Ablesbarkeitsmesser: A System for Assessing the Readability of German Text

Florian Pickelmann[1], Michael Färber[2], and Adam Jatowt[1]

[1] University of Innsbruck
florian.picklemann@student.uibk.ac.at, adam.jatowt@uibk.ac.at
[2] Karlsruhe Institute of Technology (KIT)
michael.faerber@kit.edu

Abstract. While several approaches have been proposed for estimating the readability of English texts, there is much less work for other languages. In this paper, we present an online service, available at `https://readability-check.org/`, that provides five well-established statistical methods and two machine learning models for measuring the readability of texts in German. For the machine learning methods, we train two BERT models. To bring all the measures together, we provide an interactive website that allows users to evaluate the readability of German texts at the sentence level. Our research can be useful for anyone who wants to know whether the text content at hand is easy or difficult and therefore can be used in certain situations or rather needs to be adapted and improved. In education, for example, it can help to assess the suitability of a particular teaching material for a particular grade.

Keywords: Readability · German language · online service.

1 Introduction

When we communicate information in text form, how do we ensure that the recipient can understand it? For example, an explanation of how to use a product might be written in a language that is too difficult for average users. What might help mitigate this problem is to assess the readability of the text before using it. However, manual assessment of readability is difficult, especially for long texts, and requires a certain level of expertise. Moreover, a written text typically consists of many complex relations or hidden assumptions as for the knowledge of the reader that make a sentence or paragraph difficult or easy to read and understand. Hence, dedicated and sometimes complex measures need to be used to properly estimate the reading level of texts. However, most studies on readability have focused on English, while fewer studies have focused on other languages such as German, French, or Polish [5].

In this paper, we present an online service, available at `https://readability-check.org/`, that allows to assess the readability of German texts based on various metrics. The goal is to provide a platform where you can paste a text and then easily assess how readable it is and which parts of the text cause readability problems.

This would help people who need to publish or hand over written work that needs to be easily readable, such as legal documents, texts on difficult topics, or texts for non-native speakers. While we focus on German, similar solutions can be used to estimate readability in other languages. Our work demonstrates a feasible model that also has an explanation-oriented function in a sense that the difficulty of a document can be explained by the set of its difficult sentences indicated in the output.

Related Work. Automatically assessing text by means of readability metrics started in the early 1900s when readability formulas based on statistics (e.g., word frequency, word length, sentence length) were proposed. However, these measures have been criticized for their weak statistical bases and inability to capture more complex aspects of a language. With the development of state-of-the-art natural language processing methods, attempts where made to improve the measures created. The next leap in readability assessment was made when machine learning (ML) became mainstream [6,10]. The ML models were generally able to produce better results than the traditional formulas, but require more effort, such as creating data sets and training the model until they are ready to use. Specifically, for the German language, methods that use traditional language models [8,3] or which use semantic networks in comparison to simple surface-level indicators to calculate text readability [4] were created. However, to our knowledge there is no research work describing the process of building of a working tool for assessment of text readability, especially for German.

Statistical Readability Measures. Traditionally, readability of texts was measured by statistical readability formulas, which try to construct a simple human-comprehensible formula with a good correlation to what humans perceive as the degree of readability. The simplest of them is the average sentence length (ASL); others they take into account various other statistical factors, such as word length and word frequency. Most of these formulas were originally developed for English but could be also applicable to other languages with some modifications [10]. For our system we have chosen four formulas, the first three are among the most popular ones used [7] and the fourth one was designed for the German language [2].

(1) The *Gunning fog index* (GFI) [12] calculates the years of education a person needs to understand a text on its first reading. It is calculated with the following expression:

$$GFI = (\frac{numberOfWords}{numberOfSentences} + 100 * \frac{numberOfPolysyllables}{numberOfWords}) * 0.4,$$

where $numberOfPolysyllables$ indicates the number of words with three or more syllables.

