HHUplexity at Text Complexity DE Challenge 2022

David Arps, Jan Kels, Florian Krämer, Yunus Renz, Regina Stodden, Wiebke Petersen
Heinrich Heine Universität
Düsseldorf, Germany
first.last@hhu.de
all authors contributed equally

Abstract

In this paper, we describe our submission to the 'Text Complexity DE Challenge 2022' shared task on predicting the complexity of German sentences. We compare performance of different feature-based regression architectures and transformer language models. Our best candidate is a fine-tuned German Distilbert model that ignores linguistic features of the sentences. Our model ranks 7th place in the shared task.[1]

1 Introduction

Texts are a basic form of human information exchange. Too high of a text complexity, however, can result in text comprehension failures (Bormuth, 1966) and therefore miscommunication. Text complexity and readability assessment are a long known problem and several computational approaches and metrics have been proposed (Dascalu, 2012; Hancke et al., 2012; Collins-Thompson, 2014), relying on different linguistic features and primarily aiming at English.

Among other application objectives, an adequate, quantificational metric for text complexity can be of high benefit to the educational domain (as a means of providing textual material according to student levels), writing support systems (as feedback) or for other natural language processing tasks like estimating the complexity of the output of text simplification systems or chatbots. Most related tasks focus either on the prediction of complex words (Paetzold and Specia, 2016; Shardlow et al., 2021) or the assessment of readability levels (Collins-Thompson, 2014). However, the goal of the 'Text Complexity DE 2022' shared task is the prediction of an empirically determined complexity score called 'Mean Opinion Score' (MOS) for German sentences. Overall, our best model is ranked on the 7th place out of 10. In the following, we present the approach and results of our team "HHUplexity" in more detail.

[1]Our code is available at https://github.com/Vipitis/HHUplexity

1.1 Shared Task Data

The training data (Naderi et al., 2019) for the shared task contains 1000 sentences from 25 Wikipedia texts. The development data and test data contain 100 and 210 sentences, respectively, for which the document distribution is not known. The sentences were rated by German language learners (between CEFR level A and B) on a 7 point Likert-scale regarding their complexity (1 – very easy to 7 – very complex). The arithmetic mean of these ratings is the target score – MOS score – of the shared task. 7.6% of the training samples are rated as very easy (score = 1), whereas 20.3% are rated as rather complex (score > 4) and 3.4% have a score higher than 5.

The root mean squared error (RMSE) after third order mapping as well as a more balanced RMSE score (RMSE_{mapped}) are used to evaluate the predicted MOS scores (Mohtaj et al., 2022).

2 Method

Our main approach is to combine hand-crafted features with text embeddings of language models. Therefore, we have calculated several features as described in subsection 2.1. To compare the effect of these features in combination with language models, we follow two baseline approaches: i) training different regression models with the features (see subsection 2.2), and ii) fine-tuning language models without features (see subsection 2.3). Afterwards, we combine the features with the language models in a multimodal model (see subsection 2.4).

2.1 Features

We calculate 349 features of seven main categories: features based on length, readability assessment features, features based on language proficiency, morphological features, syntactic features, morphosyn-

	Feature
Length-based	number of words [♠]
	number of types
	number of characters [♠]
	number of syllables [♠]
	avg. word length in characters
	max. word length in characters
	avg. word length in syllables
	number of sentences
Readability	Flesch Reading Ease Score [♥]
	Flesch-Kincaid Grade Level [♥]
	Dale-Chall Readability Score
	Linsear Write Formula
	Automated Readability Index
	difficult words
Morphological	ratio of negations & negated words
	ratio of compounded words & nouns
	number of nominalizations
	N-gram frequencies
	ratio of nouns in cases
Morphosyntactic	number of verbs, auxiliaries, nouns, pronouns
	ratio of coarse-grained POS-tags [◆] [♣]
	ratio of fine-grained POS-tags (STTS) [◆] [♣]
	noun-to-verb ratio
	number of stop words [◀]
	ratio of function words [◀]
	ratio of named entities

