TUM Social Computing at GermEval 2022: Towards the Significance of Text Statistics and Neural Embeddings in Text Complexity Prediction

Miriam Anschütz
Technical University of Munich
Department of Informatics
Germany
miriam.anschuetz@tum.de

Georg Groh
Technical University of Munich
Department of Informatics
Germany
grohg@in.tum.de

Abstract

In this paper, we describe our submission to the GermEval 2022 Shared Task on Text Complexity Assessment of German Text. It addresses the problem of predicting the complexity of German sentences on a continuous scale. While many related works still rely on handcrafted statistical features, neural networks have emerged as state-of-the-art in other natural language processing tasks. Therefore, we investigate how both can complement each other and which features are most relevant for text complexity prediction in German. We propose a fine-tuned German DistilBERT model enriched with statistical text features that achieved fourth place in the shared task with a RMSE of 0.481 on the competition's test data.

1 Introduction

Text readability describes how easy a given text is understood by a specific reader (Hancke et al., 2012). Factors that influence the readability are, for example, the number of technical terms in the text or the length and convolution of the sentences. Assessing a text's readability can be used to select the proper texts for a specific user group or provide authors feedback about their texts. Moreover, it can be integrated into an automatic text simplification system. On the one hand, it helps to decide whether and, if so, how much a text should be simplified. On the other hand, readability assessment is a measure to evaluate a simplification system by checking if the output has a higher readability (Garbacea et al., 2021; Martinc et al., 2021). Text complexity is inversely related to text readability; thus, in this work, the terms text complexity prediction and readability assessment are used interchangeably.

This paper is a contribution to the GermEval 2022 Shared Task on Text Complexity Assessment of German Text that aims to predict the complexity of a German text on a continuous scale (Mohtaj et al., 2022). We propose a model based on fine-tuned German DistilBERT (Sanh et al., 2019) combined with traditional readability formulas and statistical text features. This model achieved fourth place in the competition. Moreover, we used SHAP (Lundberg and Lee, 2017) to explain our model's predictions and discuss which features contribute to higher complexity. By knowing the feature relevance, authors and machine learning engineers can pay attention to them when generating new texts. Our code is released on Github for further research and development.[1]

This paper is structured as follows: Section 2 gives an overview of existing readability formulas and prediction models. In section 3, we present the organization of the shared task and introduce its dataset. Then, section 4 walks through our proposed approaches and entails their performance. Finally, in section 5, we apply explainability methods to discuss text features relevant to complexity prediction.

2 Related work

We investigated two approaches for readability assessment, traditional readability formulas, and deep learning. Therefore, this section gives an overview of existing formulas and models. Moreover, we analyze which text features yielded promising prediction results in previous work.

2.1 Traditional complexity measures

Multiple formulas exist to calculate the readability of a text based on statistical values such as word counts or average word length. Flesch (1948) proposed the Flesh reading ease (FRE) score that calculates a value between $0 - 100$, where a higher value indicates a lower complexity. Similarly, the readability index (LIX) (Björnsson, 1983) returns a readability estimate ranging from 20 to 60. However, with this score, an easier text gets a lower

[1] https://github.com/MiriUll/text_complexity

value. As German words tend to be longer than English words on average, Amstad (1978) adapted the FRE measure to the German language by adapting the weight of the average word length measure. Kincaid et al. (1975) used the FRE score as a basis for a new measure, the Flesch-Kincaid-Grade-Level (FKGL). In contrast to the previous scores, this returns the U.S. school grade in which the text can be understood. Other complexity scores returning the number of years in education needed to grasp the content of a text are SMOG (Laughlin, 1969) and Gunning fog index (Gunning et al., 1952). The Wiener Sachtext formulas are four slightly varying formulas returning the required grade adapted to the German school system and specificities of the German language (Bamberger and Vanacek, 1984).

These formulas are based on an analysis of textual features. In the literature, different text properties are distinguished (Santucci et al., 2020; vor der Brück et al., 2008; Hancke et al., 2012): Statistical features analyze the number of sentences or the number of words in a sentence, while syntactic features investigate the sentence structure, e.g., the depth of the dependency tree. Other categories are lexical features, such as the number of unique words, or semantic features, i.e., the length of causal chains. As indicated by Solnyshkina et al. (2017), using the plain text properties as features can outperform the complexity estimation of readability formulas.

