Overview of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text

Salar Mohtaj[1,2], Babak Naderi[1], and Sebastian Möller[1,2]

[1]Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
[2]German Research Centre for Artificial Intelligence (DFKI), Labor Berlin, Germany
{salar.mohtaj|babak.naderi|sebastian.moeller} @ tu-berlin.de

Abstract

In this paper we present the GermEval 2022 shared task on Text Complexity Assessment of German text. Text forms an integral part of exchanging information and interacting with the world, correlating with quality and experience of life. Text complexity is one of the factors which affects a reader's understanding of a text. The mapping of a body of text to a mathematical unit quantifying the degree of readability is the basis of complexity assessment. As readability might be influenced by representation, we only target the text complexity for readers in this task. We designed the task as text regression in which participants developed models to predict complexity of pieces of text for a German learner in a range from 1 to 7. The shared task was organized in two phases; the development and the test phases. Among 24 participants who registered for the shared task, ten teams submitted their results on the test data.

1 Introduction

Text forms an integral part of exchanging information and interacting with the world. Along with the other types of content (e.g., image and video), textual content has been increased drastically in amount and importance during recent years. Text complexity (in the following used interchangeably with text readability) is one of the factors which affects a reader's understanding of text (Dale and Chall, 1949). Readability is concerned with the relation between a given text and the cognitive load of a reader to comprehend it. This complex relation is influenced by many factors, such as a degree of lexical and syntactic sophistication, discourse cohesion, and background knowledge (Crossley et al., 2017; Martinc et al., 2021). A readability score is the mapping of a body of text to a mathematical unit quantifying the degree of readability. It is the basis of readability assessment. Readability assessment has diverse use cases and applications, such as helping people with disabilities and also facilitate choosing of learning material for second language learners (Aluisio et al., 2010).

In this paper, we present the challenge and results from the task of German text complexity assessment in *GermEval* 2022. The task includes developing Natural Language Processing (NLP) models to automatically assign a complexity score in the range from 1 to 7 to German texts, where 1 represent an easy to understand (i.e., simple) text/sentence and 7 shows a complex text for German learners. In other words, the shared task is a text regression task in which the output is a continuous variable between 1 and 7.

GermEval is a series of shared task evaluation campaigns that focus on Natural Language Processing for the German language. It started in 2014 with a shared task on German Named Entity Recognition (Benikova et al., 2014) and continued in the years after with different tasks from lexical substitution (Miller et al., 2015) to the task of identification of toxic, engaging, and fact-claiming comments (Risch et al., 2021) and German scene segmentation (Zehe et al., 2021).

The rest of the paper is organized as follow; Section 2 presents recent research on text readability and complexity assessment and related tasks. An overview of the shared task and the data set, resources and the evaluation metrics that have been used in the shared task are presented in Sections 3 and 4, respectively. We briefly review the submitted models and discussed the results in Sections 5. Finally, we conclude the paper and the *German Text Complexity Assessment* shared task in Section 6.

2 Related Work

In this section we provide an overview of related shared tasks in different languages, and also highlight a number of the recent approaches for the task of text complexity assessment.

2.1 Shared Tasks

To the best of our knowledge, no shared task has been held so far on text complexity assessment at a sentence level. However, there are a few competitions on word level complexity assessment.

Paetzold and Specia organized the a shared task on complex word identification as a *SemEval* 2016 task (Paetzold and Specia, 2016). The task was to develop systems that can predict whether a target word is complex for a non-native English speaker, knowing the context sentence. In other words, it was a binary classification task in which 1 means the target word is complex in the given context sentence, and 0 means it's a simple word for a non-native English speaker.

The next complex word identification shared task was organized at the BEA workshop in 2018 for different languages including English, German, Spanish and French. The shared task included two subtasks: The first task was a binary classification of a target word in a context sentence as being complex or not complex. The second task was a probabilistic classification in which the participants were asked to assign the probability of a target word being considered complex (Yimam et al., 2018).

There were two subtasks of complexity prediction of single words and multi-word expressions as a regression task in the *SemEval* 2021 lexical complexity prediction task (Shardlow et al., 2021). The data includes around 10,000 instances for lexical complexity in which the target words were annotated on a five point Likert scale.

