation candidates. We benchmarked the performance of State-of-the-Art pre-trained multilingual language models XLM RoBERTa using WADER and achieved promising results. Our findings demonstrate the importance of data augmentation for mitigating data imbalance and scarcity in text regression tasks. This study's contributions provide a direction for future research in the field of computational linguistic and its applications to social information analysis.

References

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2021. A survey on data augmentation for text classification. *CoRR*, abs/2107.03158.

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. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Harry F. Harlow and Robert R. Zimmermann. 1959. Affectional response in the infant monkey. *Science*, 130(3373):421–432.

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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Pei Liu, Xuemin Wang, Chao Xiang, and Weiye Meng. 2020. A survey of text data augmentation. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pages 191–195.

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. *CoRR*, abs/1907.11692.

Abraham Harold Maslow. 1981. *Motivation and personality*. Prabhat Prakashan.

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

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*.

Karen Jean Prager. 1995. *The psychology of intimacy*. Guilford Press.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Harry Stack Sullivan. 1993. *The Interpersonal Theory of Psychiatry*. Routledge.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*, abs/1901.11196.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2018. Conditional BERT contextual augmentation. *CoRR*, abs/1812.06705.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.