(2) The *Flesch reading ease* (FRE) [9] indicates the United States (US) grade level a reader needs in order to be able to understand the text. It is calculated

in the following way:

$$FRE = 206.835 - (1.015 * ASL) - (84,6 * ASW),$$

where ASL is the average sentence length and ASW is average number of syllables per word. For the use in German language, Amstad [1] proposed the formula

$$FRE_{deutsch} = 180 - ASL - (58.5 * ASW).$$

(3) The *Simple Measure of Gobbledygook* (SMOG grade) [11] is a readability formula originally used for checking health messages. It roughly corresponds to the years of education needed to understand the text. It is calculated as follows:

$$SMOG = 1.0430\sqrt{numberOfPolysyllables * \frac{30}{totalSentances}} + 3.1291,$$

(4) The *Wiener Sachtextformel* (WSTF) [2] calculates readability in terms of Austrian/German grade levels. It ranges from four to fifteen with four being easy and 15 being very hard.

$$WSTF = 0.1935 * MS + 0.1672 * SL + 0.1297 * IW - 0.0327 * ES - 0.875.$$

MS is the percentage of words with more than three syllables, SL represents the mean of the number of words, IW denotes the percentage of words with more than six letters and ES is the percentage of words with one syllable.

2 Datasets

We use two datasets for training our machine learning components.

TextcomplexityDE Dataset.[3] This dataset consists of 1,000 sentences in German taken from 23 Wikipedia articles. It was created to be used for generating models for text-complexity prediction and text simplification. Sentences were rated in three scales (complexity, understandability, and lexical difficulty) by 267 non-native German learners.

Deutsche Welle Dataset. This dataset was created to assist in a Deep Text Evaluation Project[4]. It was created by using different articles of the online magazine "Deutsche Welle" (DW) to create a classification for the reading level zero for B1 and one for B2 and C1. After splitting all the sentences and classifying them to the corresponding readability classes of their documents, we obtained 17,395 sentences (249,206 words) with readability class 0, being easy, and 43,814 sentences (685,577 words) with readability class 1, being hard.

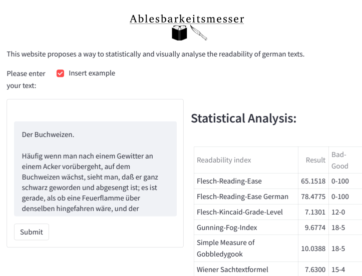
Training BERT models. We used the pre-trained BERT model $BERT_{BASE}$ and fine-tuned models for both the datasets described above. A user can select either of the models to be used as well as obtain readability scores by the statistical readability metrics discussed in Sec. 1.

[3] https://github.com/babaknaderi/TextComplexityDE
[4] https://github.com/shlomihod/deep-text-eval

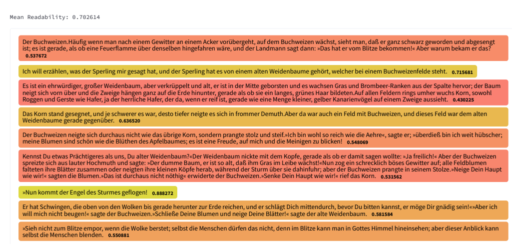
3 Demonstration System

All the implementation code together with the processed datasets can be found on our GitLab[5]. The service can be accessed online[6] and the screenshots of the website are shown in Fig. 1a and 1b. A short video presentation of the website can be also found online.[7] To create a platform where everyone can check input text for readability we used the python package `streamlit`.[8] To let users input text we chose a simple textbox with a submit button. Having a text input, we then incorporated the classic measures into the website, showing their results as an output table.

We used the BERT models to highlight sentences with different complexities. To check different complexities, we decided to implement a slider that defines the lower threshold for displaying the predictions. A user is able to choose with the slider the minimal value that should be highlighted. Furthermore, the model output was normalized to be from zero to one using a min-max normalization. For highlighting a range of values we decided to color-code them from green (zero) to red (one). A range of colours was created using the `Color` objects from the `colours` package. Note that since the model trained on Deutsche Welle displays just two distinct values (0 or 1), we vary the color depending on the confidence of the prediction.



(a) Screenshot of the website displaying the text input field and the output table of the classic measures (explained in Sec. 1).



(b) Screenshot of the text highlighting using the BERT model trained on TextComplexityDE.

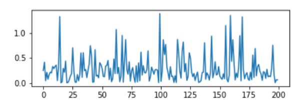
4 Evaluation

We calculated the mean square error (MSE) and the mean absolute error (MAE) of our models with 5-fold cross-validation. The average results over all the five

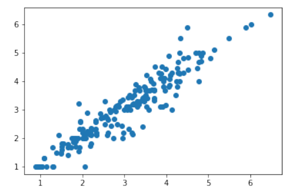
[5] https://git.uibk.ac.at/csaw3616/readability-detection
[6] https://readability-check.org/
[7] https://youtu.be/jtYJH2XxLl4
[8] https://streamlit.io/

Table 1: MSE and MAE for the TextComplexityDE and the Deutsche Welle models.