Russian simple sentence evaluation in 2021 is another related activity in which the task was developing systems to generate a simplified version of a given input complex sentence in Russian (Sakhovskiy et al., 2021). The proposed data set for the task includes around 3,000 complex sentences, each have 2.2 corresponding simplified sentences on an average.

As another related effort, Stajner et al. organized a shared task for the assessment of text simplification in which systems should automatically assign a label (e.g., good, OK, and bad) to four aspect of the pairs of original and simplified sentences (Stajner et al., 2016). The four aspects of interests include the quality of the generated sentences from grammar, meaning preservation, simplicity, and overall quality point of views.

2.2 Approaches

In this section we overview some of the recent approaches and models for automatic text complexity and readability assessment. We review the state-of-the-art models for English and German texts.

As one of the recent models for English text readability assessment, (Lee et al., 2021) developed different hybrid models using traditional machine learning approaches based on handcrafted features, and also transformer-based models. Based on their experiments, the combination of RoBERTA and Random Forrest models could outperforms the other models and achieved almost perfect classification accuracy (Lee et al., 2021). Hybrid models show promising results for the task in different languages and were the main trend among the submitted models for *GermEval* 2022.

Naderi et al. proposed a model for German text readability assessment based on linguistic features (Naderi et al., 2019b). They extracted traditional, lexical and morphological linguistic features (73 features in total). Their experiments show that again the Random Forest Regressor outperforms the other supervised models including SVM, Linear Regression, and Polynomial Regression models for the task (Naderi et al., 2019b).

In another study Weiss and Meurers proposed a model for sentence-wise German readability assessment for L2 readers (Weiss and Meurers, 2022). They compared different machine learning models in two different tasks for readability assessment; predictive regression and sentence pair ranking. The obtained results in their experiments show that a Bayesian Ridge Regression model achieved the best performance against the other models including the proposed model in (Naderi et al., 2019b) and also against the widely used readability formulae for the task of predictive regression. Moreover, regarding the document level text complexity assessment, their findings show that the readability of texts is driven by the maximum rather than the overall readability scores on the sentence level.

3 Task Description

In this section we describe the proposed task in detail. The data set and the evaluation metrics are presented in the next section.

The mapping of a body of text to a mathematical unit quantifying the degree of readability is the basis of readability assessment. This quantified unit is significant in informing the reader about how difficult the text content is to read. We defined the task of German text complexity assessment as a text regression task in which the participants were asked to develop systems to automatically assign a variable in the range from 1 to 7 to given German texts. We considered German learners at the B level as the target group. This means the system should predict the complexity/difficulty of a piece of text for a person who learns German at a B level.

The shared task is organized on the *Codalab* platform (Pavao et al., 2022), where the participants could access the data and submit their prediction on the provided data sets and get informed about the obtained results via the platform. More information about the competition is accessible via the corresponding web-page on the Codalab website [1].

Although the task is defined as a text regression task, there is no restriction on re-formulation of the task. Moreover, there was no restriction about using additional data sets for training purposes.

The shared task is organized in two phases; the development and the test phases. During the development phase the teams could develop their systems and test it against a validation data set. There was no restriction on the number of submissions during the development phase. The obtained results on the validation set were accessible for the teams immediately after submitting the predictions.

The test phase was a one week time period in which the participants could submit their results on the provided test data set. The test data was shared with the participants one week before the start of the test phase. During the test phase each team could submit a maximum number of two submissions per day on the test data set. The participants could only know about the achieved results on the test data (i.e., the leaderboard) when the competition ended. The detailed information about the provided data set and the evaluation metrics are presented in Section 4.

4 Data Set and Evaluation

In this section we discuss briefly the compiled data set for the competition and also overview the evaluation metrics that have been used to assess and ranked the submitted results.

4.1 Data Set

Three different data sets were available to the participants during the competition. We provided a training data set with complexity scores that could be used to train and tune the models and the systems. Moreover, two collections of sentences without the complexity score were shared as the validation and the test sets. The participants could evaluate their models using this data set during the development phase.

4.1.1 Train set

The training data set consisting of 1,000 German sentences taken from 23 Wikipedia articles. The data set includes subjective assessment of different text-complexity aspects provided by German learners at level A and B (Naderi et al., 2019a).

