Multi-Task Learning for German Text Readability Assessment

Salar Mohtaj[1,2,*], Vera Schmitt[1,2], Razieh Khamsehashari[1] and Sebastian Möller[1,2]

[1]Technische Universität Berlin, Berlin, Germany
[2]German Research Centre for Artificial Intelligence (DFKI), Labor Berlin, Germany

Abstract

Automated text readability assessment is the process of assigning a number to the level of difficulty of a piece of text automatically. Machine learning and natural language processing techniques made it possible to measure the readability and complexity of the fast-growing textual content on the web. In this paper, we proposed a multi-task learning approach to predict the readability of German text based on pre-trained models. The proposed multi-task model has been trained on three tasks: text complexity, understandability, and lexical difficulty assessment. The results show a significant improvement in the model's performance in the multi-task learning setting compared to single-task learning, where each model has been trained separately for each task.

Keywords

Text readability assessment, Multi-task learning, Transfer learning, Text complexity

1. Introduction

Automated text readability assessment is the task of analyzing the difficulty of a piece of text for a target group. Text readability assessment has a wide range of applications, from empowering language learners to find proper reading materials to learn a new language [1] to helping people with disabilities [2]. However, manual assessment of text readability is not an option nowadays due to the fast pace of online content creation on the web. Automated techniques use machine learning and Natural Language Processing (NLP) models to analyze the complexity of a piece of text and spontaneously assign a readability score to textual contents. Automated text readability is the task of assigning a difficulty level to an input text. The readability score is the mapping of a piece of text (e.g., a short sentence or a paragraph) to a mathematical unit (i.e., text regression) which is the basis of the readability assessment. Text readability assessment could be designed as a text classification [3] or regression [4] task, depending on the input labels.

In this paper, we present a Multi-Task Learning (MTL) approach based on pre-trained language models for the task of German text readability assessment. We used three metrics that present the readability to train our proposed model. These metrics include complexity, understandability, and lexical difficulty of German texts in the sentence level. Recently, pre-trained large language models showed promising results and could outperform state-of-the-art deep neural network-based models in different NLP tasks either in fine-tuning [5] and feature extraction settings [6]. On the other side, MTL models have had successes not only in NLP tasks but also in speech recognition and computer vision [7].

The proposed MTL model is based on the available readability scores in the *TextComplexityDE* data set [8]. The data set includes three readability-related scores (i.e., complexity, understandability, and lexical difficulty scores) for 1,000 German text samples. We assumed that the knowledge in the prediction of one of these scores could be used and transferred into the prediction of the others, due to the relatedness of these scores. As a result, we propose an MTL model in which some layers are shared between the tasks. The obtained results from the experiments show that the MTL approach could significantly improve the overall performance of the prediction of all three scores compared to the single-learning setting, where each task has been trained separately.

The rest of this paper is organized as follows; Section 2 reviews the recent research on automated German text readability assessment. The *TextComplexityDE* data set is briefly explained in Section 3. The proposed MTL model and the obtained results on the tasks of text complexity, understandability, and lexical difficulty prediction are presented in Sections 4 and 5, respectively. Finally, in Section 6, we conclude the paper and discuss the potential future research directions.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
salar.mohtaj@tu-berlin.de (S. Mohtaj);
vera.schmitt@tu-berlin.de (V. Schmitt);
razieh.khamsehashari@tu-berlin.de (R. Khamsehashari);
sebastian.moeller@tu-berlin.de (S. Möller)
0000-0002-0032-3833 (S. Mohtaj); 0000-0002-9735-6956 (V. Schmitt); 0000-0003-3057-0760 (S. Möller)

2. Related Work

In this section, we review some of the recent efforts in using NLP and machine learning models for the evaluation of text readability.

A supervised model for German text readability assessment is proposed in [9]. They have extracted more than 70 features grouped in traditional, lexical, and morphological-based features to train text regression models. They have selected the top 20 features for the training phase based on different criteria, such as the low ratio of missing values and also low correlation between features. The obtained results show that the *Random Forest* model could outperform *Linear Regression* and *Polynomial Regression* models with respect to the Root Mean Squared Error (RMSE) metric. They improved the results on the same data set by fine-tuning pre-trained language models in [10]. They used pre-trained models in feature extraction and fine-tuning settings and came to the conclusion that the fine-tuning approach could outperform the feature extraction as well as classical machine learning models.

