Supporting Cognitive and Emotional Empathic Writing of Students

Thiemo Wambsganss[1,2], **Christina Niklaus**[1, 3], **Matthias Söllner**[4],
Siegfried Handschuh[1, 3] and **Jan Marco Leimeister**[1, 4]
[1] University of St.Gallen
{thiemo.wambsganss, christina.niklaus,
siegfried.handschuh, janmarco.leimeister}@unisg.ch
[2] Carnegie Mellon University
twambsga@andrew.cmu.edu
[3] University of Passau
{christina.niklaus, siegfried.handschuh}@uni-passau.de
[4] University of Kassel
{soellner, leimeister}@uni-kassel.de

Abstract

We present an annotation approach to capturing emotional and cognitive empathy in student-written peer reviews on business models in German. We propose an annotation scheme that allows us to model emotional and cognitive empathy scores based on three types of review components. Also, we conducted an annotation study with three annotators based on 92 student essays to evaluate our annotation scheme. The obtained inter-rater agreement of α=0.79 for the components and the multi-π=0.41 for the empathy scores indicate that the proposed annotation scheme successfully guides annotators to a substantial to moderate agreement. Moreover, we trained predictive models to detect the annotated empathy structures and embedded them in an adaptive writing support system for students to receive individual empathy feedback independent of an instructor, time, and location. We evaluated our tool in a peer learning exercise with 58 students and found promising results for perceived empathy skill learning, perceived feedback accuracy, and intention to use. Finally, we present our freely available corpus of 500 empathy-annotated, student-written peer reviews on business models and our annotation guidelines to encourage future research on the design and development of empathy support systems.

1 Introduction

Empathy is an elementary skill in society for daily interaction and professional communication and is therefore elementary for educational curricula (e.g., Learning Framework 2030 (OECD, 2018)). It is the *"ability to simply understand the other person's perspective [...] and to react to the observed experiences of another,"* (Davis, 1983, p.1)[1]. Empathy skills not only pave the foundation for successful interactions in digital companies, e.g., in agile work environments (Luca and Tarricone, 2001), but they are also one of the key abilities in the future that will distinguish the human workforce and artificial intelligence agents from one another (Poser and Bittner, 2020). However, besides the growing importance of empathy, research has shown that empathy skills of US college students decreased from 1979 to 2009 by more than thirty percent and even more rapidly between 2000 to 2009 (Konrath et al., 2011). On these grounds, the Organization for Economic Cooperation and Development (OECD) claims that the training for empathy skills should receive a more prominent role in today's higher education (OECD, 2018).

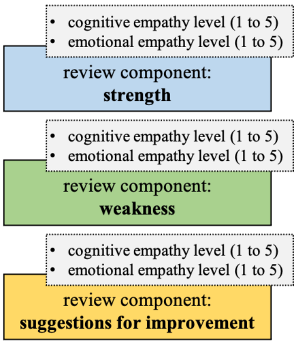

• cognitive empathy level (1 to 5)
• emotional empathy level (1 to 5)
review component: **strength**
• cognitive empathy level (1 to 5)
• emotional empathy level (1 to 5)
review component: **weakness**
• cognitive empathy level (1 to 5)
• emotional empathy level (1 to 5)
review component: **suggestions for improvement**

Figure 1: Empathy annotation scheme. First, a text paragraph is classified into a peer review component (*strengths, weakness, improvement suggestions*). Second, the same annotator is then scoring the cognitive and emotional empathy level of the components based on our annotation guideline on a 1-to-5 scale.

[1] Being aware that empathy is a multidimensional construct, in this study, we focus on emotional and cognitive empathy (Spreng et al., 2009; Davis, 1983).

To train students with regard to empathy, educational institutions traditionally rely on experiential learning scenarios, such as shadowing, communication skills training, or role playing (Lok and Foster, 2019; van Berkhout and Malouff, 2016). Individual empathy training is only available for a limited number of students since individual feedback through a student's learning journey is often hindered due to large-scale lectures or the growing field of distance learning scenarios such as Massive Open Online Classes (MOOCs) (Seaman et al., 2018; Hattie and Timperley, 2007).

One possible path for providing individual learning conditions is to leverage recent developments in computational linguistics. Language-based models enable the development of writing support systems that provide tailored feedback and recommendations (Santos et al., 2018), e.g., like those already used for argumentation skill learning (Wambsganss et al., 2020a, 2021b). Recently, studies have started investigating elaborated models of human emotions (e.g., Wang et al. (2016), Abdul-Mageed and Ungar (2017), Buechel and Hahn (2018), or Sharma et al. (2020)), but available corpora for empathy detection are still rare. Only a few studies address the detection and prediction of empathy in natural texts (Khanpour et al., 2017; Xiao et al., 2012), and, to the best of our knowledge, only one corpus is publicly available for empathy modelling based on news story reactions (Buechel et al., 2018). Past literature therefore lacks 1) publicly available empathy annotated data sets, 2) empathy annotation models based on rigorous annotation guidelines combined with annotation studies to assess the quality of the data, 3) the alignment of empathy in literature on psychological constructs and theories, and 4) an embedding and real-world evaluation of novel modelling approaches in collaborative learning scenarios (Rosé et al., 2008).

We introduce an empathy annotation scheme and a corpus of 500 student-written reviews that are annotated for the three types of review components, *strengths, weaknesses*, and *suggestions for improvements*, and their embedded *emotional* and *cognitive empathy level* based on psychological theory (Davis, 1983; Spreng et al., 2009). We trained different models and embedded them as feedback algorithms in a novel writing support tool, which provided students with individual empathy feedback and recommendations in peer learning scenarios. The measured empathy skill learning (Spreng et al., 2009), the perceived feedback accuracy (Podsakoff and Farh, 1989), and the intention to use (Venkatesh and Bala, 2008) in a controlled evaluation with 58 students provided promising results for using our approach in different peer learning scenarios to offer quality education independent of an instructor, time, and location.

