Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2177–2187
Marseille, 20-25 June 2022

Argument Similarity Assessment in German for Intelligent Tutoring: Crowdsourced Dataset and First Experiments

Xiaoyu Bai, Manfred Stede
Applied Computational Linguistics, University of Potsdam
Karl-Liebknecht-Straße 24-25, 14476, Potsdam, Germany
{xiaoyu.bai, stede}@uni-potsdam.de

Abstract

NLP technologies such as text similarity assessment, question answering and text classification are increasingly being used to develop intelligent educational applications. The long-term goal of our work is an intelligent tutoring system for German secondary schools, which will support students in a school exercise that requires them to identify arguments in an argumentative source text. The present paper presents our work on a central subtask, viz. the automatic assessment of similarity between a pair of argumentative text snippets in German. In the designated use case, students write out key arguments from a given source text; the tutoring system then evaluates them against a target reference, assessing the similarity level between student work and the reference. We collect a dataset for our similarity assessment task through crowdsourcing as authentic German student data are scarce; we label the collected text pairs with similarity scores on a 5-point scale and run first experiments on the task. We see that a model based on BERT shows promising results, while we also discuss some challenges that we observe.

Keywords: Semantic textual similarity, Argument similarity, Intelligent tutoring systems

1. Introduction

Recent years have seen increasing interest in applying natural language processing (NLP) applications to the field of education. Topics such as automated essay scoring (Uto et al., 2020), grammatical error correction (Bryant et al., 2019) and automated writing assistance (Zhang et al., 2019; Madnani et al., 2018) have attracted much attention. While focus has been on English, research has also be done on other languages, including Chinese (Gong et al., 2021), Spanish (González-López et al., 2020), Japanese (Mizumoto et al., 2019) and German (Horbach et al., 2017).

In the context of intelligent tutoring in German, we aim to implement an NLP-based system that supports secondary school students in a common source-based argumentative writing exercise known in German as *textgebundene Erörterung*[1] (TE). In TE, students read a text discussing a controversial social topic, such as whether or not German universities should introduce tuition fees. Students then compose an essay in which they analyse the arguments given by the author of the text and illustrate their own position on the topic. A preparation for writing the essay is the correct identification of the arguments found in the source text, guided by questions such as *what is the author's core message?* or *what arguments are presented by the author to support their stance?* Writing out the main ideas and arguments in the source text is commonly considered as a useful first-step exercise before producing the actual essay.[2]

Our long-term goal is to implement a tutoring system which performs fine-grained evaluation of students' reproduction of the arguments in the source text against a given reference and which gives formative content-related feedback. The system would compare each of the arguments identified by a student against each of the arguments in the reference answer, recognising core points from the source text that the student might have missed, and pointing them out to the student. Conversely, arguments identified by the student but not covered in the reference can be flagged to a human tutor for verification.

For such a tutoring system, a key technological challenge is automatically assessing the similarity level between a given sentential or phrasal argument produced by a student and an argument from the reference answer. This is the task that we address in this paper. Concretely, our contributions are as follows:

- To approximate student data, which are scarce, we collect a dedicated dataset through crowdsourcing, where crowdworkers complete TE-like exercises that we formulate and for which we provide reference answers for comparison (see Table 1 for an example).
- We pair arguments written by crowdworkers and corresponding reference arguments and manually annotate each pair with a semantic similarity score on a 5-point scale. Examples are shown in Table 2 below and will be explained in due course.
- We run first automatic similarity prediction experiments trained on our labelled dataset and compare various techniques.

[1] Roughly meaning "text-bound argumentation".

[2] See e.g. exam preparation materials such as Stark (2021) and educational practice materials at `https://www.tutory.de/entdecken/dokument/6ec42d20`, `https://www.cornelsen.de/magazin/beitraege/analyse-vs-eroerterung`.

Our dataset is publicly available under a Creative Commons license.[3] In follow-up work in the future, we plan to implement a working demo system that will provide fine-grained feedback messages based on the similarity detected between students' texts and the reference.

2. Related Work

The automatic assessment of students' free-form texts, including essays and short answers to prompt questions, has been a field of interest to the NLP community for many years. The English-language datasets released by Kaggle as part of the Automated Student Assessment Prize[4] (ASAP) are among the most prominent datasets for student text assessment and have been experimented on in countless works. We refer to Uto (2021), Blessing et al. (2021) and Ramesh and Sanampudi (2021) for comprehensive surveys of the vast body of literature on the subject.

Our present task is closely related to reference-based content scoring of students' short answers. Scoring systems addressing this task take text pairs, viz. matching pairs of student and reference answers, as their input and model the relation and similarity between them. For instance, building on the recent success of BERT, Sung et al. (2019) fine-tune a pre-trained BERT model on student-reference text pairs from the prominent Student Response Analysis (SRA) dataset by Dzikovska et al. (2012). They feed the output representation for the special `[CLS]` token[5] to a classifier and obtain good results on SRA. Maharjan and Rus (2019), on the other hand, represent both student and expert answers as concept maps and compare them with each other, which allows them to reveal missing information in the student's answers and to alert the student to them.