	Feature
Syntactic	max. depth of the dependency parse tree [◀]
	max. & avg. distance between tokens in the parse tree
	max. & avg. distance between verbs and verb particles in the parse tree
	avg. length of NP & VP & PP [◆]
	± projective parse tree [▶]
	± head of the parse tree is a noun or verb [▶]
	± one child of the head of the parse tree is a subject [▶]
	± passive voice [◆]
	± subjunctive mood [◆]
	ratio of multi-word expressions [▶]
	number of clauses
	ratio of all tokens of coordinating & subordinating clauses [◆]
	ratio of tokens marking relative clauses [◆]
	ratio of tokens marking prepositional phrases [◆]
	ratio of tokens marking referential phrases [▶]
Lexical	ratio of words that are in the vocabulary lists for CEFR levels A1, A2, & B1
	type-token ratio [◀]
	avg. lemma frequency & rank (based on deCOW)
	lexical complexity based on ranks of German FastText embeddings [♥]
	max. and avg. rank in the German FastText embeddings [♥]
Other	perplexity score (based on GerPT2)
	label of target group and their softmax scores predicted by a fine-tuned model on this labeling task
	cosine similarity between original sentence and backtranslated sentences into German from English, Turkish, Hungarian, Chinese, and Georgian
	avg. imagebility and concreteness score [◀]

Table 1: Overview table of all features per category. The symbols stand for the papers in which the features were introduced: [♠] Scarton et al. (2018), [♥] Martin et al. (2018), [♣] Kauchak et al. (2014), [◆] Gasperin et al. (2009), [◀] Collins-Thompson (2014), [▶] Stodden and Kallmeyer (2020).

tactic features, and other features. An overview of all features is provided in Table 1. In general we find that 78% of the features have a significant Pearson correlation with the MOS target value (p-value > 0.05). Of those 66% have a weak correlation ($|r| < .4$), 21% a moderate correlation ($.4 \leq |r| < .6$) and 12% a strong correlation ($.6 \leq |r|$).

However, several features are absolute count features such as e.g. syllables or character count that depend on sentence length. If one transforms these features into proportional features the rate of features having a significant Pearson correlation with the MOS target value drops to 57%.

Features based on Length. As basic features to estimate the complexity of a sentence, we consider the length of the sentence (in words, syllables and characters) and the length of the words (in syllables and characters).

Readability Assessment Features. The length of words and sentences can also be jointly used to estimate text complexity within traditional readability formulas for texts, e.g., Flesch Reading Ease score or Flesch-Kincaid Grade Level.[2] The established readability metrics have been calculated for the original German sentences as well as for automatically translated English sentences (altogether 24 features). It turns out that the German scores correlate better than or equally well as the English scores with the exception of the Dale-Chall Readability Score. While Dale-Chall shows no significant correlation with the MOS-values for the German sentences it correlates with $r = 0.392$ for the English sentences.

It turns out that these quite simple formulas lead to the features with the strongest MOS-correlations. Only four significant features have a Pearson correlation r-value above 0.7 of which three are established readability scores: Linsear Write Formula with $r = 0.745$, difficult words with $r = 0.741$, Automated Readability Index (ARI) with $r = 0.706$, and number of words $r = 0.701$.

Lexical Features and Features based on Language Proficiency. Even if a word is short, it can be still unknown to a user and, therefore, difficult to understand. In our work, we include some

[2] We use several readability metrics of the textstat package (https://pypi.org/project/textstat/).

lexical and language proficiency-based features to estimate the complexity of a sentence based on the choice of words. Simple words are often frequent and complex words more infrequent, so word frequency might help to estimate the complexity of a sentence (Martin et al., 2018; Collins-Thompson, 2014). We follow two approaches, first, we obtain the frequency and rank per lemma based on the deCOW-corpus (Bildhauer and Schäfer, 2014) and build the average of them per sentence. Second, we measure the lexical complexity based on the word ranks in the German FastText Embeddings as well as obtaining the highest and average position of the tokens in the sentence.