An online survey system was created to collect the subjective assessment of the 1,000 sentences using three items each rated on a 7-point Likert scale. A survey session consisted of training and rating sections. The training section was containing three sentences which participants needed to rate on the same scale as the main section. The sentences in the training section were constant and represent very easy, average and very complex sentences. Afterward, participants rated *complexity*, *understandability* and *lexical difficulty* of ten sentences. For each sentence in the data set the Mean Opinion Score (MOS) is calculated. The MOS score is the arithmetic mean over the all ratings of a particular aspect (complexity, understandability or lexical difficulty) provided for that sentence. The data set is published as *TextComplexityDE* in (Naderi et al., 2019a). For this shared task we only used the *complexity* scores of the sentences.

Figure 1 shows a few sample sentences from the training set. The training data set is freely available in a GitHub repository[2]. Moreover, a more detailed description of the *TextComplexityDE* data

[1] https://codalab.lisn.upsaclay.fr/competitions/4964

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

Als Nebenprodukt entstand damals natürlich auch die erste Seifenblase. *MOS complexity score: 1.60* Translation: As a by-product, of course, the first soap bubble was created at that time.
In Abgrenzung zum klassischen Rasiermesser wird ein Rasiermesser mit Wechselklinge als Shavette bezeichnet. *MOS complexity score: 3.25* Translation: In distinction from the classic razor, a razor with interchangeable blade is called a shavette.
In Pompeji gefundene Exemplare von frühen Klapp-Rasiermessern mit 12 Zentimeter langen trapezförmigen Klingen und Griffen aus Elfenbein gehörten als Luxusobjekte zum Hausstand höherer Schichten. *MOS complexity score: 4.36* Translation: Specimens of early folding razors with 12-centimeter-long trapezoidal blades and ivory handles found in Pompeii belonged to the household of higher classes as luxury objects.

Figure 1: Sample sentences from the training set

set including the conducted pilot study to determine relevant dimensions of text complexity and the manually simplified sentences are presented in (Naderi et al., 2019a).

4.1.2 Test set

The ratings for the validation and test data sets are collected in four different experiments. For each experiment, 100 sentences were complied in which 80 sentences were from 18 different Wikipedia articles, and 20 sentences were shared between all experiments and taken from the *TextComplexityDE* data set. Participants are recruited through online German learner groups in social media and also language schools. For online participants, there was a short mandatory listening and comprehensive language test to make sure they have basic to intermediate knowledge of German. We used a same 7-point Likert Scale as it was used in the training data set (*TextComplexityDE*). In the data cleansing step, all submissions from users with one of the following conditions were removed from the data.

- Users with wrong answer to the gold standard question[3]
- Users who failed in the language test
- Users with specific click patterns (i.e. small variance) or those who were too fast in finishing a session

[3]gold standard question contains a text that its complexity is known to organizers (i.e. very simple or very complex ones) and used to filter participants who not following the instructions.

Like the *TextComplexityDE* data set, a MOS score for complexity is calculated for each sentence. Using the 20 shared sentences in each experiment, a first-order mapping function for MOS values from each experiment to the MOS values of the *TextComplexityDE* data set are fitted. This was done to remove the well-known bias and gradient between different subjective tests.

The final data set includes 310 new sentences from 18 Wikipedia articles which were rated by a minimum of 16 participants. 100 sentences of this data have been used as the validation set. The participating teams used the validation set to tune their models and parameters during the development phase. The reminding 210 sentences were used as test data set to assess the performance of the submission in the test phase. All the reported results in this paper are the achieved results on the test data set with 210 instances.

Table 1 provides a summary of statistics and frequency distribution of the training and test data sets. Moreover, the histogram of MOS values in the training and test data sets are presented in Figure 2. As it is highlighted in the figure, the sentences in the training set tend to be more balanced. In other words, more complex and difficult sentences are presented in the training data set, compared to the test data set.