Outside of the educational domain, our task closely relates to the well-established tasks of sentential semantic textual similarity (STS) assessment and paraphrase identification (PI), which have been the focus of multiple shared tasks at SemEval (Agirre et al., 2012; Agirre et al., 2014; Xu et al., 2015; Agirre et al., 2016). An overview of various noteworthy neural approaches to the tasks is provided by Lan and Xu (2018). More recently, fine-tuning pre-trained language models such as BERT has also been shown to be successful on STS tasks (Devlin et al., 2019; Sung et al., 2019).

A special case of sentential STS is the similarity between two sentential *arguments* in the context of a debate. The Argument Facet Similarity (AFS) dataset (Misra et al., 2017) has been compiled for this task and formulates similar arguments as those that express similar propositions or similar aspects of an argument. STS in an argumentative context is in turn relevant to automated argumentation mining, where it is the basis for the extraction of argument clusters from raw texts (Boltužić and Šnajder, 2015) or for comparison and clustering of arguments represented as structured argument graphs (Lenz et al., 2019; Bergmann et al., 2019; Block et al., 2019).

Many resources, including the Kaggle datasets for student text scoring and the SemEval and AFS datasets for text and argument similarity assessment, are limited to English. Data for these tasks in German is rare. Examples known to us include ASAP-DE (Horbach et al., 2018), a German version of Kaggle's short-answer scoring dataset that the authors collected through crowdsourcing, and the CREG corpus (Ott et al., 2012; Meurers et al., 2011), which contains answers to reading comprehension questions by learners of German and corresponding target answers. However, we are not aware of existing datasets that are immediately suitable for our task.

3. Data Collection

Given our eventual goal to implement a tutoring system to support secondary school students in TE-exercises and to provide reference-based evaluation as described in Section 1, it would be ideal to create a similarity assessment model based on data from real students. However, authentic student data in German is extremely rare and difficult to obtain. As an alternative, we collect data through crowdsourcing, where crowd workers are given a TE-like exercise resembling real school exercises. Skeppstedt et al. (2018), Kim et al. (2017) and Horbach et al. (2018), among others, have shown that crowdsourcing can be a viable option for collecting textual training data for machine learning tasks in NLP, including student text evaluation.

3.1. Data Collection via Crowdsourcing

3.1.1. TE Exercise Formulation

For our formulation of the TE-oriented exercise, we choose three short openly accessible news articles which range between 590 and 714 words in length as our source text. The chosen topics are appropriate for minors and deal with issues that school students can easily relate to; thus the exercise could be presented to real students in the future. The articles deal with the following disputes:

1. Should Twitter be integrated into school classes as a tool?[6]
2. Should climate change be taught at school in a subject of its own?[7]

[3] `https://zenodo.org/record/6499223`

[4] `https://www.kaggle.com/c/asap-aes`, `https://www.kaggle.com/c/asap-sas`

[5] The special `[CLS]` token in the BERT family of models is pre-trained to capture the relation between two input text sequences based on the next-sentence-prediction task. See Devlin et al. (2019) for details.

[6] From Zeit-Online, at `https://www.zeit.de/digital/internet/2011-06/twitter-unterricht/komplettansicht`, last accessed on 17.12.2021.

[7] From Zeit-Online, at `https://www.zeit.de/gesellschaft/schule/2020-01/klimawandel-schulfach-bildung-unterricht-konkurrenz`, last accessed on 17.12.2021.

3. Should school start later in the morning?[8]

In what follows, we refer to these articles and topics as *Twitter*, *Climate* and *LateSchool*, respectively, for short.

For each topic, we ask crowd-workers to read the article and to answer the same open questions about the source text:

1. What is the discussion topic?
2. What is the main stance of the author?
3. Which arguments presented by the author support her position?
4. Which arguments in the article undermine her position?

For *LateSchool*, Question 2 is omitted because the author does not express a clear stance of her own, and Questions 3 and 4 are reformulated to *Which arguments in the article support/undermine the suggestion of starting school later?*

Since the exercise targets the identification and reproduction of relevant content from the source text and does not target specific writing skills, the crowd workers are asked to answer each of the questions in (single or multiple) bullet points consisting of concise phrases or single sentences. Henceforth, we refer to the text content in each bullet point as a *(text) snippet*. To ensure that workers fully understand the exercise, prior to working on it, they are shown an example with a different article[9] and example responses to the same questions.

Reference answers to which the crowd workers' responses are compared later are provided by the first author and a research assistant; both individually answered the questions, discussed them and agreed on a final set of reference responses. Thus, crowd workers' responses approximate student responses in a real prospective use case for intelligent tutoring, while our reference answers approximate target responses that will be supplied by teachers and educational experts in the real-world scenario. Reference responses share the format of crowd workers' responses and consist of snippets consisting of simple sentences. To illustrate, Table 1 shows the response snippets to Question 3 for the topic *Twitter* by one crowd worker, opposite our reference for that question. The texts have been translated into English and slightly simplified for brevity.