Additionally, we select vocabulary lists per CEFR level A1, A2, B1 by the Goethe institute[3] and measure the ratio of words in the input sentence that can be found in the CEFR vocabulary lists. Vocabulary lists for other CEFR levels have not been available. The correlations with the empirical MOS-values indicate that the study participants judging the complexity are familiar with the vocabulary up to the B1 level. All three correlations (ratio of A1 / A2 / B1 vocabulary words) are negative and lie in the range $-0.35 \leq r \leq -0.4$. That is the higher the proportion of A1/A2/B1 vocabulary words, the less complex the participants judged the sentence.

Morphological Features. Besides the length and the choice of the words, a morphological analysis of words can be helpful to assess the complexity of the sentences. For example, some morphemes can drastically change the meaning of a word, e.g., negation prefixes ("irr-" or "un-"), nominalization suffixes ("-heit" or "-keit"), or one-token compound nouns ("Staubecken", "Dampfschiff"). Therefore, we calculate the number of nominalizations, negations based on a fix list of affixes, count the number of n-grams[4], as well as the ratio of compounded words[5]. Furthermore, we include the ratio of nouns per case, as the genitive is often difficult to understand. The non-ngram morphological features exhibit a significant but weak correlation with the MOS-values.

Syntactic Features. Besides an analysis of the words, an analysis of the structure of a sentence can give additional insights into its complexity because some syntactic structures take longer to process and comprehend (Gibson, 1998). To reflect syntactic complexity in our features, we measure the maximum depth of the dependency parse tree, maximum and average distances between words and number of clauses.[6] Based on Gasperin et al. (2009), we add the average length of noun, verb, and prepositional phrases, and whether the sentence is written in active or passive voice and indicative or subjunctive mood. Furthermore, we check some regularities in the parse tree based on Stodden and Kallmeyer (2020) (see Table 1). Based on the parse tree, we also count the ratio of multi-word expressions and ratio of all tokens of some clauses (see Table 1). Maximum tree depth has the strongest correlation ($r = 0.583$) with the MOS-values. Tree width features like average NP or VP length show a moderate positive correlation as well ($r \approx 0.4$). A negative correlation is found for the number of clauses normalized by sentence length ($r = -0.46$).

Morphosyntactic Features. Part-of-speech (POS) tags combine some morphological information with syntactic information, therefore we use the number and ratio of coarse-grained / fine-grained POS tags and noun-to-verb ratio to estimate the sentence complexity as similar as in Gasperin et al. (2009) and Kauchak et al. (2014). Following Collins-Thompson (2014), we also include the ratio of function words and stop words to all tokens as a feature. However, none of these features has a moderate or strong correlation with the MOS-score.

Psycholinguistic Features. In readability literature, psycholinguistic-based features are often named as relevant features (Collins-Thompson, 2014; Davoodi and Kosseim, 2016). In our work, we obtain the imageability and concreteness of each word per sentence based on the Concreteness and imageability lexicon MEGA.HR-Crossling (Ljubešić, 2018) and measure the average per sentence as another feature. Both of these features do not show a moderate or strong correlation which might be due to the absence of the words of the sentences in the chosen resources.

Perplexity Feature. We calculate the perplexity of the sentence with "GerPT-2"[7]. The higher the perplexity score, the harder to predict the seen sen-

[3] https://www.goethe.de/de/spr/kup/prf/prf.html

[4] We consider all uni-grams in the training data (char vocab), and top k n-grams for k=20, n=2, 3, 4, 5.

[5] The compounded words are obtained by https://github.com/repodiac/german_compound_splitter.

[6] The number of clauses is heuristically derived by the clause-splitting method described in Dönicke (2020).

[7] https://huggingface.co/benjamin/gerpt2

tence and the more unlikely is the input sentence for the model. Hence, we hypothesize, the higher the perplexity score, the more uncommon/complex is the sentence. For the training data the hypothesis can be confirmed but only by a weak correlation ($r = 0.214$).

Translation-based Features. The idea is to test whether translation difficulties indicate higher MOS-values. Therefore, with GoogleTranslator the sentences have been translated into English, Turkish, Hungarian, Chinese, and Georgian and backtranslated into German. These languages vary in their morphological and syntactic similarity and in their degree of genetic relationship to German. For the original and the backtranslated sentences contextualized embedding vectors have been determined with a transformer language model[8]. Finally, the cosine similarity for the sentence pairs has been calculated and added as a feature. It turns out that only Georgian leads to a non significant feature (p-value = 0.08), all others are significantly correlating with the MOS-values indeed only weakly. The highest correlation is found for Chinese Simplified with $r = 0.146$ and $p = 0.00$.

