HIIG at GermEval 2022: Best of Both Worlds Ensemble for Automatic Text Complexity Assessment

Hadi Asghari [*1]
[1]AI & Society Lab
Humboldt Institute for Internet and Society
Berlin, Germany
firstname.lastname@hiig.de

Freya Hewett [*1,2]
[2]Applied Computational Linguistics
University of Potsdam
Potsdam, Germany
firstname.lastname@hiig.de

Abstract

In this paper we explain HIIG's contribution to the shared task Text Complexity DE Challenge 2022. Our best-performing model for the task of automatically determining the complexity level of a German-language sentence is a combination of a transformer model and a classic feature-based model, which achieves a mapped root square mean error of 0.446 on the test data.

1 Introduction

Text complexity is not only a highly interesting topic from a linguistic perspective; it also has several implications on a societal level. A text that has the appropriate complexity level for a specific reader not only ensures that the reader can fully understand the information presented in the text, but it also keeps the reader engaged and can help the reader to learn new structures and expand their vocabulary. This last point is particularly relevant for language learners and readers who are reading text in a language that is not their native language. The Text Complexity DE Challenge focuses on this specific target group as the task involves predicting the complexity of a sentence in German, which have been annotated on a scale of 1 to 7 by German learners whose language proficiency is at B level (on the CEFR scale). An overview of the shared task and the results from all the teams can be found in (Mohtaj et al., 2022).

In this paper we briefly report on related work in Section 2, before describing the dataset used in the shared task in Section 3. In Section 4 we outline our various approaches to the task, before reporting on the results and briefly discussing them in Section 5. In Section 6 we conclude the paper.

2 Related work

Previous work aimed at automatically assessing the text complexity level of sentences has focused mostly on the English language. Stajner et al. (2017) use the Newsela corpus (English-language newspaper articles, simplified at multiple levels for different aged school children) and calculate scores for unigrams, bigrams and trigrams by looking at what levels of the corpus they occur in. They experiment with different classifiers and achieve the best results with a Random Forest. Pitler and Nenkova (2008) conduct a small-scale analysis on 30 articles from the Wall Street Journal which have been manually annotated on a scale from 1-5 for the question of how well-written the article is. They investigate how various linguistic features correlate with these scores. Vocabulary and discourse relations are the strongest predictors of readability, followed by average number of verb phrases and length of the text. Lee et al. (2021) work with three English-language datasets and produce hybrid models which consist of a transformer based model combined with a feature-based model. They predict 3 and 5 classes (depending on the dataset) and achieve the state of the art, with a ROBERTA-based transformer model performing best.

Work on German-language text complexity assessment is fairly rare. Hancke et al. (2012) look at text-level binary readability classification using a corpus of 1627 articles in original form and a version aimed at children. Their classifier uses the Sequential Minimal Optimization algorithm with five groups of features (traditional readability formulas, lexical, syntactic, language model, and morphological), with a best accuracy score of 89.7%. Stodden and Kallmeyer (2020) work with various corpora from different languages from the text simplification domain and evaluate 104 different features using statistical tests, with the aim to determine differences between simplified and complex texts. They also work with a German-language corpus of 1888 texts (Klaper et al., 2013) and find that the feature lexical complexity, in particular, is relevant specifically for German texts. Battisti

* Equal contribution

et al. (2020) build on the same corpus and release a newer version with 6217 documents. Hewett and Stede (2021) create a corpus of 2655 texts from online lexica at three different levels (adults, children, children who are beginner readers) and use knowledge graph based features to estimate conceptual complexity. In a pairwise classification task they achieve an accuracy score of 91%.

3 Dataset

The dataset for the challenge consists of sentences that have been taken from 41 Wikipedia articles from different article genres. Groups of German learners, with language levels between A2 and B2, rated the sentences according to complexity, understandability and lexical difficulty on a scale from 1 to 7. For each aspect, the arithmetic mean (or Mean Opinion Score; MOS) was calculated and the task was to predict the MOS complexity score of the sentences. More information on the dataset can be found in (Naderi et al., 2019).

The training dataset consists of 1000 sentences, the validation set (for development phase) of 100 and the test set (for the evaluation phase) of 210 sentences. Figure 1 shows a histogram of the target variable (MOS) in the training set (mean=3.02, stdev=1.18). Some examples from the training set can be seen in Table 1.