Our contribution is fourfold: 1) we derive a novel annotation scheme for empathy modeling based on psychological theory and previous work on empathy annotation (Buechel et al., 2018); 2) we present an annotation study based on 92 student peer reviews and three annotators to show that the annotation of empathy in student peer reviews is reliably possible; 3) to the best of our knowledge, we present the second freely available corpus for empathy detection in general and the first corpus for empathy detection in the educational domain based on 500 student peer reviews collected in our lecture about business innovation in German; 4) we embedded our annotation approach as predictive models in a writing support system and evaluated it with 58 students in a controlled peer learning scenario. We hope to encourage research on student-written empathetic texts and writing support systems to train students' empathy skills based on NLP towards a quality education independent of a student's location or instructors.

2 Background

The Construct of Empathy The ability to perceive the feelings of another person and react to their emotions in the right way requires empathy – the ability "*of one individual to react to the observed experiences of another*" (Davis (1983), p.1). Empathy plays an essential role in daily life in many practical situations, such as client communication, leadership, or agile teamwork. Despite the interdisciplinary research interest, the term empathy is defined from multiple perspectives in terms of its dimensions or components (Decety and Jackson, 2004). Aware of the multiple perspectives on empathy, in this annotation study, we focused on the cognitive and emotional components of empathy as defined by Davis (1983) and Lawrence et al. (2004). Therefore, we follow the *'Toronto Empathy Scale'* (Spreng et al., 2009) as a synthesis of instruments for measuring and validating empathy. Hence, empathy consists of both emotional and cognitive components (Spreng et al., 2009). While emotional empathy lets us perceive what other people feel, cognitive empathy is the human ability to recognize and understand other individuals (Lawrence et al., 2004).

Emotion and Empathy Detection In NLP, the detection of empathy in texts is usually regarded as a subset of emotion detection, which in turn is often referred to as part of sentiment analysis. The detection of emotions in texts has made major progress, with sentiment analysis being one of the most prominent areas in recent years (Liu, 2015). However, most scientific studies have been focusing on the prediction of the polarity of words for assessing negative and positive notions (e.g., in online forums (Abbasi et al., 2008) or twitter postings (Rosenthal et al., 2018)). Moreover, researchers have also started investigating more elaborated models of human emotions (e.g., Wang et al. (2016), Abdul-Mageed and Ungar (2017), and Mohammad and Bravo-Marquez (2017)). Several corpora exist where researchers have annotated and assessed the emotional level of texts. For example, Scherer and Wallbott (1994) published an emotion-labelled corpus based on seven different emotional states. Strapparava and Mihalcea (2007) classified news headlines based on the basic emotions scale of Ekman (1992) (i.e., *anger, disgust, fear, happiness, sadness* and *surprise*). More recently, Chen et al. (2018) published *EmotionLines*, an emotion corpus of multi-party conversations, as the first data set with emotion labels for all utterances was only based on their textual content. Bostan and Klinger (2018) presented a novel unified domain-independent corpus based on eleven emotions as the common label set. However, besides the multiple corpora available for emotion detection in texts, corpora for empathy detection are rather rare. As Buechel et al. (2018) also outline, the construction of corpora for empathy detection and empathy modelling might be less investigated due to various psychological perspectives on the construct of empathy. Most of the works for empathy detection focus, therefore, on spoken dialogue, addressing conversational agents, psychological interventions, or call center applications (e.g., McQuiggan and Lester (2007), Pérez-Rosas et al. (2017), Alam et al. (2018), Sharma et al. (2020)) rather than written texts. Consequently, there are hardly any corpora available in different domains and languages that enable researchers in training models to detect the empathy level in texts, e.g., by providing students with individual empathy feedback (Buechel et al., 2018).

Empathy Annotated Corpora and Annotation Schemes Only a few studies address the detection and prediction of empathy in natural language texts (e.g., Khanpour et al. (2017) and Xiao et al. (2012)). Presenting the first and only available gold standard data set for empathy detection, Buechel et al. (2018) constructed a corpus in which crowdworkers were asked to write emphatic reactions to news stories. Before the writing tasks, the crowdworkers were asked to conduct a short survey with self-reported items to measure their empathy level and their personal distress based on Batson et al. (1987). The scores from the survey were then taken as the annotation score for the overall news reaction message. The final corpus consisted of 1,860 annotated messages (Buechel et al., 2018). Nevertheless, previous empathy annotations on natural texts merely focused on intuition-based labels instead of rigorous annotation guidelines combined with annotation studies by researchers to assess the quality of the corpora (i.e., as is done for corpora of other writing support tasks, e.g., argumentative student essays by Stab and Gurevych (2017)). Moreover, previous annotations have mostly been conducted at the overall document level, resulting in one generic score for the whole document, which makes the corpus harder to apply to writing support systems.

Consequently, there is a *lack of linguistic corpora for empathy detection in general* and, more specifically, for training models that provide students with adaptive support and feedback about their empathy in common pedagogical scenarios like large-scale lectures or the growing field of MOOCs (Wambsganss et al., 2021c, 2020b). In fact, in the literature about computer-supported collaborative learning (Dillenbourg et al., 2009), we found only one approach by Santos et al. (2018) that used a dictionary-based approach to provide students with feedback on the empathy level of their texts. We aim to address this literature gap by presenting and evaluating an annotation scheme and an annotated empathy corpus built on student-written texts with the objective to develop intelligent and accurate empathy writing support systems for students.