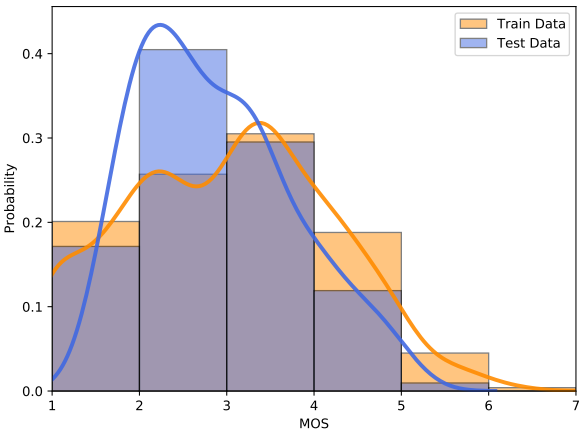

Figure 2: The distribution of MOS values in the training and test data sets

	Training data	Test data
Number of records (i.e., sentences)	1,000	210
Max length of sentences (in character)	487	486
Min length of sentences (in character)	19	38
Average length of sentences (in character)	147.3	160.03
Number of terms	20077	4400
Number of unique terms	7539	2249
Average of the complexity score	3.01	2.87
Standard Deviation	1.18	0.87

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

4.2 Evaluation Metrics

We used the Root Mean Square Error (RMSE) MAPPED metric to evaluate and rank the submitted results. Moreover, the normal RMSE scores were evaluated and reported.

RMSE shows the root of average squared difference between the estimated values $\widehat{y}_i$ (complexity scores) and the actual value y for the sentence i, as presented in the following equation.

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

Since slightly different ratings and consequently different MOS values could be obtained by repeating a subjective test and adding bias to the data, the RMSE MAPPED score has been used to assess the submitted runs. We used a mapping function to get ride of this offset/bias. The RMSE MAPPED is calculated based by the following steps:

1. A team submits its predictions (mos_pre).
2. A f(mos_pre) function is created by minimizing the absolute value between (true_mos) and f(mos_pre).
3. We call the outcome of the function f to be mapped_mos_pre:
   mappend_mos_pre = f(mos_pre)
4. We calculate the RMSE between the mappend_mos_pre and the true_mos.

The f function is created for each model, and is a linear function.

5 Results

In this section we present the baseline model and also survey the submitted models for the shared task.

5.1 Baseline Model

For the baseline model we fine-tuned a GBERT pre-trained model (Chan et al., 2020) on the training set. After feeding the input text into the model the last hidden state is passed through a dense linear layer by applying a *Tanh* activation. A dropout layer is also put on top before the output layer.

Regarding the hyper parameters, the *AdamW* optimizer (Loshchilov and Hutter, 2019) was used with a learning rate of $5e-5$. The model was fine-tuned in 3 epochs.

5.2 Proposed Models

In this section we highlight the main contributions of the proposed models in the shared task. The overall performance of the submitted results is presented in Table 2.

Among the submitted models, hybrid approaches in which the traditional machine learning models based on linguistic feature extraction are combined with state-of-the-art pre-trained language models show promising results for the task.

The top ranked team (Mosquera, 2022), HHUplexity team (Arps et al., 2022) and HIIG team (Asghari and Hewett, 2022) proposed hybrid models that combine a feature engineering approach and transfer learning via pre-trained transformers. Although the approaches are similar in general, different features and models have been used by different teams. For instance while Bert (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are fine-tuned in (Mosquera, 2022), the HHUplexity team extracted features from Bert and DistilBERT (Sanh et al., 2019) and the HIIG team fine-tuned XLM-R (Conneau et al., 2020). Moreover, different approaches have been used by different teams to combine the outcome of the feature engineering models and the pre-trained models. However, the hybrid models couldn't always

outperform the simple models. For instance, the obtained results from the HHUplexity team show that fine-tuning DistilBERT can outperform the other models including the hybrid model based on linguistic features. Also, the experiments from the HIIG team show that data augmentation could not increase the overall performance of the proposed model.

The AComplexity team (Blaneck et al., 2022), TUMuch Complexity team (Vladika et al., 2022) and TUM Social Computing team (Anschütz and Groh, 2022) used a similar approach of hybrid models. The AComplexity team extracted 154 features for each sentence and fine-tuned GBERT and GPT-2-Wechsel (Minixhofer et al., 2022) models. They combined the output of the pre-trained model with the readability features calculated for each sentence using a multi-layer perceptron with two layers (Blaneck et al., 2022). On the other side, the TUMuch Complexity team stacked RoBERTa and Gaussian process models as the proposed hybrid approach. As the stacking approach, they averaged the output predictions of the Gaussian process model and the fine-tuned XLM-RoBERTa (Conneau et al., 2020). The TUM Social Computing team (Anschütz and Groh, 2022) computed 6 different readability formulae based on some statistics and combined them with the fine-tuned DistilBERT model. Their analysis on the relevance of different features on the predictions highlight the importance of pre-trained models and also some statistics from text like the average sentence length (Anschütz and Groh, 2022).