[8]From ÄrzteZeitung Online, at https://www.aerztezeitung.de/Panorama/Ist-es-vernuenftig-die-Schule-um-8-zu-beginnen-402238.html, last accessed on 17.12.2021.

[9]This article, from the Süddeutsche Zeitung Online (https://www.sueddeutsche.de/karriere/schule-oder-lehre-nichts-wie-raus-hier-1.128827, last accessed on 17.12.2021) discusses the pros and cons of leaving secondary school early for vocational training.

3.1.2. Crowdsourcing Process

We recruited crowd workers on the Prolific[10] platform. Workers had to be over 18 years old according to Prolific requirements; they also had to be fluent in German and have had secondary school education at minimum[11]. A separate crowd sourcing task was created for each topic, and workers could submit to multiple topics if they wished to. They were paid according to German minimum wage, based on the estimated time needed for completing the exercise.

We conducted the crowdsourcing process in two phases. Phase One was a pilot study involving volunteers and a small set of crowd workers. The goal was to verify that the source articles and the formulation of the exercise were clear, as well as to gain a realistic time estimation for exercise completion. Here, we received the feedback that the article for the *Climate* topic is comparatively difficult and more time-consuming to work on. Therefore, in Phase Two, which involved a larger number of crowd workers who had not participated in Phase One, we only collected submissions for the topics *Twitter* and *LateSchool*. For our final corpus, we joined all submissions that we received in the two phases.

The submissions were manually inspected. Around eight of them were considered inadequate in that they included some nonsensical submissions or showed either a poor level of German or failure to follow the exercise instructions. They were discarded and replaced with submissions by new crowd workers. In total, we obtained 50 submissions each for *Twitter* and *LateSchool*, and 17 for *Climate*. Overall, although the task for the crowd workers is more cognitively demanding than simpler annotation tasks, we consider the quality of our collected data to be satisfactory and believe that similar processes can be an option for other languages and tasks.

3.2. Data Annotation

In what follows, we refer to the crowd workers' responses as *candidate responses*. Comparison between candidate and reference responses is done on the level of snippets, in order to enable the fine-grained evaluation that we target in an intelligent tutoring context (see Section 1). For each question and each topic, we pair each snippet in the collected candidate responses with each of the snippets for that question-topic combination in the reference responses. We thereby obtain a total of 2940 candidate-reference pairs of snippets across all topics and questions. Each snippet pair has been composed in response to the same question on the same topic; we do not, for instance, pair a student snippet in response to Question 2 for topic *LateSchool* to a reference snippet for Question 4 of *Twitter*.

In the spirit of the annotation scheme for the SemEval

[10]https://www.prolific.co/

[11]This was to ensure that they would likely be familiar with TE-like exercises from school.

Twitter Question 3: Which arguments presented by the author support her position [of using Twitter in class]?	
Worker Submission	Reference
- Reticent students will feel braver - Results will be better remembered - Progress for society	- Class will become more interesting - Class will be remembered for a longer period - Shy students are encouraged to voice their opinion - Studies in the US show success of using social media in class - Educational experts recommend embracing digitalisation - ...

Table 1: Submission sample for Question 3 of topic *Twitter*, juxtaposed with our corresponding reference answer; translated into English.

STS shared tasks (Xu et al., 2015; Agirre et al., 2016), we define a 5-point similarity scale with the scores [0, 1, 2, 3, 4], where 0 indicates no similarity between two text snippets and 4 indicates near-complete semantic identity such that the two snippets are paraphrases or near-paraphrases of each other. Table 2 shows a description of each similarity level with example snippets from the *LateSchool* topic that cover arguments against starting school later, translated into English.

Each snippet pair is labelled with a similarity score. Two annotators – the first author and a research assistant – each separately annotated all 2940 pairs. Annotation results were then compared; where the difference between the two annotations was greater than 2 (around 1.4% of all cases), the sample was discussed and the annotation adjudicated. After this process, our inter-annotator agreement achieved a Pearson's correlation of 0.870, a Spearman's rank correlation of 0.835 and a quadratic weighted kappa of 0.868. To obtain the final gold-standard, we average the two annotations for each data point, again following the SemEval shared task practice (Agirre et al., 2016).

Figure 1 shows the distribution of the similarity scores across our full dataset, which range from 0 to 4 in steps of 0.5 due to averaging. Evidently, the distribution is highly imbalanced and immensely biased towards the score 0.0. This is expected since we extracted each possible snippet pairing between a candidate and a reference response within a given question and topic. They include large amounts of pairings where the candidate snippet expresses a fully valid proposition in response to the question, but where it is not being compared to the corresponding snippet in the reference response, such as in the example for the score 0 in Table 2. As such, this imbalance will remain an expected observation even when additional data collection efforts are made. We address this imbalance issue in the sections to come.