3 Corpus Construction

We compiled a corpus of 500 student-generated peer reviews in which students provided each other with feedback on previously developed business models. Peer reviews are a modern learning scenario in large-scale lectures, enabling students to reflect on their content, receive individual feedback from peers, and thus deepen their understanding of the content (Rietsche and Söllner, 2019). Moreover, they are easy to set up in traditional large-scale learning scenarios or the growing field of distance-learning scenarios such as MOOCs. This can be leveraged to train skills such as *the ability to appropriately react to other students' perspectives* (e.g., Santos et al. (2018)). Therefore, we aim to create an annotated corpus to provide empathy feedback based on a data set that A) is based on real-world student peer reviews, B) consists of a sufficient corpus size to be able to train models in a real-world scenario and C) follows a novel annotation guideline for guiding the annotators towards an adequate agreement. Hence, we propose a new annotation scheme to model peer review components and their emotional and cognitive empathy levels that reflect the feedback discourse in peer review texts. We base our empathy annotation scheme on emotional and cognitive empathy following Davis (1983) and Spreng et al. (2009) guided by the study of Buechel et al. (2018). To build a reliable corpus, we followed a 4-step methodology: 1) we examined scientific literature and theory on the construct of empathy and on how to model empathy structures in texts from different domains; 2) we randomly sampled 92 student-generated peer reviews and, on the basis of our findings from literature and theory, developed a set of annotation guidelines consisting of rules and limitations on how to annotate emphatic review discourse structures; 3) we applied, evaluated, and improved our guidelines with three native speakers of German in a total of eight consecutive workshops to resolve annotation ambiguities; 4) we followed the final annotation scheme based on our 14-page guidelines to annotate a corpus of 500 student-generated peer reviews.[2]

3.1 Data Source

We gathered a corpus of 500 student-generated peer reviews written in German. The data was collected in a business innovation lecture in a master's program at a Western European university. In this lecture, around 200 students develop and present a new business model for which they receive three peer reviews each. Here, a fellow student from the same course elaborates on the strengths and weaknesses of the business model and gives recommendations on what could be improved. We collected a random subset of 500 of these reviews from around 7,000 documents collected from the years 2014 to 2018 in line with the ethical guidelines of our university and with approval from the students to utilize the writings for scientific purposes. An average peer review consists of 200 to 300 tokens (in our corpus we counted a mean of 19 sentences and 254 tokens per document). A peer review example is displayed in Figure 2.

3.2 Annotation Scheme

Our objective is to model the empathy structures of student-generated peer reviews by annotating the review components and their emotional and cognitive empathy levels. Most of the peer reviews in our corpus followed a similar structure. They described several strengths or weaknesses of the business model under consideration, backing them up by examples or further elaboration. Moreover, the students formulated certain suggestions for improvements of the business model. These review components (i.e., *strengths, weaknesses, and suggestions for improvement*) were written with different empathetic levels, sometimes directly criticizing the content harshly, sometimes empathetically referring to weaknesses as further potentials for improvement with examples and explanation. We aim to capture these empathic differences between the peer reviews with two empathy level scores, the *cognitive empathy level* of a certain review component and the *emotional empathy* level of a certain component. Our basic annotation scheme is illustrated in Figure 1.

3.2.1 Review Components

For the review components, we follow established models of feedback structures suggested by feedback theory (e.g., Hattie and Timperley (2007) or Black and Wiliam (2009)). A typical peer review, therefore, consists of three parts: 1) elaboration of strengths, 2) elaboration of weaknesses, and 3) suggestions for improvements (to answer *"Where am I going and how am I going?"* and *"Where do I go next?"*, i.e., Hattie and Timperley (2007)). Accordingly, the content of a review consists of multiple components, including several controversial statements (e.g., a claim about a strength or

[2] The annotation guidelines as well as the entire corpus can be accessed at `https://github.com/thiemowa/empathy_annotated_peer_reviews`.

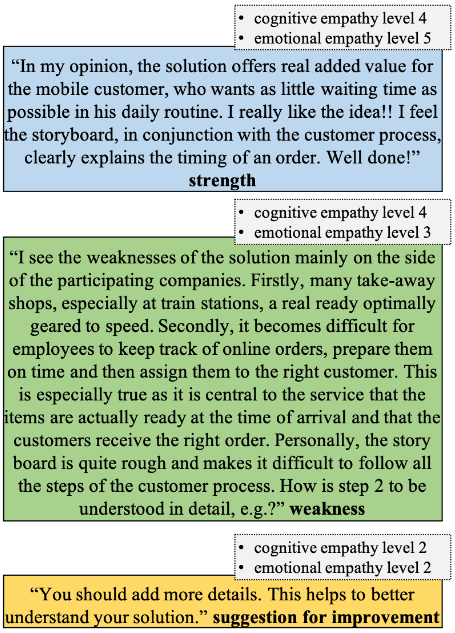
weakness of a business model) that are usually supported by elaborations or examples (i.e., a *premise*) (Toulmin, 1984). Also, in the domain of student-written peer reviews, we found that a standpoint and its elaboration are the central element of a review component. Accordingly, we summarized all the claims and premises which described positive aspects of a business model as *strengths*. All content (claims and premises) describing negative aspects were modelled as *weaknesses*, while claims and premises with certain content for improvement were modelled as *suggestions for improvement*, following the structure of a typical review. Besides the content, syntactical elements and key words were used as characteristics for the compound classification, e.g., most students introduced a review component by starting with structural indications such as *"Strengths:"* or *"Weaknesses:"* in their peer review texts.