2.2 Learning complexity prediction models

Syntactic, semantic, or lexical text features have been exploited for readability prediction in different languages such as Italian (Santucci et al., 2020) or English (Štajner and Hulpus, 2020). Other approaches use neural language models like BERT for their predictions (Martinc et al., 2021). For the German language, Weiß and Meurers (2018) proposed a binary prediction model based on linguistic features, such as lexical or morphological complexity, and psycholinguistic features, i.e., cognitive complexity and language use. Their work was based on the binary prediction model by Hancke et al. (2012). In a very recent work (Anonymous, 2021), the neural approaches by Martinc et al. (2021) were transferred to German, yielding promising results in a language-level prediction task. These approaches focus on a classification task, while vor der Brück et al. (2008) worked on a seven-point Likert scale, similar to the Shared Task data. They used syntactic and semantic features together with a nearest neighbor model for their predictions.

2.3 Feature relevance analysis

To understand why a model deems a sentence complex, but also to use the complexity scores for further tasks such as text simplification (Garbacea et al., 2021), the features that contributed to the predictions are of interest. Santucci et al. (2020) used the Gini measure and permutation importance to inspect which text property was important for their predictions. They reported that the most relevant features were the syntactic and morphosyntactic ones. Similarly, Hancke et al. (2012) discovered the essential features for their classification were the average word length or the number of complex nominals in the sentences.

3 Shared task and Dataset

This paper explains our submission to the GermEval 2022 Shared Task on Text Complexity Assessment of German Text (Mohtaj et al., 2022). The shared task was split into two different phases, a development and a final phase. During development, participants were provided a labeled training and an unlabeled validation dataset. Predictions on this validation data could be uploaded to the competition page with immediate evaluation feedback. In contrast, during the final phase, the results on the final test dataset were only published at the end of the competition. The two evaluation datasets, the validation and the final test data, consist of 100 sentences each. The training dataset for this shared task originates in work by Naderi et al. (2019). It contains 1000 sentences from the German Wikipedia together with a complexity score ranging from 1 to 7. Naderi et al. (2019) used crowdsourcing to let non-native speakers of a B level annotate the respective sentences by their perceived readability and averaged the scores among the participants. The mean complexity value is 3.016 with a standard deviation of 1.181. There are 76 sentences with an observed complexity of 1.0, but only two samples with a complexity higher than six, making the dataset unbalanced towards the easier sentences. To counteract this imbalance, we replicated sentences with a complexity higher than 5.5 multiple times, yielding a dataset with 1054 samples.

The rooted mean squared error (RMSE) between

predicted and correct complexity scores was used to evaluate a model's performance. In addition, a third-order polynomial function was fitted between the predicted and correct scores to counteract the bias by subjective annotation of text complexity. Then, the predicted scores were projected using this function, and the error was calculated on the mapped predictions as well (Mohtaj et al., 2022).

4 Approaches

In this section, we explain the three approaches we explored to predict the complexity score of a sentence. We did not apply any preprocessing to the data, i.e., fed the sentences into the model's tokenizer directly.

4.1 Learning from text statistics

We analyzed different textual features and readability scores calculated based on them. Table 1 shows which statistics were calculated. On the one hand, statistics on a sentence level were investigated, such as the average sentence length or the maximal depth of the dependency tree. We assumed that a more complex sentence holds subclauses or multi-word expressions that show in a high dependency tree depth. For our data, the average sentence length is similar to the number of words in a sentence, as our data samples contain only one sentence. On the other hand, we examined the characteristics of the words in a sentence, e.g., the average number of syllables among all words. Moreover, the percentage of words consisting of only one syllable was calculated. These are very short and easy-to-understand words, i.e., a high percentage can indicate a simple sentence.

Feature	Description
asl	Average sentence length
mtd	Maximal dependency tree depth
pw6	Percentage of words with at least six letters
asc	Average number of syllables
ps1	Percentage of words with only one syllable
ps3	Percentage of words with at least three syllables

Table 1: Statistical features calculated from sentences.