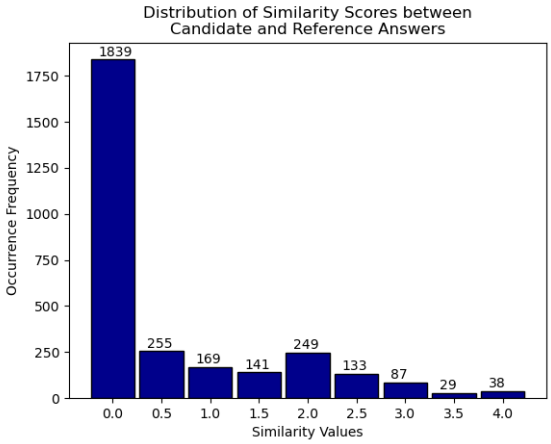

Figure 1: Similarity score distribution across the full dataset.

4. Experiments

Our machine learning task is defined as textual argument similarity judgement[12] in German. Given a pair of candidate - reference snippets, the task is to automatically predict a similarity score in the range of [0, 5]. We describe in this section our experiments using a range of regression models. Due to the small size of our dataset, we performed all experiments with randomly split 5-fold cross-validation.

4.1. Pre-processing

We performed a few steps of pre-processing:

- We removed all punctuation tokens and special tokens such as the arrow signs "->" or "-->" which some crowd workers have used.
- Crowd workers used different gender-neutral spelling variants to refer to students and teachers in German.[13] We respectively mapped each of them to the single surface form *Schüler* and *Lehrer* for the benefit of further processing.

[12] Strictly speaking, snippets in response to Question 1 for all topics cannot be considered argumentative since the question only asks students to identify the discussion topic. Snippets for all other questions express argumentative content and are the majority in the dataset. In both cases, our approach to similarity assessment is the same.

[13] To our knowledge gender-neutral spelling is not widely standardised in German. Variants used for the term *students*, for instance, include *Schüler:innen*, *Schüler_innen*, *Schüler*innen* etc..

Score	Description and *Example*
0	Snippets share no similarity and address different topics *School busses would be harder to organise* *Children would have less free time in the afternoon*
1	Snippets share no linguistic similarity but address related topics *School finishing later in the day is unpopular with students* *Children would have less free time in the afternoon*
2	Snippets share little linguistic similarity but express same arguments given source text context *School busses would have to pick up children who start school at different times* *Organisation of school busses in the countryside would be problematic*
3	Snippets share significant linguistic similarity and express same arguments given source text context *Scheduling issues with school busses* *Organisation of school busses in the countryside would be problematic*
4	Snippets are linguistically nearly equivalent, with differences being immaterial *Problems with school busses, especially in the countryside* *Organisation of school busses in the countryside would be problematic*

Table 2: Description and examples of similarity scores used in our annotation for snippets from *LateSchool* that address arguments against starting school later; translated into English.

- We lowercased all texts and normalised relevant occurrences of the umlauts *oe*, *ae* and *ue* to *ö*, *ä*, *ü*, respectively.
- We performed basic spelling correction using the Python implementation of Hunspell[14], leaving, however, words longer than 15 characters unchanged since preliminary checking revealed that Hunspell did not perform reliably on long compound words in German. Performing spelling correction as a pre-processing step is motivated by Riordan et al. (2019), who find that spelling correction in pre-processing helps to address misspellings in content-oriented student text assessment.

One exception to the steps above is that we did not lowercase the texts in our neural transformer model described in Section 4.2.3 since we used a pre-trained, *cased* language model for fine-tuning.

4.2. Models

Models we experimented with are described below. All models are implemented in Python. For the non-neural, classical models, we used Scikit-learn (Pedregosa et al., 2011); for the neural transformer one, we used PyTorch (Paszke et al., 2019) and Huggingface (Wolf et al., 2020).

4.2.1. LSA with cosine similarity

We use Latent Semantic Analysis (LSA) to embed both snippets as vectors, calculate the cosine similarity between the two vectors and scale the output similarity value to our desired range of [0, 4]. To obtain the LSA vectors, we fit a TF-IDF weighted word unigram vectoriser and a character n-gram vectoriser with character n-grams in the range of [3, 6] on all snippets in the training partition. After using them to obtain sparse vector representations of each snippet, Singular Value Decomposition (SVD) is used to reduce the dimensionality of the vectors to a fixed size n, where $n = 500$ is revealed to give the best results for us. LSA vector representation combined with cosine similarity is used by some of the earlier influential intelligent tutoring systems such as AutoTutor (Graesser et al., 2004) and DeepTutor (Rus et al., 2013) for performing reference-based assessment of students' input texts. We have therefore chosen it as a baseline.