3.2.2 Empathy Level

To capture the differences in the empathy levels of the peer reviews (i.e., the way the writer was conveying their feedback (Hattie and Timperley, 2007)), we followed the approach of Davis (1983) and Spreng et al. (2009) for cognitive and emotional empathy. Cognitive empathy (perspective taking) is the writer's ability to use cognitive processes, such as role taking, perspective taking, or "*decentering*," while evaluating the peers' submitted tasks. The student sets aside their own perspective and *"steps into the shoes of the other."* Cognitive empathy can happen purely cognitively, in that there is no reference to any affective state, (Baron-Cohen and Wheelwright, 2004) but it mostly includes understanding the other's emotional state as well. The following example displays high cognitive empathy: *"You could then say, for example, 'Since market services are not differentiated according to customer segments and locations, the following business areas result... And that due to the given scope of this task you will focus on the Concierge-Service business segment.' After that, you have correctly only dealt with this business segment."* Emotional empathy (emphatic concern) is the writer's emotional response to the peers' affective state. The students can either show the same emotions as read in the review or simply state an appropriate feeling towards the peer. Typical examples include sharing excitement with the peer about the business model submitted or showing concern over the peer's opinion. The following example depicts high emotional empathy: *"I think your idea is brilliant!"*.

Both constructs are measured on a scale from 1-5 following the empathy scale range of Moyers and Martin (2010), with every level being precisely defined in our annotation guidelines. A summary of the definitions for both empathy level scores are displayed in Table 1 and Table 2. A more detailed description of both scores can be found in the appendix in Table 7 and Table 8.[3]

Figure 2 illustrates an example of an entire peer review that is annotated for *strength, weakness and suggestion for improvement* and the cognitive and emotional empathy scores.[4]



- cognitive empathy level 4
- emotional empathy level 5

"In my opinion, the solution offers real added value for the mobile customer, who wants as little waiting time as possible in his daily routine. I really like the idea!! I feel the storyboard, in conjunction with the customer process, clearly explains the timing of an order. Well done!" **strength**

- cognitive empathy level 4
- emotional empathy level 3

"I see the weaknesses of the solution mainly on the side of the participating companies. Firstly, many take-away shops, especially at train stations, a real ready optimally geared to speed. Secondly, it becomes difficult for employees to keep track of online orders, prepare them on time and then assign them to the right customer. This is especially true as it is central to the service that the items are actually ready at the time of arrival and that the customers receive the right order. Personally, the story board is quite rough and makes it difficult to follow all the steps of the customer process. How is step 2 to be understood in detail, e.g.?" **weakness**

- cognitive empathy level 2
- emotional empathy level 2

"You should add more details. This helps to better understand your solution." **suggestion for improvement**

Figure 2: Fully annotated example of a peer review.

3.3 Annotation Process

Three native German speakers annotated the peer reviews independently from each other for the components *strengths, weaknesses* and *suggestions for improvement*, as well as their *cognitive and emotional empathy levels* according to the annotation guidelines we specified. The annotators were master's students in business innovation from a European university with bachelor's degrees in business administration and were, therefore, domain experts in the field of business models. Inspired by Stab

[3] More elaborated definitions, examples, and key word lists for both empathy scales can be found in our annotation guidelines.

[4] Since the original texts are written in German, we translated the examples to English for the sake of this paper.

Score	Description
5	The student fully understands the peer's thoughts. She completely stepped outside her own perspective and thinks from the peer's perspective. She does that by carefully evaluating the peer's idea with rich explanations. Questions, personal pronouns, or direct addressing of the author could be used in order to better understand and elaborate on the peer's perspective.
4	The student thinks from the perspective of the peer. She elaborates in a way that serves the peer best to further establish the idea or activity. Each component is affirmed with further explanations.
3	The student tries to understand the perspective of the peer and adds further elaborations to her statements. However, her elaborations are not completely thought through, and her feedback is missing some essential explanations, examples, or questions to make sure she understood everything correctly.
2	The student did not try to understand the peer's perspective. The student rather just tried to accomplish the task of giving feedback.
1	The student's feedback is very short and does not include the peer's perspective. She does not add any further elaboration in her thoughts.

Table 1: Description of the cognitive empathy scores.

Score	Description
5	The student was able to respond very emotionally to the peer's work and fully represents the affectional state in her entire review. She illustrates this by writing in a very emotional and personal manner and expressing her feelings (positive or negative) throughout the review. Strong expressions include exclamation marks (!).
4	The student was able to respond emotionally to the peer's submitted activity with suitable emotions (positive or negative). She returns emotions in her feedback on various locations and expresses her feelings by using the personal pronouns ("I", "You"). Some sentences might include exclamations marks (!).
3	The student occasionally includes emotions or personal emotional statements in the peer review. They could be quite strong. However, the student's review is missing personal pronouns ("I", "You") and is mostly written in third person. Emotions can both be positive or negative. Negative emotions can be demonstrated with concern, missing understanding or insecurity (e.g., with modal verbs or words such as rather, perhaps).
2	Mostly, the student does not respond emotionally to the peer's work. Only very minor and weak emotions or personal emotional statements are integrated. The student writes mostly objectively (e.g., "Okay", "This should be added", "The task was done correctly", etc.). In comparison to level 1, she might be using modal verbs (might, could, etc.) or words to show insecurity in her feedback (rather, maybe, possibly).
1	The student does not respond emotionally to the peer's work at all. She does not show her feelings towards the peer and writes objectively (e.g., no "I feel", "Personally" "I find this..." and no emotions such as "good", "great", "fantastic", "concerned", etc.). Typical examples would be "Add a picture." or "The value gap XY is missing.".