A sentence-wise readability assessment model for German L2 readers is introduced in [11]. They extracted 373 features from different types (e.g., syntax) to train machine learning models for the regression and ranking tasks. The Bayesian Ridge Regression model outperforms the widely used readability formulae in the regression task in their experiments. They also analyzed the complexity at the document level and found that the maximum complexity in the sentence level impacts the document complexity.

A hybrid model combining a feature engineering approach and transfer learning for German text complexity assessment is proposed in [12]. They have extracted word level and sentence level features from text and ensemble it with transformer-based models like Bert [13] and RoBERTa [14]. The proposed model achieved the first ranking in the *Text Complexity DE Challenge 2022* [15].

An online service for assessing the readability of German text based on machine learning models is presented in [16]. The authors provided the model as an online service that is publicly available to use. The online service provides five statistical metrics and two machine learning models for an input text. The machine learning models are based on the BERT and the fine-tuned BERT. They achieved promising results on two different data sets based on Mean Square Error (MSE) and Mean Absolute Error (MAE) metrics [16].

To the best of our knowledge, there is no text readability prediction model for German text based on MTL approaches. The proposed model uses the benefits of pre-trained language models as well as a multi-task learning approach where features that form good predictors for multiple tasks are favoured over those that don't.

3. Data Set

In this section, we describe the data set that has been used to train and test the proposed models in this paper.

We used *TextComplexityDE*[1] data set [8] to train the proposed model and also to test it against single-task learning approaches. In this section, we briefly describe the data set, especially the available readability scores in the data that make it possible to train multi-task learning models.

As thoroughly explained in [8], *TextComplexityDE* data set contains 1,000 sentences in the German language taken from 23 Wikipedia articles from three different topics. The sentences were annotated by German learners in levels A and B who were 32 years old on average and mostly held a university degree. Each sentence is mapped to the Mean Opinion Score (MOS) of three different readability metrics, namely complexity, understandability, and lexical difficulty. All the sentences have been rated by multiple annotators on a 7-point Likert scale. The complexity shows how complex a sentence for an annotator was in the range of very easy (1) to very complex (7). The understandability metric shows how well the participants were able to understand a sentence, and the lexical difficulty presents the difficulty of the most difficult word in a sentence.

This data set has been used as the training set in the *Text Complexity Challenge on German Text* in 2022. In order to train and also evaluate the single- and multi-task learning models in this paper, we split the data set into the train, validation, and test parts (60%, 20%, and 20%, respectively).

Figure 1 shows the distribution of MOS values over the training and test data sets for the three metrics. As presented in the figure, there are more easy instances in the data set than complex ones.

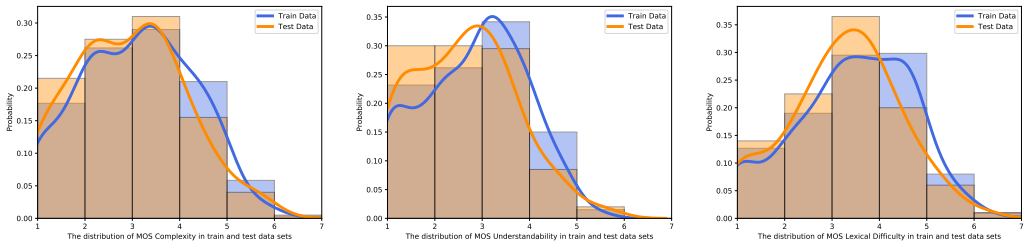
Table 1 provides a summary of statistics and frequency distribution of the training and test data sets. As described in the table, the training and test sets follow a similar distribution from the textual content and readability scores point of view.

4. Multi-task Learning Model

In this section, we present our model based on a multi-task learning approach to predict the complexity score of textual input and the understandability and lexical difficulty scores. We use pre-trained language models to extract features from the input text and feed the extracted features into a Recurrent Neural Network (RNN) as the initial hidden state.

Due to the fact that MTL can learn features that generalize better across tasks and considering the relation

[1] https://github.com/babaknaderi/TextComplexityDE



(a) (b) (c)

Figure 1: The distribution of MOS values over the train and test data sets for the a) **complexity**, b) **understandability**, and c) **lexical difficulty** scores.