It is also worth mentioning that 'complexity' can be subjective. For example as can be seen in Table 1, the second sentence '*Das Meerwasser ist leicht basisch*' has a score of 1; whilst the sentence is clearly short and has a very simple structure, arguably the words alkaline (*basisch*) and even seawater (*Meerwasser*) are not usually part of a language learner's vocabulary. The sentence structure may not be 'complex' but the lexical items do seem more advanced. These kinds of scores may be due to the fact that participants were also asked to rate the understandability of a sentence, a score which was not used in this shared task. The subjective nature of complexity is a limitation of the dataset which the shared task organisers try to compensate for by using a mapped root mean squared error as a metric, more information can be found in the task overview paper (Mohtaj et al., 2022).

4 Approaches

In this section we outline our different approaches. As a baseline, we take the simple approach of predicting the mean MOS value (3.02) for all samples. Using this baseline, the root mean squared error (RMSE) is 1.18.[1]

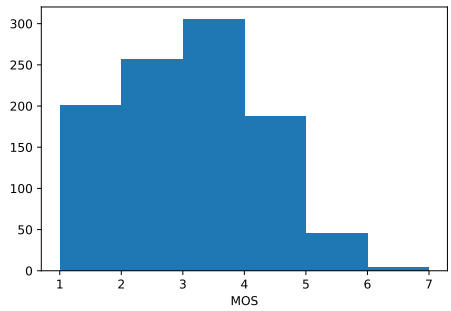

Figure 1: Histogram of mean opinion scores.

4.1 Additional Augmented Data

As the training dataset is not particularly large (1000 sentences), one of our approaches was to create additional training data. We used the extended lexica corpus from Hewett and Stede (2021) which consists of 86613 sentences at three different levels. We created artificial scores on a scale of 1 to 7 using a simple method. We took the original labels (1-3) and scaled them using the feature of sentence length, which we found to be a strong predictor for complexity. For example, a sentence with the original label of 1, with one of the longest sentence lengths for this class would have a transformed score of around 2.3, whereas a short sentence originally labelled with 1 would have a transformed score closer to 1.

We used this additional data together with the training data in several different models and the results were consistently worse than the basic baseline using only the training data. This is most likely due to the noise that is introduced when producing these artificial labels, and the fine-grained nature of the labels. Another reason could also be the different target groups; the shared task data has been labelled by non-native speakers whereas the lexica corpus has children as its target group. Children and non-native speakers are two different target groups of simplified language with different needs.

4.2 Neural Approach

The neural method we use is to fine-tune a pre-trained transformer model for our given task. As

[1] Aside from the final results in Table 2, all reported scores refer to 20% of the training set and to the non-mapped RMSE.

Original Sentence	Literal Translation	MOS Complexity
Die Folgen dieser Versauerung betreffen zunächst kalkskelettbildende Lebewesen, deren Fähigkeit, sich Schutzhüllen bzw Innenskelette zu bilden, bei sinkendem pH-Wert nachlässt.	The consequences of this acidification have an effect on calcium-skeleton-forming organisms, whose ability to form protective shells or internal skeletons diminishes with decreasing pH.	5.25
Das Meerwasser ist leicht basisch.	Seawater is slightly alkaline.	1

Table 1: Example sentences from the dataset.

our base model, we chose XLM-R (also known as XLM-RoBERTa) by Conneau et al. (2019).

XLM-R is a self-supervised cross-lingual transformer model – trained on 2.5TB of filtered CommonCrawl data containing 100 languages – using a masked language modeling objective. It is mostly intended to be fine-tuned on downstream tasks (HuggingFace, 2022), and offers state-of-the-art performance for many language tasks. Specifically it outperforms multilingual BERT on a variety of metrics (Conneau et al., 2019).

The fact that XLM-R has great performance out of the box and is multilingual, make it a suitable choice for the challenge. We downloaded the pretrained model using the Hugging Face Python library.[2] We changed the model head to a (single) regressor layer plus a dropout, inspired by Kozodoi (2022). (As is typical, the weights for the new layers are randomly assigned, while the rest of the model is initialized to the pretrained weights.) We used a custom trainer to set RMSE as the loss function, and did not freeze any of the layers for higher accuracy. For preprocessing, was used the XLM-R Tokenizer with padding and truncation, which is how this model expects the data.

During the earlier phases of the Text Complexity DE challenge, we used a simple 80:20 data split for training and validation; and observed that our modified XLM-R model performed quite well after 10 training epochs with the default AdamW optimizer. For the final stage of the challenge, we adopted k-fold validation (with k=5) to ensure that all the available data was used during training. Thus we ended up with five models (with RMSEs between 0.55 and 0.70). For the actual predictions on the test dataset, we averaged the prediction of these five models.