Table 2: Description of the emotional empathy scores.

and Gurevych (2017), our guidelines consisted of 14 pages, including definitions and rules for how the review components should be composed, which annotation scheme was to be used, and how the cognitive and emotional empathy level were to be judged. Several individual training sessions and eight team workshops were performed to resolve disagreements among the annotators and to reach a common understanding of the annotation guidelines on the cognitive and emotional empathy structures. We used the *tagtog* annotation tool,[5] which offers an environment for cloud-based annotation in a team. First, a text was classified into peer review components (*strengths, weaknesses, suggestions for improvement*, or *none*) by the trained annotators. Second, the same annotator then scored the cognitive and emotional empathy levels of each component based on our annotation guideline on a one to five scale. After the first 92 reviews were annotated by all three annotators, we calculated the inter-annotator agreement (IAA) scores (see Section 4.1).[6] As we obtained satisfying results, we proceeded with two annotators annotating 130 remaining documents each and the senior annotator annotating 148 peer reviews, resulting in 408 additional annotated documents. Together with the 92 annotations of the annotation study of the senior annotator (the annotator with the most reviewing experience), we counted 500 annotated documents in our final corpus.

4 Corpus Analysis

4.1 Inter-Annotator Agreement

To evaluate the reliability of the review components and empathy level annotations, we followed the approach of Stab and Gurevych (2014).

[5] https://tagtog.net/

[6] Our intention was to capture the annotation of 100 randomly selected essays. However, we discarded 8 of the 100 essays as they contained less than 2 review components.

Review Components Concerning the review components, two strategies were used. Since there were no predefined markables, annotators not only had to identify the *type of review component* but also its *boundaries*. In order to assess the latter, we use Krippendorff's α_{U} (Krippendorff, 2004), which allows for an assessment of the reliability of an annotated corpus, considering the differences in the markable boundaries. To evaluate the annotators' agreement in terms of the selected category of a review component for a given sentence, we calculated the percentage agreement and two chance-corrected measures, multi-π (Fleiss, 1971) and Krippendorff's α (Krippendorff, 1980). Since each annotation always covered a full sentence (or a sequence of sentences), we operated at the sentence level for calculating the reliability of the annotations in terms of the IAA.

	%	Multi-π	Kripp. α	Kripp. α_{U}
Strength	0.9641	0.8871	0.8871	0.5181
Weakness	0.8893	0.7434	0.7434	0.3109
Suggestions	0.8948	0.6875	0.6875	0.3512
None	0.9330	0.8312	0.8312	0.9032

Table 3: IAA of review component annotations.

Table 3 displays the resulting IAA scores. The obtained scores for Krippendorff's α indicated an almost perfect agreement for the *strengths* components and a substantial agreement for both the *weaknesses* and the *suggestions for improvement* components. The unitized α of strengths, weaknesses and suggestions for improvement annotations was slightly smaller compared to the sentence-level agreement. Thus, the boundaries of review components were less precisely identified in comparison to the classification into review components. Yet the scores still suggest that there was a moderate level of agreement between the annotators for the strengths and a fair agreement for the weaknesses and the suggestions for improvement. With a score of α_{U}=90.32%, the boundaries of the non-annotated text units were more reliably detected, indicating an almost perfect agreement between the annotators. Percentage agreement, multi-π, and Krippendorff's α were considerably higher for the non-annotated spans as compared to the strengths, weaknesses, and suggestions for improvement, indicating an almost perfect agreement between the annotators. Hence, we conclude that the annotation of the review components in student-written peer reviews is reliably possible .

Empathy Level To assess the reliability of the cognitive and emotional empathy level annotations, we calculated the multi-π for both scales. For the cognitive empathy level, we received a multi-π of 0.41 for both the emotional and cognitive empathy level, suggesting a moderate agreement between the annotators in both cases. Thus, we conclude that the empathy level can also be reliably annotated in student-generated peer reviews.

To analyze the disagreement between the three annotators, we created a confusion probability matrix (CPM) (Cinková et al., 2012) for the review components and the empathy level scores. The results can be found in Section C of the appendix.

4.2 Corpus Statistics

The corpus we compiled consists of 500 student-written peer reviews in German that were composed of 9,614 sentences with 126,887 tokens in total. Hence, on average, each document had 19 sentences and 254 tokens. A total of 2,107 strengths, 3,505 weaknesses and 2,140 suggestions for improvement were annotated.

Tables 4, 5, and 6 present some detailed statistics on the final corpus.

	total	mean	std dev	min	max	median
Sentences	9,614	19.23	10.39	1	85	17
Tokens	126,887	253.77	134.18	10	1026	228

Table 4: Distribution of *sentences* and *tokens* in the created corpus. Mean, std dev, min, max and median refer to the number of sentences and tokens per document.

	total	mean	std dev	min	max	median	%
Str.	2,107	4.21	2.71	1	20	4	0.27
Weak.	3,505	7.01	6.10	0	41	5	0.45
Sug.	2,140	4.28	5.49	0	59	3	0.28

Table 5: Distribution of the *review components*.

	mean	std dev	min	max	median
Cognitive EL	2.94	0.99	1	5	3
Emotional EL	3.22	1.03	1	5	3

Table 6: Distribution of the *empathy level (EL) scores*.

Moreover, Figure 3 displays the distribution of the empathy scores in the annotated dataset. Both the cognitive and the emotional empathy levels approximately follow a normal distribution with a mean score of 2.94 and 3.22, respectively (see Table 6). We measured only a low correlation of 0.38 between the scores of cognitive and emotional empathy.