	Training data	Test data
Number of records (i.e., sentences)	600	200
Max length of sentences (in character)	439	487
Min length of sentences (in character)	23	19
Average length of sentences (in character)	151.3	143.7
Number of words	12366	3886
Number of unique words	5258	2075
Mean complexity score (Standard Deviation)	3.10 (1.19)	2.93 (1.17)
Mean understandability score (StD)	2.84 (1.08)	2.63 (1.08)
Mean lexical difficulty score (StD)	3.45 (1.22)	3.25 (1.16)

Table 1
Summary of statistics and frequency distribution of the training and test data sets

between three readability scores in the *TextComplexityDE* data set, we propose a joint model for the task. Considering the similarity between the three tasks and in order to enable knowledge sharing among tasks, we used a parallel architecture (i.e., tree-like architecture) [17] in this work.

We use the German BERT model [18] (i.e., *bert-base-german-cased*) in a feature extraction setting where the input text is fed into the model to convert textual input into vectors. The model includes a shared layers part that is shared between three regression models (i.e., complexity, understandability, and lexical difficulty prediction) and a unique task-specific layer for each task. The overall architecture of the model is depicted in Figure 2 (a).

As presented in Figure 2, the output of the BERT model is fed into a two layers Bi-GRU model [19]. As an RNN model GRUs can handle sequence input very well and showed promising results in text readability prediction in the previous studies [20]. A fully connected layer is on top as the last layer of the shared layers.

The task-specific layer includes a separated, fully connected layer that is connected to the task-specific output layer. The following hyper-parameters are tested during the training phase in order to find the best configuration for this task. The best-performing parameters are highlighted.

- Learning rate: **0.001**, 0.0005, 0.0001
- Batch size: **32**, 64
- Dropout probability: 0.3, **0.4**, 0.5
- Size of the hidden layer: 64, **128**, 256

Moreover, we trained all the models in 50 epochs and set the early stopping patience to 10 checkpoints to prevent over-fitting. In other words, the training has stopped in case of no improvement in ten continuous epochs. The model has *110,125,315* parameters in total and *1,043,971* trainable parameters since the parameters from the pre-trained model are frozen and didn't change during the training phase.

Regarding the loss weighting strategy, we used the "optimizing worst-case task loss" strategy, in which the worst-performing task has been chosen in each step as the optimization target. The importance of worst-case task loss compared to the vanilla average task loss when training an MTL model is analyzed in [21]. The achieved results on the test data set are presented in the next section.

5. Evaluation and Results

In this section, we briefly describe the evaluation metric used to measure the performance of the proposed model

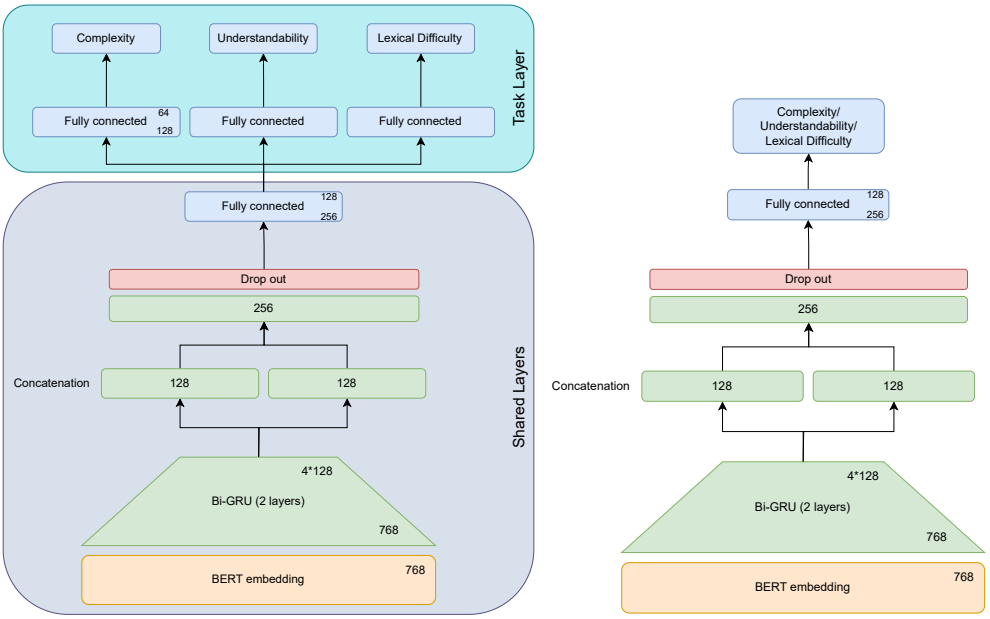


Figure 2: The architecture of the *a)* multi-task learning, and *b)* single learning setting. The same architecture is used to train models for the tasks of **complexity**, **understandability**, and **lexical difficulty** prediction in the single learning setting.