These statistics are part of different readability formulas. Equations 1 to 6 show the formulas for the scores used in this work. We propose calculating the Flesh reading easy (FRE) by Amstad (Amstad, 1978), the four Wiener Sachtext formulas (Bamberger and Vanacek, 1984) and the SMOG score (Laughlin, 1969). The FRE formula uses the average sentence length and the average number of syllables among all words and returns a value between 0 and 100, where a higher score indicates better readability. The Wiener Sachtext formulas are a collection of four formulas that slightly vary the statistics they use and their weights. The formulas calculate for which school grade between four and 15 the text is suited. Similarly, the SMOG score returns how many years of education the reader needs to understand the text. Thus, a lower value is desirable for the Wiener Sachtext formulas and the SMOG score. In contrast to the other formulas, the SMOG score only uses the number of words with at least three syllables (ns3) as a statistical measure.

$$\textbf{fre\_amstad} = 180 - \text{asl} - (58.5 \cdot \text{asc}) \quad (1)$$

$$\textbf{wstf1} = 0.1935 \cdot \text{ps3} + 0.1672 \cdot \text{asl} + 0.1297 \cdot \text{pw6} - 0.875 - 0.0327 \cdot \text{ps1} \quad (2)$$

$$\textbf{wstf2} = 0.2007 \cdot \text{ps3} + 0.1682 \cdot \text{asl} + 0.1373 \cdot \text{pw6} - 2.779 \quad (3)$$

$$\textbf{wstf3} = 0.2963 \cdot \text{ps3} + 0.1905 \cdot \text{asl} - 1.1144 \quad (4)$$

$$\textbf{wstf4} = 0.2744 \cdot \text{ps3} + 0.2656 \cdot \text{asl} - 1.6930 \quad (5)$$

$$\textbf{SMOG} = 1.0430 \cdot \sqrt{\text{ns3}} + 3.1291 \quad (6)$$

We computed the statistics in Table 1 and scores in Equations 1 to 6 for all samples in our data. Then, we fitted a support vector regression based on these statistical vectors as a prediction baseline. For this, we used the implementation by sklearn and its default hyperparameters parameters.[2] The model achieved a RMSE of 0.657 and mapped RMSE of 0.647 on the training data.

4.2 Fine-tuning a transformer model

To investigate the complexity prediction quality of neural networks, we fine-tuned a German DistilBERT model. We utilized Huggingface (Wolf et al., 2020) to load and fine-tune the distilbert-base-german-cased (von Platen, 2020) model. We

[2] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

trained the model on the shared tasks' training data with the default setup of Huggingface's trainer API[3] for two epochs. Table 2 shows the promising results achieved by this model on the training, validation, and final test data. The model outperformed the statistics-only SVR baseline model by far.

Dataset	RMSE	RMSE_mapped
training	0.402	0.399
validation	0.405	0.404
final test	0.481	0.460

Table 2: Complexity prediction results by fine-tuned DistilBERT model.

4.3 Combining DistilBERT embedding with textual features

The pure text statistics model and the fine-tuned DistilBERT model yielded promising results. To take advantage of both their handcrafted features and deep textual understanding, we combined both models. We used the last hidden state of the DistilBERT model as an embedding of size 768. Then, we concatenated the embedding with the vector of statistical measures and readability scores. Finally, we trained a support vector regression model on these representations with the same setup as the statistical SVR. Table 3 highlights the performance on the three different datasets. With this model, we achieved fourth place in both the competition's development and final evaluation phase.

Dataset	RMSE	RMSE_mapped
training	0.404	0.403
validation	0.395	0.390
final test	0.466	0.449

Table 3: Complexity prediction results by SVR with DistilBERT embedding and statistical features.