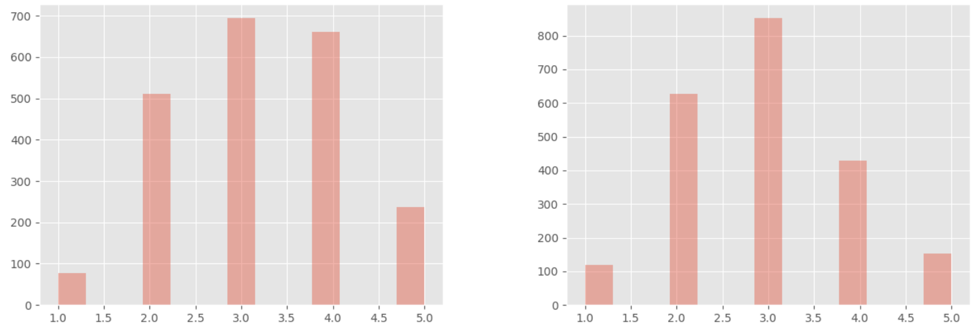


Figure 3: Distribution of the cognitive *(left)* and emotional *(right)* empathy scores (1-5 scale).

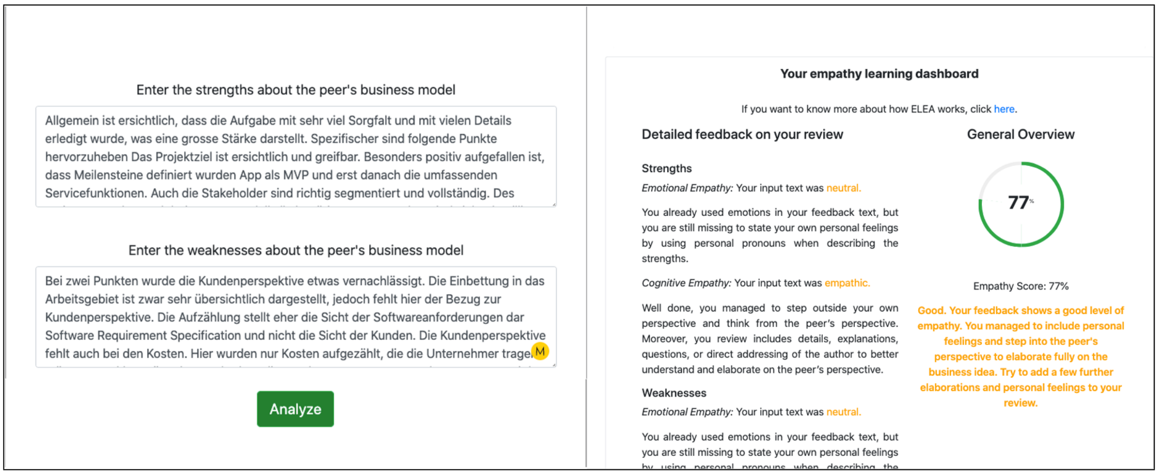
5 Providing Students Adaptive Feedback

Modelling Cognitive and Emotional Empathy The empathy detection task is considered a paragraph-based, multi-class classification task, where each paragraph is either considered to be a *strength, weakness*, or a *suggestion for improvement* and has a "non-empathic", "neutral", or "empathic" cognitive and emotional empathy level. Therefore, we assigned the levels of our cognitive and emotional empathy scores to three different labels: level 1 and 2 were assigned to a "non-empathic" text label, level 3 to a "neutral" label, and levels 4 and 5 to an"empathic" label . We split the data into 70% training, 20% validation, and 10% test data. To apply the model, the corpus texts were split into word tokens. The model performances were measured in terms of accuracy, precision, recall, and f1-score.

We trained a predictive model following the architecture of Bidirectional Encoder Representations from Transformers (BERT) proposed by Devlin et al. (2018). We used the BERT model from *deepset*,[7] since it is available in German and provides a deep pretrained model that was unsupervised while training on domain-agnostic German corpora (e.g., the German Wikipedia). The best performing paramenter combination for our BERT model incorporated a dropout probability of 10% and a learning rate of 3e^{-5}, and the number of epochs were 3. After several iterations, we reached a micro f1-score of 74.96% for the detection of the emotional empathy level and 69.98% for the detection of the cognitive empathy level of a text paragraph. Moreover, we reached an f1-score of 94.83% to predict a text paragraph as a strength, a 64.28% to predict a text paragraph as a weakness, and 59.79% to predict suggestions for improvement. To ensure the validity of our BERT model, we benchmarked against bidirectional Long-Short-Term-Memory-Conditional-Random-Fields classifiers (BiLSTM-CRF). In combination with the corresponding embeddings vocabulary (GloVe) (Pennington et al., 2014), our LSTM reached an unsatisfying f1-score of 61% for detecting the emotional empathy level and 51% for detecting the cognitive empathy level.

Evaluation in a Peer Learning Setting We designed and built an adaptive writing support system that provides students with individual feedback on their cognitive and emotional empathy skills. The application is illustrated in Figure 4. We embedded our system into a peer writing exercise where students were asked to write a peer review on a business model. During this writing task, they received adaptive feedback on the cognitive and emotional empathy level based on our model. The evaluation was conducted as a web experiment facilitated by the behavioral lab of our university, and thus, designed and reviewed according to the ethical guidelines of the lab and the university. We received 58 valid results (mean age = 23.89, SD= 3.07, 30 were male, 28 female). The participants were told to read an essay about a business model of a peer student. Afterwards, they were asked to write a business model review for the peer by providing feedback on the strengths, weaknesses, and suggestions for improvement of the particular business model. After the treatment, we measured the intention to use (ITU) (Venkatesh and Bala, 2008) by asking three items. We also asked the participants to judge their perceived empathy skill learning (PESL) by asking two items that covered cognitive and emotional empathy skills (Spreng

[7] https://github.com/deepset-ai/FARM



Enter the strengths about the peer's business model

Allgemein ist ersichtlich, dass die Aufgabe mit sehr viel Sorgfalt und mit vielen Details erledigt wurde, was eine grosse Stärke darstellt. Spezifischer sind folgende Punkte hervorzuheben Das Projektziel ist ersichtlich und greifbar. Besonders positiv aufgefallen ist, dass Meilensteine definiert wurden App als MVP und erst danach die umfassenden Servicefunktionen. Auch die Stakeholder sind richtig segmentiert und vollständig. Des

Enter the weaknesses about the peer's business model

Bei zwei Punkten wurde die Kundenperspektive etwas vernachlässigt. Die Einbettung in das Arbeitsgebiet ist zwar sehr übersichtlich dargestellt, jedoch fehlt hier der Bezug zur Kundenperspektive. Die Aufzählung stellt eher die Sicht der Softwareanforderungen dar Software Requirement Specification und nicht die Sicht der Kunden. Die Kundenperspektive fehlt auch bei den Kosten. Hier wurden nur Kosten aufgezählt, die die Unternehmer trage

Your empathy learning dashboard

If you want to know more about how ELEA works, click here.