4.2.2. Feature-based regression models

We use two traditional statistical machine learning models: the Support Vector Regressor (SVR) and the Random Forest Regressor. Both models are fed the following hand-crafted features, which are designed to capture the similarity between the input pair of snippets:

- Direct overlap between word unigrams in the two snippets. We use the openly available tool CharSplit[15] to de-compound compound words in the text for which CharSplit produces a confidence level above a pre-defined threshold.
- Overlap between lemmatised content word unigrams in the two snippets. Lemmatisation and Part-of-Speech (POS) tagging to identify content words are done using SpaCy (Honnibal and Montani, 2017).
- Overlap between character n-grams in range [3, 5] between the snippets.
- Overlap between character n-grams in range [3, 5], where all non-content words are removed from each snippet based on Spacy POS-tagging.

[14] https://pypi.org/project/hunspell/

[15] https://github.com/dtuggener/CharSplit

- For each snippet, we obtain 100-dimensional pre-trained word embeddings from FastText (Bojanowski et al., 2017) for each token in the snippet.[16] We average across the word embeddings for each token of the snippet to obtain an embedding of the whole snippet. After doing so for both snippets, we take the dot product, the cosine similarity and the euclidean distance between the two vectors as features.
- LSA vectors and cosine similarity between the two snippets as described in the previous section is used as a single feature.

For all overlap features, we take both absolute numbers of word and character n-gram overlaps and overlaps normalised by the length of each snippet.
For the SVR, we use the default implementation by Scikit-learn, which uses a radial basis kernel. For Random Forest, we also use Scikit-learn's default parameters, apart from setting the minimum number of samples at leaf nodes to 3 to avoid overfitting. To address the strong imbalance in the data as shown in Figure 1, we set the sample weight of each training sample with the score 0.0 to 0.3 and the weight of all other training samples to 0.7.

4.2.3. Fine-tuning pre-trained BERT

Our last approach is fine-tuning a pre-trained BERT model. BERT has been shown to perform well on sentence-pair modelling tasks such as reference-based student answer assessment (Sung et al., 2019) and STS (Devlin et al., 2019). We use the pre-trained BERT model "bert-base-german-cased", released by deepset[17] and integrated into Huggingface (Wolf et al., 2020).
Similarly to Sung et al. (2019), we feed candidate–reference snippet pairs to BERT and extract the dense representation of the special `[CLS]` token, which has been pre-trained on next sentence prediction. We then send it through a linear layer with a single output and apply the sigmoid function to it, which yields a single model output score in the range of [0, 1]. During training, we scale our target similarity scores to the range of [0, 1] for loss computation, whereas at inference time on the test partition, we scale the model's output to our desired range of [0, 4].
We adopt mean squared error (MSE) as our loss function. Moreover, we use Adam optimisation with weight decay (Loshchilov and Hutter, 2017) with the default parameters implemented by Huggingface, except for setting the learning rate to $5e^{-5}$. We set our batch size to 16 and train for 11 epochs, which is where test error stopped decreasing when we performed a preliminary run on a small test set.

[16]The downloaded embeddings are 300-dimensional. We reduce the vectors to 100 dimensions with FastText's built-in dimension reducer for efficiency. Using 300-D vectors does not yield better performance than 100-D ones in our case.

[17]`https://www.deepset.ai/german-bert`

4.3. Evaluation Metrics

We use MSE, Spearman's rank correlation and Pearson's correlation with the gold standard similarity scores as regression evaluation metrics. Pearson is used as the main evaluation metric in the SemEval shared tasks on STS (Agirre et al., 2014; Agirre et al., 2016), while Reimers et al. (2016) find Pearson's correlation to be insufficient on its own and recommend using other metrics such as Spearman's rank correlation in addition.
Furthermore, we also evaluate model performances using classification metrics, for which both predicted and gold-standard scores must be mapped into pre-defined bins (see also Reimers et al. (2016)):

- In a fine-grained 5-class setting, we round all scores to integers in the range [0, 4], which corresponds to the 5-level similarity scores used for data annotation.
- In a second setting, we map all scores to three coarse-grained similarity levels. To do so, we arbitrarily set the boundaries for the three classes at 0.9 and 3.0, such that scores lower than 0.9 map to the level *low*, scores between 0.9 and 3.0 to *medium* and scores above or equal to 3.0 to *high*.

We see two main motivations for using classification metrics in addition: First, we can separately calculate precision, recall and F1-scores for each class and compute confusion matrices to investigate prediction errors in different areas on the similarity scale. Second, in the prospective use case in intelligent tutoring, we aim to give pre-designed formative feedback on individual snippets in students' responses, where the feedback message to display will be determined by the discrete similarity level that is detected between the snippets in the student and the reference response. For instance, in the 3-class setting, where a reference snippet r for a given question-topic combination displays low similarity with all student snippets for that same question and topic, feedback can inform the student that she has missed a relevant main idea in her answer and prompt her to look for the proposition expressed by r in the source text; where r shows high similarity with at least one student snippet, the feedback can congratulate the student on having correctly identified the idea expressed by r. Similarly, more fine-grained feedback messages can be designed when five similarity classes are used.
In both classification settings, the metrics used are macro-averaged F1 and quadratic weighted kappa (QWK). QWK measures the amount of agreement between two annotations that use discrete, ordered labels and takes into account the extent of misclassification with respect to the label scale when computing the metric. A QWK value of 0 indicates no agreement and 1 indicates perfect agreement. It is the official evaluation metric used in the Kaggle ASAP essay scoring and