and the obtained results from the MTL model as well as a single-task learning model as the baseline.

5.1. Evaluation Metric

The Root Mean Square Error (RMSE) metric is used to evaluate the models' performance. It measures the root of the average squared difference between the estimated values (e.g., complexity scores) and the actual value. It is a common metric for regression analysis including text readability assessment.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \widehat{y}_i)^2}{N}} \quad (1)$$

where y_i is ith actual value, $\widehat{y}_i$ is the ith predicted value and N is the number of data points.

5.2. Results

We evaluated the performance of the proposed MTL model on the test set of the data. We compared the obtained results in the MTL setting with the single-task learning setting as the baseline. The overall performance of the single-task and multi-task learning modules are presented in Table 2.

We used a similar architecture for the single-task learning model. The single-task learning model includes the same embedding layer (i.e., the German BERT model) and the same 2-layers Bi-GRU layers on top. In this model, the output of the fully connected layer is fed directly to the output layer as depicted in Figure 2 (b). The single-task learning model has *1,019,137* trainable parameters (compared to *1,043,971* trainable parameters of the MTL model). We used the same model to train the text regression to predict text complexity, understandability, and lexical difficulty scores, separately.

As presented in the table, the MTL setting significantly outperforms the single-learning model in all three tasks. Moreover, the average error of the three tasks (*0.7945*) is much lower in the MTL model compared to the situation where each model is trained separately (*0.8379*).

It also should be noted that the number of trainable parameters is almost the same in both models ($\sim 0.025\%$ more parameters in the MTL model). In contrast, the single-task learning model undergoes three separate training sessions, one for each task. So, in addition to achieving a better performance in predicting German text readability, the MTL model also demonstrates higher computational efficiency.

The obtained results from the MTL setting highlight the importance of the prediction of text readability score from different perspectives. In other words, the results show that the performance of a text complexity predictor could be improved by introducing other related metrics

Task	Single-task setting	Multi-task setting
Complexity	0.7558	**0.7155**
Understandability	0.9436	**0.9287**
Lexical difficulty	0.8143	**0.7393**
Average	0.8379	**0.7945**

Table 2
The performance of single-task learning and multi-task learning approaches on the prediction of complexity, understandability, and lexical difficulty scores.

such as understandability and lexical difficulty to the model.

6. Conclusion

In this paper, we proposed a model based on a multi-task learning approach for the task of text readability assessment in German text. The model is trained and tested on the *TextComplexityDE* data set. It is simultaneously trained on three different readability scores, namely complexity, understandability, and lexical difficulty. Our results showed that the MTL model outperforms the common single-task learning models in all three scores. The obtained results in this experiment reveal the importance of the annotation of text readability from different perspectives.

As the direction for future studies, different multi-task learning architectures (e.g., hierarchical architectures) could be tested in the task. Moreover, in this study, we exclusively tested the BERT model to extract features from the input text. However, exploring and assessing the impact and the performance of other pre-trained models is a question for future works. Finally, the performance of fine-tuning approaches of transfer learning can be compared to the feature extraction approach in future studies.

Acknowledgments

The present study was funded by the Deutsche Forschungsgemeinschaft (DFG) through the project "Analyse und automatische Abschätzung der Qualität maschinell generierter Texte", project number 436813723.

References

[1] M. Xia, E. Kochmar, T. Briscoe, Text readability assessment for second language learners, in: J. R. Tetreault, J. Burstein, C. Leacock, H. Yannakoudakis (Eds.), Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA, The Association for Computer Linguistics, 2016, pp. 12–22.

[2] S. Aluisio, L. Specia, C. Gasperin, C. Scarton, Readability assessment for text simplification, in: Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications, 2010, pp. 1–9.