5 Explaining the predictions

To evaluate which of the suggested statistics and formulas help to predict the complexity of German texts, we calculated the SHapley Additive exPlanations (SHAP) values (Lundberg and Lee, 2017) for each of our models. SHAP measures each feature's contribution by masking their different combinations and rerunning the predictions with these masks. Features for which the masked predictions deviate strongly from the initial prediction have a substantial impact and are, thus, the most relevant ones. The SHAP values are calculated per sample and averaged among them. Figure 1 shows the

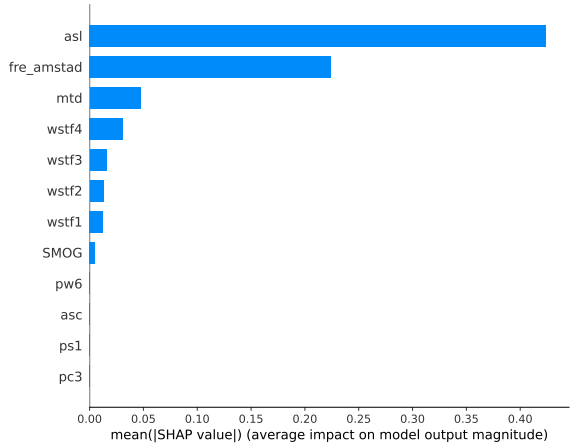

Figure 1: SHAP values for statistical text features in our support vector regression model, sorted in descending order.

mean SHAP values for each feature in the statistical SVR model (Section 4.1). The most relevant statistic is the average sentence length, i.e., the longer a sentence is, the more likely it is complex. The FRE score uses this statistic; thus, it is reasonable that it has high importance. Even though the Wiener Sachtext formulas also include this statistic, their contribution to the predicted score is smaller. They incorporate more advanced measures like the percentage of words with more than three syllables. As indicated by the small SHAP values, these additional statistics are not helping our complexity prediction model. The third most relevant feature is the maximum tree depth, indicating how convoluted a sentence is.

For a neural network, it is unknown what functionality a specific neuron models. Therefore, a feature-relevance analysis is not beneficial for interpreting a neural network. Instead, we selected the example sentence "Dieser Vorgang wird Gletscherschwund oder Gletscherschmelze genannt." (*"This process is called glacier recession or glacier melt."*) and investigated which words have an impact on the prediction. The correct complexity for this sentence is 2.266667, and our model (Section 4.2) predicts a complexity of 2.373029. Figure 2

[3] https://huggingface.co/docs/transformers/main_classes/trainer#transformers.TrainingArguments

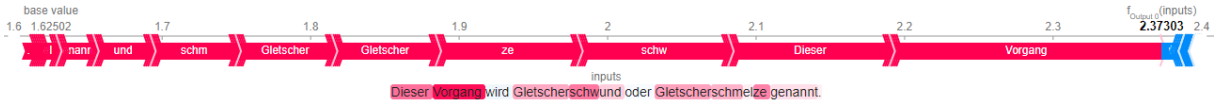

Figure 2: DistilBERT prediction on an example sentence (English translation: "This process is called glacier recession or glacier melt."): contribution of each word and word chunk to the prediction result.

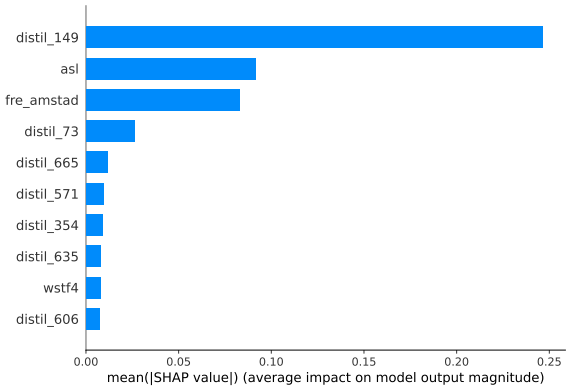

Figure 3: Feature relevance analysis for combined model: the ten features with highest SHAP values. "distil_i" indicates the ith index in the DistilBERT embedding.

shows which words and parts of words increase or decrease the predicted score compared to a base value. Words like "wird" (*"is"*) and "oder" (*"or"*) have a negative contribution, i.e., they indicate an easier sentence. Contrary, the word "Vorgang" (*"process"*) has the highest positive impact. The word itself is not very difficult, but it is often used to describe complex procedures and, thus, can be seen as a signal for a complex sentence. In German, compound nouns such as "Gletscherschwund" (*"glacier recession"*) are very common. However, the DistilBERT tokenizer splits them into multiple tokens. Therefore, different parts of these compound words have different contributions to the prediction, making it harder to identify their overall contribution.