Detailed feedback on your review

Strengths

Emotional Empathy: Your input text was neutral.

You already used emotions in your feedback text, but you are still missing to state your own personal feelings by using personal pronouns when describing the strengths.

Cognitive Empathy: Your input text was empathic.

Well done, you managed to step outside your own perspective and think from the peer's perspective. Moreover, you review includes details, explanations, questions, or direct addressing of the author to better understand and elaborate on the peer's perspective.

Weaknesses

Emotional Empathy: Your input text was neutral.

You already used emotions in your feedback text, but you are still missing to state your own personal feelings by using personal pronouns when describing the

General Overview

Empathy Score: 77%

Good. Your feedback shows a good level of empathy. You managed to include personal feelings and step into the peer's perspective to elaborate fully on the business idea. Try to add a few further elaborations and personal feelings to your review.

Figure 4: Screenshot of a trained model on our corpus as an adaptive writing support system.

et al., 2009; Davis, 1983). Finally, we surveyed the perceived feedback accuracy (PFA) (Podsakoff and Farh, 1989) to control the accuracy of our model. All constructs were measured with a 1-to-7 point Likert scale (1: totally disagree to 7: totally agree, with 4 being a neutral statement).[8] Furthermore, we asked three qualitative questions: "*What did you particularly like about the use of the tool?*", "*What else could be improved?*", and "*Do you have any other ideas?*" and captured the demographics. In total, we asked 13 questions. All participants were compensated with an equivalent of about 12 USD for a 25 to 30 minute experiment.

Results Participants judged their empathy skill learning with a mean of 5.03 (SD= 1.05). Concerning the PFA, the subjects rated the construct with a mean of 4.93 (SD= 0.94). The mean value of intention to use of the participants using our application as a writing support tool in peer learning scenarios was 5.14 (SD= 1.14). The mean values of all three constructs were very promising when comparing the results to the midpoints. All results were better than the neutral value of 4, indicating a positive evaluation of our application for peer learning tasks. We also asked open questions in our survey to receive the participants' opinions about the tool they used. The general attitude was very positive. Participants positively mentioned the simple and easy interaction, the distinction between cognitive and emotional empathy feedback, and the overall empathy score together with the adaptive feedback message several times. However, participants also said that the tool should provide even more detailed feedback based on more categories and should provide concrete text examples on how to improve their empathy score. We translated the responses from German and clustered the most representative responses in Table 16 in the appendix.

6 Conclusion

We introduce a novel empathy annotation scheme and an annotated corpus of student-written peer reviews extracted from a real-world learning scenario. Our corpus consisted of 500 student-written peer reviews that were annotated for review components and their emotional and cognitive empathy levels. Our contribution is threefold: 1) we derived a novel annotation scheme for empathy modeling based on psychological theory and previous work for empathy modeling (Buechel et al., 2018); 2) we present an annotation study based on 92 student peer reviews and three annotators to show that the annotation of empathy in student peer reviews is reliably possible ; and 3) to the best of our knowledge, we present the second freely available corpus for empathy detection and the first corpus for empathy detection in the educational domain based on 500 student peer reviews in German. For future research, this corpus could be leveraged to support students' learning processes, e.g., through a conversational interaction (Zierau et al., 2020). However, we would also encourage research on the ethical considerations of empathy detection models in user-based research (i.e., Wambsganss et al. (2021a)). We, therefore, hope to encourage future research on student-generated empathetic texts and on writing support systems to train empathy skills of students based on NLP towards quality education independent of a student's location or instructors.

[8] The exact items are listed in the appendix.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems*, 26(3):1–34.

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:718–728.

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech and Language*, 50:40–61.

Simon Baron-Cohen and Sally Wheelwright. 2004. The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. Technical Report 2.

C. Daniel Batson, Jim Fultz, and Patricia A. Schoenrade. 1987. Distress and Empathy: Two Qualitatively Distinct Vicarious Emotions with Different Motivational Consequences. *Journal of Personality*, 55(1):19–39.

Emily Teding van Berkhout and John M. Malouff. 2016. The efficacy of empathy training: A meta-analysis of randomized controlled trials. *Journal of Counseling Psychology*, 63(1):32–41.

Paul Black and Dylan Wiliam. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1):5–31.

Laura Ana Maria Bostan and Roman Klinger. 2018. An Analysis of Annotated Corpora for Emotion Classification in Text Title and Abstract in German. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 4758–4765.

Sven Buechel and Udo Hahn. 2018. Emotion Representation Mapping for Automatic Lexicon Construction (Mostly) Performs on Human Level. pages 2892–2904.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao, Huang, and Lun-Wei Ku. 2018. EmotionLines: An Emotion Corpus of Multi-Party Conversations. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 1597–1601.

Silvie Cinková, Martin Holub, and Vincent Kríž. 2012. Managing uncertainty in semantic tagging. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 840–850, Avignon, France. Association for Computational Linguistics.

Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113–126.