4.3 Feature-based Approach

A further approach was to use the 43 'single features' which Stodden and Kallmeyer (2020) applied in their cross-lingual study on text complexity (see Section 2). These features are calculated using sentences as input; we therefore did not perform any additional pre-processing. We applied feature ranking using the recursive feature elimination implementation from scikit-learn (Pedregosa et al., 2011) and used the top 34 features; the full list of features can be found in Appendix A. The most important features were number of words per sentence, number of syllables per sentence and number of characters per word. We used these with a linear regression model, using the default parameters from scikit-learn. When applied to the training data in a 80/20 split, the RMSE was 0.7. We then retrained the model in the whole training set before using the official test data set as input (the results of which can be seen in Table 2). Approaches using sentence embeddings or lexical complexity values derived from our additional data did not beat our simple baseline on the training set and so were therefore not pursued any further.

5 Ensemble Results & Discussion

Our final approach is to combine our feature-based and neural approach by averaging the outputs of these two models.[3]

While both our transformer and feature-based models perform better than the baseline RMSE, among them, the transformer model does generally better on different data splits. Thus, it might seem paradoxical that our final model is a weighted average (ensemble) of the two. Using an ensemble method is, however, a theoretically sound practice, and quite common in machine learning competitions.

In the words of Page (2018), "To rely on a single model is hubris. It invites disaster. [...] Wisdom can be achieved by averaging models." Simply explained, due to both over-fitting and under-fitting, any one model will predict some samples (especially among the unobserved) quite wrongly. As long as the individual models in the collection do

[2]We chose the base model, not large, so that the training could be done efficiently on our laptop GPU.

[3]Our implementation can be found at https://github.com/hadiasghari/konvens22-shared-task

not share a common bias, then any diverse collection of the models will be more accurate than the average member—an implication of the so called diversity prediction theorem (Page, 2018).

To illustrate the point, we can compare the predictions from both models on the test dataset (Figure 2). The Pearson correlation coefficient between the two is 0.85. On average, the predictions are close, with the transformer model predicting slightly lower scores. In about ten percent of the samples, the difference between the predictions is bigger than 1, and crucially, in both directions.[4]

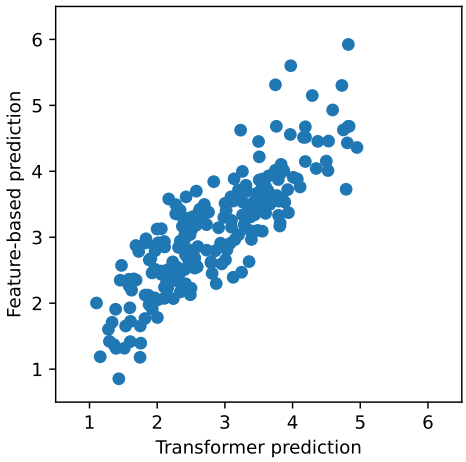

Figure 2: Histogram of mean opinion scores.

Closer to home and in the readability assessment (RA) literature, Lee et al. (2021) also propose using ensemble methods. In particular "[when] a transformer shows weak performance on small datasets, there must be some additional measures done to supply the final model (e.g. ensemble) with more linguistic information", adding that such studies are rare in RA.

The final results on the test dataset are presented in Table 2. We hypothesized that since our transformer model does slightly better than our feature based model, a weighted average favoring the former might yield better accuracy, which turned out to be the case.[5]

[4]Without the MOS scores for the test dataset, we can only speculate about this discrepancy between the two model predictions. See Appendix B for a few examples.

[5]In future work, the averaging weights could themselves be learnt from the data, and obviously, more models be added to the ensemble.

RMSE	Model Description
0.541	Linear regression (feature based)
0.484	Ensemble 70:30 (lr:xlmr)
0.479	XLM-R (without k-fold)
0.458	XLM-R (with k-fold)
0.457	Ensemble 50:50
0.450	Ensemble 40:60
0.446	**Ensemble 30:70** (lr:xlmr)

Table 2: Mapped RMSE results for different models (more information on the mapping can be found in the shared task overview paper (Mohtaj et al., 2022))

6 Conclusion

In this paper we explained our contribution to the shared task Text Complexity DE Challenge 2022. We experimented with both neural and feature-based approaches. Our best-performing model is a weighted average of a fine-tuned XLM-R transformer model and a classic feature-based model with linear regression. The ensemble achieves a mapped root square mean error of 0.446 on the test data which is better than either of the models alone.