Text Level. We fine-tune a 3-class text level classifier on the Lexica corpus (Hewett and Stede, 2021), a dataset with German Wikipedia texts for three different target groups: younger children, children and adults. From this dataset we sample roughly 38k sentences (taken from roughly 1650 different texts), and fine-tune a German BERT model[9] to predict one of the three labels: child, youth, adult. The fine-tuned language model is applied to the shared task dataset and the softmax scores for the three labels, as well as the predicted labels are used as additional text level features. All four text level features have a rather high moderate correlation with the MOS values: softmax adult $r = 0.589$, softmax youth $r = -0.447$, softmax child $r = -0.519$, and predicted label $r = 0.565$. The correlations show that the study participants judge sentences as less complex if they have a higher probability of being labeled as 'child' or 'youth' and as more complex the higher the probability of the label 'adult' is. This indicates that the German language proficiency level of the participants is in between the youth and the adult level.

[8] https://huggingface.co/Sahajtomar/German-semantic
[9] https://huggingface.co/deepset/gbert-base

2.2 Predicting MOS from features

We have compared different methods to predict MOS based on the features from the previous section. To choose an appropriate model architecture and hyperparameters, we train and test models on a 5-fold crossvalidation split of the shared task training data for which MOS scores are available. We compare linear regression models with different regularization (Ridge, ElasticNet), and XGBoost (Chen and Guestrin, 2016). Because XGBoost achieved the best crossvalidation performance by a margin of $>$ .05 RMSE compared to the other models, we only report results for this model. Using the same 5-fold crossvalidation split, the best hyperparameters are determined. The best model is an XGBoostRegressor[10] with `n_estimators`=2500, a learning rate of `eta`=.005, and a `max_depth` of 5. This model achieves a RMSE of .545 (RMSE mapped .502) on the final test data.

2.3 Fine-tuning

We have explored fine-tuning a language model directly on the regression task using Huggingface's *AutoModelForSequenceClassification* for various models available on Huggingface including English, German and multilingual versions of BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019). To select the best pre-trained model, we have trained on 900 sentences of the training data and evaluated RMSE on the remaining 100. With a learning rate of $2 * 10^{-5}$ and 5 epochs. Trading smaller batch sizes for more steps led to better results where 10 did better than 30 or 50.

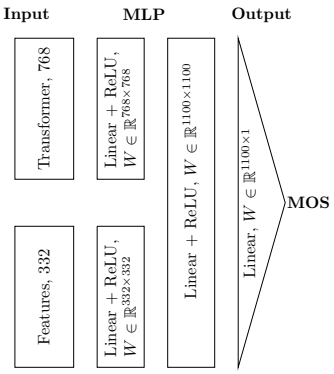

Figure 1: Architecture of the multimodal model that combines BERT embeddings with feature vectors

[10] https://docs.getml.com/1.1.0/api/getml.predictors.XGBoostRegressor.html

2.4 Multimodal model

To combine text embeddings and numerical features, we have written a custom version of Huggingface's *DistilbertForSequenceClassification* heavily inspired by Multimodal-Toolkit (Gu and Budhkar, 2021). BERT embeddings and text features are combined by a feedforward neural network, the architecture of which is displayed in Figure 1.

3 Results

	RMSE
XGBoost no ngrams	**.545**
XGBoost all feats	.639
ElasticNet no ngrams	.672
ElasticNet all feats	.659
Ridge no ngrams	.669
Ridge all feats	.713
bert-base-cased	.601
bert-base-german-cased	.552
bert-base-german-dbmdz-cased	.638
bert-base-multilingual-cased	.565
distilbert-base-cased	.600
distilbert-base-german-cased	**.486**
xlm-roberta-base	.511
distilbert-base-german-cased multimodal	**.622**

Table 2: Results for all models. Boldface results indicate performance on test data via a submission to the evaluation system. In all other cases, the performance is measured on a randomly selected held-out split of the training data. The first line separates regression models and fine-tuned language models, and the second line separates the multimodal model.