short-answer scoring competitions[18] and has since been used as a standard metric in the domain of automatic scoring of students' texts (Ke and Ng, 2019; Ramesh and Sanampudi, 2021). We measure the agreement between the gold-standard and the model-predicted similarity scores. Since the labels are highly imbalanced in our data, we consider accuracy to be uninformative and therefore do not report it.

4.4. Results

Table 3 shows the results for our experiments as averaged across all five folds in cross-validation. We also display the inter-annotator agreement reported in Section 3.2 as an upper-bound reference point.

Overall, LSA vectors with cosine similarity as a baseline show particularly low performance when evaluated in the two classification settings. Among the two feature-based models, Random Forest outperforms SVR both with regard to regression and classification metrics. However, the BERT-based model clearly beats all other models we have experimented with, reaching, in fact, a higher Pearson's correlation than the agreement between the two human annotators.

In the classification settings, performance on different classes vary. Figure 2 visualises the precision, recall and F1 scores per class of the best-performing BERT-based model in the 5-class evaluation settings, averaged across all folds. Clearly, prediction is best for the most frequent class 0.0 (see Figure 1), whereas performances for all other classes leave room for improvement even for the BERT-based model. Table 4 shows detailed results for the same model in the 3-class setting, where it is also revealed that performance on the similarity class "High", especially its recall, is particularly in need of improvement.

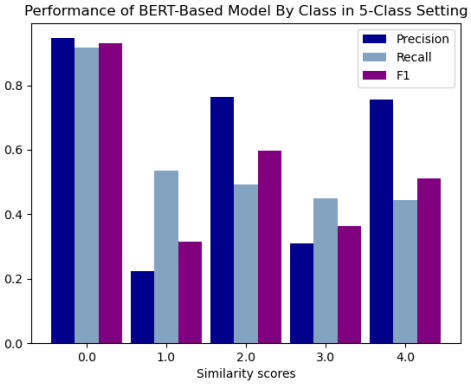

Figure 2: Performance of BERT-based model in terms of precision, recall and F1-score by class in the 5-class setting; all numbers averaged across five folds.

[18] See https://www.kaggle.com/c/asap-aes/overview/evaluation and https://www.kaggle.com/c/asap-sas/overview/evaluation for details on the metric.

5. Discussion

As mentioned, the dataset is heavily biased towards the similarity score 0 or 0.0, which applies both to the training and the test partitions in cross-validation. Two major issues arise from this imbalance: First, in training the models can pick up a bias towards this most frequent similarity level, and in general to lower similarity scores. Figure 3 shows a confusion matrix for the best feature-based model, the sample-weighted Random Forest model, normalised by the true class labels.[19] The values are averaged across five folds. We see that a larger percentage of true samples in class 1.0 are misclassified as belonging to class 0.0 (44%) than are correctly classified as belonging to class 1.0 (40%); 17% of true samples in class 2.0 are also misclassified as belonging to class 0.0; and there appears to be a clear trend for test samples to be misclassified as a lower than as a higher class.

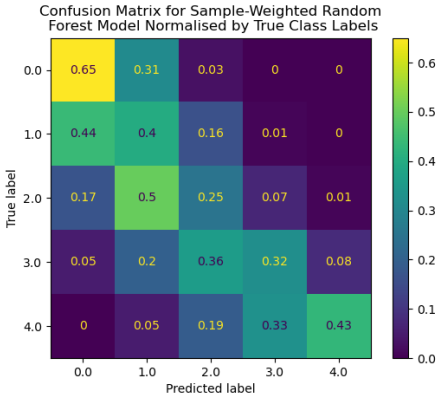

Figure 3: Confusion matrix for sample-weighted Random Forest model, normalised by true class labels (i.e. rows add up to 1); all values averaged across five folds.

The effect is less pronounced in the best-performing BERT-based model, as shown in Figure 4, although there is also a trend of misclassifying samples into the lower, neighbouring class rather than the higher one.

Furthermore, a second major problem arising from label imbalance is that the imbalance is also reflected in the test partitions, which often contain few samples of the less frequent similarity classes. For instance, Table 5 shows the gold label distribution in the 5-class setting in the test partition of each fold when we tested the BERT-based model.

The small number of test samples for some classes leads to highly fluctuating performance metrics on those classes across folds, which limits their reliability. For instance, for the least frequent class of 4.0, the F1 score in the BERT-based experiment ranges from 0.308

[19] That is, each row in the matrix adds up to approximately 1. For all true labels of a given class, the matrix shows the percentage of those labels which have been predicted as any of the possible classes.