Acknowledgements

Thank you for the organisers of the shared task for providing the data and relevant information.

References

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A Corpus for Automatic Readability Assessment and Text Simplification of German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France. European Language Resources Association. ArXiv: 1909.09067.

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.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability Classification for German using Lexical, Syntactic, and Morphological Features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

Freya Hewett and Manfred Stede. 2021. Automatically evaluating the conceptual complexity of German texts. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*,

pages 228–234, Düsseldorf, Germany. KONVENS 2021 Organizers.

HuggingFace. Xlm-roberta (base-sized model) model card [online]. 2022.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/simple German parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria. Association for Computational Linguistics.

Nikita Kozodoi. Estimating text readability with transformers [online]. 2022.

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. Association for Computational Linguistics.

Salar Mohtaj, Babak Naderi, and Sebastian Möller. 2022. Overview of the GermEval 2022 shared task on text complexity assessment of german text. In *Proceedings of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text*, Potsdam, Germany. Association for Computational Linguistics.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for german language.

Scott E. Page. 2018. *The Model Thinker: What You Need to Know to Make Data Work for You*, 1st edition edition. Basic Books, New York.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Sanja Stajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic Assessment of Absolute Sentence Complexity. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4096–4102, Melbourne, Australia. International Joint Conferences on Artificial Intelligence Organization.

Regina Stodden and Laura Kallmeyer. 2020. A multilingual and cross-domain analysis of features for text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 77–84. European Language Resources Association.

Appendix

A Features Used

The features we used in our feature-based model (discussed in Section 4.3) include[6]:
get type token ratio, get ratio of function words, get ratio of coordinating clauses, get ratio of subordinate clauses, get ratio prepositional phrases, get ratio relative phrases, get ratio clauses, get ratio named entities, check if head is noun, check if one child of root is subject, check passive voice, is non projective, get ratio of nouns, get ratio of verbs, get ratio of adjectives, get ratio of adpositions, get ratio of adverbs, get ratio of auxiliary verbs, get ratio of conjunctions, get ratio of determiners, get ratio of numerals, get ratio of particles, get ratio of pronouns, get ratio of punctuation, count words, count sentences, count syllables in sentence, count words per sentence, count syllables per sentence, count characters per word, count syllables per word, max pos in freq table, average pos in freq table, sentence fkgl.

B Discrepancy between Model Predictions

Without the MOS scores for the test dataset, we can only speculate about this discrepancy between the two model predictions. After manually inspecting some cases, we found that when the prediction of the feature-based model was higher (i.e. more complex) than the transformer model, these were long sentences which in fact were often just lists. When the prediction of the transformer model was higher, these were often shorter sentences with uncommon words (often compounds). See Table 3 for some examples.

[6]From the implementation from Stodden and Kallmeyer (2020): https://github.com/rstodden/text-simplification-evaluation

ID	Sentence	Translation	XLMR	LR
2115	"Die danach häufigsten Wohnungstypen waren Wohnungen in kleinen Apartmentkomplexen (2–9 Einheiten, 12,8 % der Bevölkerung), Wohnungen in mittleren Apartmentkomplexen (10–49 Einheiten, 7,9 %), Einfamilienreihenhäuser (5,9 %), Mobilheime (5,7 %), Wohnungen in großen Apartmentkomplexen (50+ Einheiten, 5,0 %) und Boote, Wohnmobile und Ähnliches (0,1 %)."	The next most common housing types were flats in small apartment complexes (2-9 units, 12.8 % of the population), flats in medium apartment complexes (10-49 units, 7.9 %), single-family terraced houses (5.9 %), mobile homes (5.7 %), flats in large apartment complexes (50+ units, 5.0 %), and boats, mobile homes, and the like (0.1 %).	3.233	4.624
2053	"Daneben gibt es auch konfessionelle (VkdL im CGB) und weitere Verbände (Waldorflehrkräfte, Lehrkräfte der Montessori-Schulen)."	There are also confessional (VkdL (Association of Catholic German Teachers) in the CGB (Christian Trade Union Federation of Germany)) and other associations (Waldorf teachers, Montessori school teachers).	3.123	2.393

Table 3: Example predictions on the test set with large discrepancy between the transformer (XLMR) and the feature based linear regression (LR) models.