[3] S. Chatzipanagiotidis, M. Giagkou, D. Meurers, Broad linguistic complexity analysis for greek readability classification, in: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@EACL, Online, April 20, 2021, Association for Computational Linguistics, 2021, pp. 48–58. URL: https://www.aclweb.org/anthology/2021.bea-1.5/.

[4] P. G. Blaneck, T. Bornheim, N. Grieger, S. Bialonski, Automatic readability assessment of german sentences with transformer ensembles, in: Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text, GermEval@KONVENS 2022, Potsdam, Germany, September 12, 2022, Association for Computational Linguistics, 2022, pp. 57–62. URL: https://aclanthology.org/2022.germeval-1.10.

[5] Z. Zhao, Z. Zhang, F. Hopfgartner, A comparative study of using pre-trained language models for toxic comment classification, in: J. Leskovec, M. Grobelnik, M. Najork, J. Tang, L. Zia (Eds.), Companion of The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021, ACM / IW3C2, 2021, pp. 500–507.

[6] S. Mohtaj, S. Möller, On the importance of word embedding in automated harmful information detection, in: P. Sojka, A. Horák, I. Kopecek, K. Pala (Eds.), Text, Speech, and Dialogue - 25th International Conference, TSD 2022, Brno, Czech Republic, September 6-9, 2022, Proceedings, volume 13502 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 251–262.

[7] S. Ruder, An overview of multi-task learning in deep neural networks, CoRR abs/1706.05098 (2017). URL: http://arxiv.org/abs/1706.05098. arXiv:1706.05098.

[8] B. Naderi, S. Mohtaj, K. Ensikat, S. Möller, Sub-

jective assessment of text complexity: A dataset for german language, CoRR abs/1904.07733 (2019). arXiv:1904.07733.

[9] B. Naderi, S. Mohtaj, K. Karan, S. Möller, Automated text readability assessment for german language: A quality of experience approach, in: 11th International Conference on Quality of Multimedia Experience QoMEX 2019, Berlin, Germany, June 5-7, 2019, IEEE, 2019, pp. 1–3. doi:10.1109/QoMEX.2019.8743194.

[10] S. Mohtaj, B. Naderi, S. Möller, F. Maschhur, C. Wu, M. Reinhard, A transfer learning based model for text readability assessment in german, CoRR abs/2207.06265 (2022). doi:10.48550/arXiv.2207.06265. arXiv:2207.06265.

[11] Z. Weiss, D. Meurers, Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference?, in: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), Association for Computational Linguistics, Seattle, Washington, 2022, pp. 141–153. URL: https://aclanthology.org/2022.bea-1.19. doi:10.18653/v1/2022.bea-1.19.

[12] A. Mosquera, Tackling data drift with adversarial validation: An application for German text complexity estimation, in: Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text, Association for Computational Linguistics, Potsdam, Germany, 2022, pp. 39–44.

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). arXiv:1907.11692.

[15] S. Mohtaj, B. Naderi, S. Möller, Overview of the germeval 2022 shared task on text complexity assessment of german text, in: S. Möller, S. Mohtaj, B. Naderi (Eds.), Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text, GermEval@KONVENS 2022, Potsdam, Germany, September 12, 2022, Association for Computational Linguistics, 2022, pp. 1–9. URL: https://aclanthology.org/2022.germeval-1.1.

[16] F. Pickelmann, M. Färber, A. Jatowt, Ablesbarkeitsmesser: A system for assessing the readability of german text, in: Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 288–293.

[17] S. Chen, Y. Zhang, Q. Yang, Multi-task learning in natural language processing: An overview, CoRR abs/2109.09138 (2021). URL: https://arxiv.org/abs/2109.09138. arXiv:2109.09138.

[18] B. Chan, S. Schweter, T. Möller, German's next language model, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, International Committee on Computational Linguistics, 2020, pp. 6788–6796.

[19] K. Cho, B. van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder–decoder approaches, in: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 103–111.

[20] Y. Sun, K. Chen, L. Sun, C. Hu, Attention-based deep learning model for text readability evaluation, in: 2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020, IEEE, 2020, pp. 1–8.

[21] P. Michel, S. Ruder, D. Yogatama, Balancing average and worst-case accuracy in multitask learning, CoRR abs/2110.05838 (2021). arXiv:2110.05838.