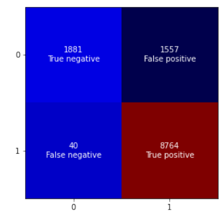
TextComplexityDE		Deutsche Welle	
MSE	**MAE**	**MSE**	**MAE**
0.172	0.413	0.197	0.435



(a) Absolute error on a single selected fold (200 sentences). Sentences are arranged on axis x.



(b) Correlation (c) of the predicted values to the test samples: c = 0.945



(c) True positive, false positive, true negative and false negative

Fig. 2: Evaluation of the TextComplexityDE models test set ((a), (b)) and the confusion matrix of the classification using the Deutsche Welle model (c).

folds are shown in Table 1 for the TextComplexityDE model and Deutsche Welle model. Figure 2a depicts also the error of TextComplexityDE model. As can be seen, there are few outliers with high absolute error values. Furthermore we also measure the Pearson correlation coefficient of the predicted values and the ground truth scores of the test samples, as shown in Figure 2b. The predicted values appear to be quite correlated (c = 0.945), which further proves the reliability of the model. Finally for Deutsche Welle model, as seen in Figure 2c, 10,645 of 12,242 samples where correctly predicted, leading to an accuracy of 86.95%. The number of false positives is way bigger than the false negatives. This was to be expected, because the dataset in general is positively biased.

5 Conclusions

While the readability of English texts has been relatively well researched, much less work has been done on the assessment of the readability of texts in other languages. Furthermore, there is no work describing ready-made, workable systems for automatically estimating the readability of texts. The goal of this work is to develop a system that can estimate the readability of German texts using two different methods: classical statistical approaches and machine learning approaches. The proposed system not only allows the estimation of the readability of an entire document, but also indicates which parts might cause difficulties for the reader, thus providing an explainable result that can assist the user in text correction. For the future, we plan to estimate the readability of text at a sub-sentence level (clauses or multi-word expressions).

References

1. Amstad, T.: Wie verständlich sind unsere Zeitungen? Abhandlung: Philosophische Fakultät I. Zürich. 1977, Studenten-Schreib-Service (1978), https://books.google.at/books?id=kiI7vwEACAAJ
2. Bamberger, R., Vanecek, E.: Lesen-Verstehen-Lernen-Schreiben: die Schwierigkeitsstufen von Texten in deutscher Sprache. Jugend und Volk (1984), https://books.google.at/books?id=TElTAAAACAAJ
3. Blaneck, P.G., Bornheim, T., Grieger, N., Bialonski, S.: Automatic readability assessment of German sentences with transformer ensembles. In: Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text. pp. 57–62. Potsdam, Germany (Sep 2022)
4. vor der Brück, T., Hartrumpf, S.: A semantically oriented readability checker for german (01 2007)
5. Collins-Thompson, K.: Computational assessment of text readability: A survey of current and future research. ITL-International Journal of Applied Linguistics **165**(2), 97–135 (2014)
6. Crossley, S.A., Skalicky, S., Dascalu, M., McNamara, D.S., Kyle, K.: Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. Discourse Processes **54**(5-6), 340–359 (2017)
7. Dubay, W.: The principles of readability. CA **92627949**, 631–3309 (01 2004)
8. Hancke, J., Vajjala, S., Meurers, D.: Readability classification for German using lexical, syntactic, and morphological features. In: Proceedings of COLING 2012. pp. 1063–1080. The COLING 2012 Organizing Committee, Mumbai, India (Dec 2012), https://aclanthology.org/C12-1065
9. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., Naval Technical Training Command Millington TN Research Branch (1975)
10. Martinc, M., Pollak, S., Robnik-Šikonja, M.: Supervised and unsupervised neural approaches to text readability. Computational Linguistics **47**(1), 141–179 (2021)
11. Mc Laughlin, G.H.: Smog grading-a new readability formula. Journal of reading **12**(8), 639–646 (1969)
12. Robert, G.: The Technique of Clear Writing. McGraw-Hill; Revised edition (1968)