The BBAW Zentrum Sprache team (Hamster, 2022) trained a random forest model on the set of extracted features like statistical, lexical, and grammatical ones. They also extracted a set of features from pre-trained NLP models like Sentence-BERT (Reimers and Gurevych, 2019). Their experiments show the linear relationship between the complexity score and the logarithm of the number of characters per sentence. Moreover, their results reveal that Sentence-BERT features also impact the complexity scores.

Due to the fact that the provided training data set was small and included only 1,000 sentences, different teams applied different strategies to increase the training data. The Deepset team used more than 220,000 pseudo-labels to train Transformer-based models in order to refrain from feature engineering step (Kostic´ et al., 2022). They used 12,562,164 distinct sentences from German Wikipedia and other corpora like news articles from Zeit Online for their semi-supervised learning approach. The proposed approach includes training a base model on the training set and pseudo-labeling the collected corpus with the base model. Finally, the pre-trained language models Fine-tuned on the pseudo-labels and the training sets and trained a linear regression model on the out-of-fold predictions from the cross-validations (Kostic´ et al., 2022).

As another approach to increase the data set size, the LGirrbach team turned the text regression task into a pairwise regression for complexity prediction (Girrbach, 2022). In this setting, instead of the direct prediction of the complexity score for the sentences, the model receive two sentences and predicts the relative difference in complexity of two sentences. However, the obtained results on the training set during the development phase show that "pairwise regression does not perform better than standard regression" (Girrbach, 2022). Unfortunately, the team could not test the proposed model on the test data set due to an error in the submission.

6 Conclusion

In this paper we described the *GermEval* 2022 task on "Complexity Assessment of German Text". The shared task is co-located with the Conference on Natural Language Processing (KONVENS) 2022. We presented the compiled data sets for the training and the test phases and the models proposed by the participants. The training and the test sets included 1,000 and 210 German sentences from Wikipedia articles, respectively, with a readability/complexity score from 1 to 7. Regarding the models, combining the traditional feature extraction models with state-of-the-art pre-trained language models was the main trend in the submitted systems. Although different teams used different feature set, pre-trained models and also different strategies to combine the outcomes of the models, there were similarities between the overall procedure from different participants. Almost all of the submissions could outperform the transfer learning based model as the competition's baseline.

For the next round of the shared task, the interpretability of the models (i.e., explainability) can be taken into account to make the predictions more

Team name	RMSE MAPPED	RMSE
Alejandro Mosquera (Mosquera, 2022)	**0.430**	0.449
AComplexity (Blaneck et al., 2022)	0.435	**0.442**
HIIG (Asghari and Hewett, 2022)	0.446	0.462
TUM Social Computing (Anschütz and Groh, 2022)	0.449	0.466
Deepset (Kostic´ et al., 2022)	0.454	0.484
TUMuch Complexity (Vladika et al., 2022)	0.457	0.489
HHUplexity (Arps et al., 2022)	0.473	0.486
Baseline	0.477	0.489
CCL	0.516	0.586
BBAW Zentrum Sprache (Hamster, 2022)	0.553	0.583
LGirrbach (Girrbach, 2022)	-	-

Table 2: The results on the test data set

understandable. Moreover, the training and the test sets can be enriched by more samples from more diverse resources.

Acknowledgments

We are grateful to the *KONVENS* 2022 conference organizers for their support on the shared task. We would also like to show our gratitude to the participants of *GermEval* 2022 whose participation and effort made *GermEval* 2022 a great success.

Finally we thank Kaspar Ensikat for his support for data preparation and Faraz Maschhur, Chuyang Wu, and Max Reinhard for developing the baseline model.

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

Miriam Anschütz and Georg Groh. 2022. TUM Social Computing at GermEval 2022: Towards the Significance of Text Statistics and Neural Embeddings in Text Complexity Prediction. In *Proceedings of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text*, Potsdam, Germany, September. Association for Computational Linguistics.