Results are presented in Table 2. For feature-based predictors, features and target MOS scores are transformed by removing the mean and scaling to unit variance. We find that gradient-boosting methods (XGBoost) work significantly better than linear models with ElasticNet or Ridge regularization. As shown in Table 2, the ablation of n-gram features clearly drops the RMSE score for XGBoost and Ridge regularization (> 0.04). XGBoost without n-gram features achieves a RMSE_{mapped} score of .502 on the test data. For fine-tuned models, distilbert-base-german-cased was trained on 990 sentences with batch size 10 and 5 epochs. The submitted result reached .486 RMSE (.473 RMSE_{mapped}) on test data. When combining a transformer language model and features (subsection 2.4), we found that our implementation did not manage to improve results over the fine-tuning baseline. The best submission for this method reached .622 RMSE (.524 RMSE_{mapped}). A smaller feature set based on their importance might improve the results, similar as shown for the ablation of n-gram features with XGBoost and Ridge regularization (see Table 2).

3.1 Distribution of predictions

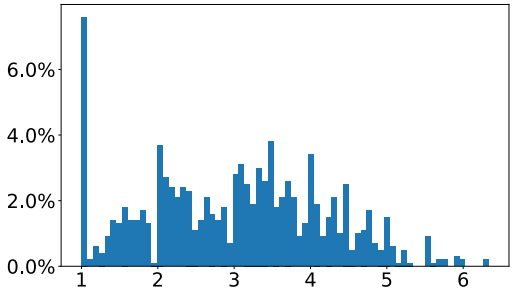

Figure 2: Distribution of MOS scores in the training data (70 bins).

On crossvalidated results, we find that many of our models do not predict the Gaussian distribution of MOS scores with an additional peak at the low end (Figure 2). All models correctly identify the mean scores of the general dataset, but generally tend to predict MOS scores of a lower standard deviation. Across feature-based models, the standard deviation of predicted scores on validation data is approximately 20% smaller than the standard deviation of the gold labels. We do not know the true labels of the validation and testing shared task data, but assume that this systematic error is also present in our submissions for these datasets.

4 Conclusion

In our contribution to the shared task, we have compared predictions based on linguistic features with an approach based on transfer learning, i.e., fine-tuning a language model. We find that even though linguistic features achieve relatively high correlation with the MOS scores, they are outperformed by a "simple" fine-tuned transformer language model.

References

Felix Bildhauer and Roland Schäfer. 2014. Decow14 lemma frequency list.

John R. Bormuth. 1966. Readability: A new approach. *Reading Research Quarterly*, 1(3):79–132.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of*

the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 785–794, New York, NY, USA. ACM.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

Mihai Dascalu. 2012. Analyzing discourse and text complexity for learning and collaborating: A cognitive approach based on natural language processing.

Elnaz Davoodi and Leila Kosseim. 2016. CLaC at SemEval-2016 task 11: Exploring linguistic and psycho-linguistic features for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 982–985, San Diego, California. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Tillmann Dönicke. 2020. Clause-level tense, mood, voice and modality tagging for German. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 1–17, Düsseldorf, Germany. Association for Computational Linguistics.

Caroline Gasperin, Lucia Specia, Tiago F. Pereira, and Ra M. Aluisio. 2009. Learning when to simplify sentences for natural text simplification. In *In Proceedings of ENIA*, pages 809–818.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Ken Gu and Akshay Budhkar. 2021. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, Mexico City, Mexico. Association for Computational Linguistics.

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.

David Kauchak, Obay Mouradi, Christopher Pentoney, and Gondy Leroy. 2014. Text simplification tools: Using machine learning to discover features that identify difficult text. *2014 47th Hawaii International Conference on System Sciences*, pages 2616–2625.

Nikola Ljubešić. 2018. Concreteness and imageability lexicon MEGA.HR-crossling. Slovenian language resource repository CLARIN.SI.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 29–38, Tilburg, the Netherlands. 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.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

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

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Text simplification from professionally produced corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

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, Marseille, France. European Language Resources Association.