Model	Regression			5-Way Classification		3-Way Classification	
	MSE	Spearman	Pearson	QWK	Macro-F1	QWK	Macro-F1
LSA + cosine similarity	0.631	0.475	0.618	0.480	0.241	0.471	0.483
Sample-weighted SVR	0.635	0.488	0.648	0.594	0.390	0.518	0.595
Sample-weighted Random Forest	0.558	0.529	0.680	0.617	0.401	0.564	0.630
BERT	**0.229**	**0.773**	**0.881**	**0.834**	**0.543**	**0.784**	**0.741**
Annotation Agreement		0.835	0.870	0.868			

Table 3: Model performances as averaged across all folds in 5-fold cross-validation, with best results in bold. Note that MSE is the only metric in which lower values indicate better performance.

	Precision	Recall	F1
Low	0.918	0.964	0.940
Medium	0.764	0.721	0.740
High	0.779	0.422	0.543

Table 4: By-class results of the BERT-based model in the 3-class setting; all numbers averaged across five folds

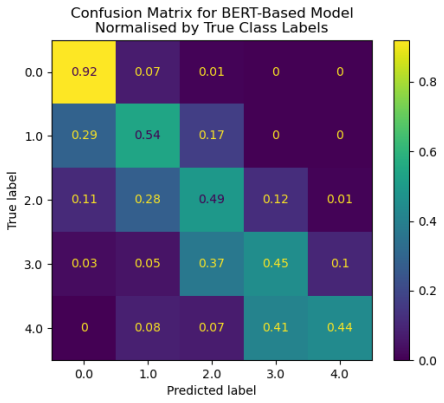

Figure 4: Confusion matrix for the BERT-based model, normalised by true class labels (i.e. rows add up to 1); all values averaged across five folds.

	0.0	1.0	2.0	3.0	4.0
Fold 1	411	40	107	18	12
Fold 2	428	28	96	21	15
Fold 3	425	34	106	12	11
Fold 4	419	38	105	16	10
Fold 5	411	29	109	20	19

Table 5: Fold-wise test label distributions in the 5-class setting in the BERT-based experiment

to 0.667 across the five folds. Collecting more data overall would increase the absolute number of samples in these similarity classes, although it would not change the relative label imbalance across similarity scores. On the whole, label imbalance remains a major challenge to our dataset and our task.

6. Conclusion

In the present paper we have presented our task of similarity assessment between two text snippets in German that express argumentative propositions in reference to a source text. The task is designed to form the technological basis for a prospective intelligent tutoring system that will support school students in source-based argumentative writing exercises by verifying that they have correctly understood key arguments in the source text. We collected a dataset for the task through crowdsourcing, provided similarity annotations for them and conducted first similarity prediction experiments on the dataset. In particular, the BERT-based model has shown promising results.

Nonetheless, our task and dataset are highly challenging. Aside from the data imbalance issue, which has been discussed in the previous section, there are two further challenges: First, in order to approximate authentic TE-like school exercises, our question prompts are extremely open. They do not suggest specific vocabulary or phrasing to students / crowd workers and therefore yield a high level of linguistic variance in candidate responses, which makes it harder for models to recognise semantically similar yet differently phrased snippets. Second, our similarity assessment task is *source-based* and involves texts written in response to the source article. Candidate - reference pairs that are not linguistically similar can still be expressing the same point *in the context of the source article*. This again makes similarity recognition more difficult.

Follow-up studies to our present work will aim for the following: We are already in the process of enlarging our dataset by repeating the data collection process on Prolific. We aim to address the label imbalance problem with different down-sampling or data augmentation methods. Finally, we will experiment with ways to incorporate the source article into the BERT model as contextual information. Due to the difficulty of the task, greater focus can be put on the 3-level classification into ”low”, ”medium” and ”high” similarity. We believe reliable performance on this coarse-grained scale can already serve as the basis of a useful tutoring tool.

7. Acknowledgements

We thank Kristin Howitt for her contributions to compiling and annotating the corpus. Moreoever, we thank the anonymous reviewers for their comments on the first draft of this paper. The first author is funded by a grant from the German Federal Ministry of Education and Research (BMBF) as part of the project "Adaptive AI-based Learning Assistant for Schools" (AKILAS), grant number 16SV8610.

8. Bibliographical References

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. (2014). Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez Agirre, A., Mihalcea, R., Rigau Claramunt, G., and Wiebe, J. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Bergmann, R., Lenz, M., Ollinger, S., and Pfister, M. (2019). Similarity measures for case-based retrieval of natural language argument graphs in argumentation machines. In *The Thirty-Second International Flairs Conference*.

Blessing, G., Azeta, A., Misra, S., Chigozie, F., and Ahuja, R., (2021). *A Machine Learning Prediction of Automatic Text Based Assessment for Open and Distance Learning: A Review*, pages 369–380. 01.