Finally, Figure 3 depicts another feature relevance analysis, but for the SVR model that combined our neural embedding with statistical text features (Section 4.3). The scores were calculated on a subset of the data, and we only highlight the values for the ten highest ranking features. The strongest impact on the prediction comes from the embedding value at index 149, but text statistics like the average sentence length and Amstad's FRE score are also relevant. This implies that both learned neural features and traditional text statistics impact text complexity prediction. Moreover, they complement each other to yield the most accurate predictions. Therefore, we have shown that neural models have not yet outperformed handcrafted features regarding German text complexity prediction.

6 Discussion

Readability is a subjective measure that depends on the reader's background knowledge and reading ability (Crossley et al., 2017). Our work is based on the shared task's dataset labeled with a crowdsourcing approach among non-native speakers. Therefore, the findings in this paper should be tested for transferability to other datasets and groups of readers. In addition, the dataset is unbalanced with an overrepresentation of simple sentences and contains some noise. For example, the sentence "Martin Luther King Jr (* 15 Januar 1929 in Atlanta als Michael King Jr; † 4 April 1968 in Memphis) war ein US-amerikanischer Baptistenpastor und Bürgerrechtler." (*"Martin Luther King Jr (born January 15, 1929 in Atlanta as Michael King Jr; † April 4, 1968 in Memphis) was a U.S. Baptist pastor and civil rights activist."*) has a complexity of 1.0, indicating it was a very easy sentence. This shows that some samples have lower complexity than they would have when relabeling the dataset.

7 Conclusion

In this paper, we have demonstrated three approaches for text complexity prediction in German, one model that relies on handcrafted statistical features only, one fine-tuned transformer network, and a combination of both. In addition, we found that the feature most indicative of a complex sentence is the sentence length and that the FRE formula by Amstad (1978) gives a good indication of text complexity. Modern transformer architectures with deep textual understanding can build accurate complexity prediction pipelines. However, they can still be improved with handcrafted statistical features, showing that they have not yet superseded traditional approaches. In future work, these findings will be extended to a paragraph and full-text level instead of a sentence-wise prediction.

References

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis, Universität Zürich.

Anonymous. 2021. Language level classification on german texts using a neural approach. ACL ARR 2021 November Blind Submission.

Richard Bamberger and Erich Vanacek. 1984. *Lesen-Verstehen-Lernen-Schreiben. Die Schwierigkeitestufen von Texten in deutscher Sprache*. Diesterweg.

C. H. Björnsson. 1983. Readability of newspapers in 11 languages. *Reading Research Quarterly*, 18(4):480–497.

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.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Cristina Garbacea, Mengtian Guo, Samuel Carton, and Qiaozhu Mei. 2021. Explainable prediction of text complexity: The missing preliminaries for text simplification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1086–1097, Online. Association for Computational Linguistics.

Robert Gunning et al. 1952. Technique of clear writing.

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.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

G. Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of Reading*, 12(8):639–646.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179.

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.

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

Valentino Santucci, Filippo Santarelli, Luciana Forti, and Stefania Spina. 2020. Automatic classification of text complexity. *Applied Sciences*, 10(20).

Marina Solnyshkina, Radif Zamaletdinov, Ludmila Gorodetskaya, and Azat Gabitov. 2017. Evaluating text complexity and flesch-kincaid grade level. *Journal of Social Studies Education Research*, 8(3):238 – 248.

Sanja Štajner and Ioana Hulpus. 2020. When shallow is good enough: Automatic assessment of conceptual text complexity using shallow semantic features. In *LREC 2020 Marseille : Twelfth International Conference on Language Resources and Evaluation : May 11-16, 2020, Palais du Pharo, Marseille, France : conference proceedings*, pages 1414–1422, Paris. European Language Resources Association, ELRA-ELDA.

Patrick von Platen. 2020. distilbert-base-german-cased.

Tim vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep indicators. *Informatica (Slovenia)*, 32(4):429–435.

Zarah Weiß and Detmar Meurers. 2018. Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 303–317, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.