David Arps, Jan Kels, Florian Krämer, Yunus Renz, Regina Stodden, and Wiebke Petersen. 2022. HHUplexity at Text Complexity DE Challenge 2022. In *Proceedings of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text*, Potsdam, Germany, September. Association for Computational Linguistics.

Hadi Asghari and Freya Hewett. 2022. HIIG at GermEval 2022: Best of Both Worlds Ensemble for Automatic Text Complexity Assessment. In *Proceedings of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text*, Potsdam, Germany, September. Association for Computational Linguistics.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pado. 2014. Germeval 2014 named entity recognition shared task: Companion paper.

Patrick Gustav Blaneck, Tobias Bornheim, Niklas Grieger, and Stephan Bialonski. 2022. Automatic Readability Assessment of German Sentences with Transformer Ensembles. In *Proceedings of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text*, Potsdam, Germany, September. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6788–6796. International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to

readability formulas. *Discourse Processes*, 54(5-6):340–359.

Edgar Dale and Jeanne S Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Leander Girrbach. 2022. Text Complexity DE Challenge 2022 Submission Description: Pairwise Regression for Complexity Prediction. In *Proceedings of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text*, pages 45–50, Potsdam, Germany, September. Association for Computational Linguistics.

Ulf A. Hamster. 2022. Everybody likes short sentences - A Data Analysis for the Text Complexity DE Challenge 2022. In *Proceedings of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text*, Potsdam, Germany, September. Association for Computational Linguistics.

Bogdan Kostic´, Mathis Lucka, and Julian Risch. 2022. Pseudo-Labels Are All You Need. In *Proceedings of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text*, Potsdam, Germany, September. Association for Computational Linguistics.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic, November. 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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Matej Martinc, Senja Pollak, and Marko Robnik-Sikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Comput. Linguistics*, 47(1):141–179.

Tristan Miller, Darina Benikova, and Sallam Abualhaija. 2015. Germeval 2015: Lexsub–a shared task for german-language lexical substitution. *Proceedings of GermEval*, pages 1–9.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3992–4006. Association for Computational Linguistics.

Alejandro Mosquera. 2022. Tackling Data Drift with Adversarial Validation: An Application for German Text Complexity Estimation. In *Proceedings of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text*, Potsdam, Germany, September. Association for Computational Linguistics.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019a. Subjective assessment of text complexity: A dataset for german language. *CoRR*, abs/1904.07733.

Babak Naderi, Salar Mohtaj, Karan Karan, and Sebastian Möller. 2019b. 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*, pages 1–3. IEEE.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 560–569. The Association for Computer Linguistics.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. CodaLab Competitions: An open source platform to organize scientific challenges. Technical report, Université Paris-Saclay, FRA., April.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany, September. Association for Computational Linguistics.

Andrey Sergeyevich Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, E. Tutubalina, Valentin Malykh, Ivan Smurov, and E. Artemova. 2021. Rusimplesenteval-2021 shared task: Evaluating sentence simplification for russian.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. In Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurélie Herbelot, and Xiaodan Zhu, editors, *Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021*, pages 1–16. Association for Computational Linguistics.

Sanja Stajner, Maja Popovic, Horacio Saggion, Lucia Specia, and Mark Fishel. 2016. Shared task on quality assessment for text simplification.

Juraj Vladika, Stephen Meisenbacher, and Florian Matthes. 2022. TUM sebis at GermEval 2022: A Hybrid Model Leveraging Gaussian Processes and Fine-Tuned XLM-RoBERTa for German Text Complexity Analysis. In *Proceedings of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text*, Potsdam, Germany, September. Association for Computational Linguistics.

Zarah Weiss and Detmar Meurers. 2022. 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)*, pages 141–153, Seattle, Washington, July. Association for Computational Linguistics.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Stajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In Joel R. Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis, editors, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications@NAACL-HLT 2018, New Orleans, LA, USA, June 5, 2018*, pages 66–78. Association for Computational Linguistics.

Albin Zehe, Leonard Konle, Svenja Guhr, Lea Katharina Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, and Annekea Schreiber. 2021. Shared task on scene segmentation @ KONVENS 2021. In *Proceedings of the Shared Task on Scene Segmentation co-located with the 17th Conference on Natural Language Processing (KONVENS 2021), Düsseldorf, Germany, September 6th, 2021*, volume 3001 of *CEUR Workshop Proceedings*, pages 1–21. CEUR-WS.org.