Block, K., Trumm, S., Sahitaj, P., Ollinger, S., and Bergmann, R. (2019). Clustering of argument graphs using semantic similarity measures. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 101–114. Springer.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*, June. arXiv: 1607.04606.

Boltužić, F. and Šnajder, J. (2015). Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115.

Bryant, C., Felice, M., Andersen, Ø. E., and Briscoe, T. (2019). The bea-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May. arXiv: 1810.04805.

Dzikovska, M. O., Nielsen, R., and Brew, C. (2012). Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 200–210.

Gong, J., Hu, X., Song, W., Fu, R., Sheng, Z., Zhu, B., Wang, S., and Liu, T. (2021). IFlyEA: A Chinese Essay Assessment System with Automated Rating, Review Generation, and Recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 240–248, Online, August. Association for Computational Linguistics.

González-López, S., Bethard, S., and Lopez-Lopez, A. (2020). Assisting Undergraduate Students in Writing Spanish Methodology Sections. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 115–123, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., and Louwerse, M. M. (2004). Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36(2):180–192.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Horbach, A., Scholten-Akoun, D., Ding, Y., and Zesch, T. (2017). Fine-grained essay scoring of a complex writing task for native speakers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 357–366, Copenhagen, Denmark, September. Association for Computational Linguistics.

Horbach, A., Stennmanns, S., and Zesch, T. (2018). Cross-Lingual Content Scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 410–419, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ke, Z. and Ng, V. (2019). Automated Essay Scoring: A Survey of the State of the Art. pages 6300–6308.

Kim, S., Lee, S., Park, D., and Kang, J. (2017). Constructing and evaluating a novel crowdsourcing-

based paraphrased opinion spam dataset. In *Proceedings of the 26th international conference on world wide web*, pages 827–836.

Lan, W. and Xu, W. (2018). Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3890–3902.

Lenz, M., Ollinger, S., Sahitaj, P., and Bergmann, R. (2019). Semantic textual similarity measures for case-based retrieval of argument graphs. In *International Conference on Case-Based Reasoning*, pages 219–234. Springer.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Madnani, N., Burstein, J., Elliot, N., Beigman Klebanov, B., Napolitano, D., Andreyev, S., and Schwartz, M. (2018). Writing Mentor: Self-Regulated Writing Feedback for Struggling Writers. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 113–117, Santa Fe, New Mexico, August. Association for Computational Linguistics.

Maharjan, N. and Rus, V. (2019). A Concept Map Based Assessment of Free Student Answers in Tutorial Dialogues. In Seiji Isotani, et al., editors, *Artificial Intelligence in Education*, Lecture Notes in Computer Science, pages 244–257, Cham. Springer International Publishing.

Meurers, D., Ziai, R., Ott, N., and Kopp, J. (2011). Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9, Edinburgh, Scottland, UK, July. Association for Computational Linguistics.

Misra, A., Ecker, B., and Walker, M. A. (2017). Measuring the similarity of sentential arguments in dialog. *arXiv preprint arXiv:1709.01887*.

Mizumoto, T., Ouchi, H., Isobe, Y., Reisert, P., Nagata, R., Sekine, S., and Inui, K. (2019). Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 316–325, Florence, Italy, August. Association for Computational Linguistics.

Ott, N., Ziai, R., and Meurers, D. (2012). Creation and analysis of a reading comprehension exercise corpus. *Multilingual corpora and multilingual corpus analysis*, 14:47.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Ramesh, D. and Sanampudi, S. K. (2021). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, pages 1–33.

Reimers, N., Beyer, P., and Gurevych, I. (2016). Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Riordan, B., Flor, M., and Pugh, R. (2019). How to account for mispellings: Quantifying the benefit of character representations in neural content scoring models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 116–126, Florence, Italy, August. Association for Computational Linguistics.

Rus, V., D'Mello, S., Hu, X., and Graesser, A. (2013). Recent advances in conversational intelligent tutoring systems. *AI magazine*, 34(3):42–54.

Skeppstedt, M., Peldszus, A., and Stede, M. (2018). More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 155–163, Brussels, Belgium. Association for Computational Linguistics.

Stark. (2021). *STARK Training MSA/eBBR 2022 - Deutsch - Berlin/Brandenburg*. Stark Verlag GmbH.

Sung, C., Dhamecha, T. I., and Mukhi, N. (2019). Improving Short Answer Grading Using Transformer-Based Pre-training. In Seiji Isotani, et al., editors, *Artificial Intelligence in Education*, volume 11625, pages 469–481. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Uto, M., Xie, Y., and Ueno, M. (2020). Neural Automated Essay Scoring Incorporating Handcrafted Features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, pages 1–26.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (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, October. Association for Computational Linguistics.

Xu, W., Callison-Burch, C., and Dolan, W. B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.

Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsumura, L. C., Howe, E., and Quintana, R. (2019). eRevise: Using Natural Language Processing to Provide Formative Feedback on Text Evidence Usage in Student Writing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9619–9625, July. arXiv: 1908.01992.