Jean Decety and Philip L. Jackson. 2004. The functional architecture of human empathy.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Pierre Dillenbourg, Sanna Järvelä, and Frank Fischer. 2009. The Evolution of Research on Computer-Supported Collaborative Learning. In Nicolas Balacheff, Sten Ludvigsen, Ton de Jong, Ard Lazonder, and Sally Barnes, editors, *Technology-Enhanced Learning: Principles and Products*, pages 3–19. Springer Netherlands, Dordrecht.

Paul Ekman. 1992. An Argument for Basic Emotions. *COGNITION AND EMOTION*, 6(3/4):169–200.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research*, 77(1):81–112.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying Empathetic Messages in Online Health Communities. Technical report.

Sara H. Konrath, Edward H. O'Brien, and Courtney Hsing. 2011. Changes in dispositional empathy in American college students over time: A meta-analysis. *Personality and Social Psychology Review*, 15(2):180–198.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Methodology*. Sage Publications, Inc., Beverly Hills, CA.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality and Quantity*, 38(6):787–800.

E. J. Lawrence, P. Shaw, D. Baker, S. Baron-Cohen, and Anthony S. David. 2004. Measuring empathy: Reliability and validity of the Empathy Quotient. *Psychological Medicine*, 34(5):911–919.

Bing Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Benjamin Lok and Adriana E. Foster. 2019. Can Virtual Humans Teach Empathy? In *Teaching Empathy in Healthcare*, pages 143–163. Springer International Publishing.

Joseph Luca and Pina Tarricone. 2001. Does Emotional Intelligence Affect Successful Teamwork? *Proceedings of the 18th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education*, (December 2001):367–376.

Scott W. McQuiggan and James C. Lester. 2007. Modeling and evaluating empathy in embodied companion agents. *International Journal of Human Computer Studies*, 65(4):348–360.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. **SEM 2017 - 6th Joint Conference on Lexical and Computational Semantics, Proceedings*, pages 65–77.

Tb Moyers and T Martin. 2010. Revised Global Scales: Motivational Interviewing Treatment Integrity 3.1.1 (MITI 3.1.1). *University of New . . .*, 1(January):1–29.

OECD. 2018. The Future of Education and Skills - Education 2030.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1532–1543. Association for Computational Linguistics (ACL).

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:1426–1435.

Philip M. Podsakoff and Jiing Lih Farh. 1989. Effects of feedback sign and credibility on goal setting and task performance. *Organizational Behavior and Human Decision Processes*, 44(1):45–67.

Mathis Poser and Eva A. C. Bittner. 2020. Hybrid Teamwork: Consideration of Teamwork Concepts to Reach Naturalistic Interaction between Humans and Conversational Agents. In *WI2020*. GITO Verlag.

Roman Rietsche and Matthias Söllner. 2019. Insights into Using IT-Based Peer Feedback to Practice the Students Providing Feedback Skill. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Carolyn Rosé, Yi Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3(3):237–271.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2018. SemEval-2017 Task 4: Sentiment Analysis in Twitter. pages 502–518.

Breno Santana Santos, Methanias Colaqo Junior, and Janisson Gois De Souza. 2018. An Experimental Evaluation of the NeuroMessenger: A Collaborative Tool to Improve the Empathy of Text Interactions. *Proceedings - IEEE Symposium on Computers and Communications*, 2018-June:573–579.

Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Julia E. Seaman, I. E. Allen, and Jeff Seaman. 2018. Higher Education Reports - Babson Survey Research Group. Technical report.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. pages 5263–5276.

R. Nathan Spreng, Margaret C. McKinnon, Raymond A. Mar, and Brian Levine. 2009. The Toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment*, 91(1):62–71.

Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers ,*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. *ACL 2007 - SemEval 2007 - Proceedings of the 4th International Workshop on Semantic Evaluations*, (June):70–74.

Stephen E. Toulmin. 1984. *Introduction to Reasoning*.

Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, 39(2):273–315.

Thiemo Wambsganss, Anne Höch, Naim Zierau, and Matthias Söllner. 2021a. Ethical Design of Conversational Agents: Towards Principles for a Value-Sensitive Design. In *Proceedings of the 16th International Conference on Wirtschaftsinformatik (WI)*.

Thiemo Wambsganss, Tobias Küng, Matthias Söllner, and Jan Marco Leimeister. 2021b. ArgueTutor: An

Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Jan Marco Leimeister, and Siegfried Handschuh. 2020a. AL : An Adaptive Learning Support System for Argumentation Skills. In *ACM CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020b. A corpus for argumentative writing support in German. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 856–869, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thiemo Wambsganss, Florian Weber, and Matthias Söllner. 2021c. Design and Evaluation of an Adaptive Empathy Learning Tool. In *Hawaii International Conference on System Sciences (HICSS)*.

Jin Wang, Liang Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers*, pages 225–230.

Bo Xiao, Dogan Can, Panayiotis G Georgiou, David Atkins, and Shrikanth S Narayanan. 2012. Analyzing the Language of Therapist Empathy in Motivational Interview based Psychotherapy. *Signal and Information Processing Association Annual Summit and Conference (APSIPA), ... Asia-Pacific. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012.

N. Zierau, T Wambsganss, Andreas Janson, Sofia Schöbel, and Jan Marco Leimeister. 2020. The Anatomy of User Experience with Conversational Agents : A Taxonomy and Propositions of Service Clues. In *International Conference on Information Systems (ICIS).*, pages 1–17.

A Details on the Description of the Annotation Scheme[9]

A more detailed description of the cognitive and emotional empathy scores can be found in Table 7 and Table 8.

B Details on the Annotation Process

The annotation process was split into three steps:

1. **Reading of the entire peer review:** The annotators are confronted with the student-written peer review and are asked to read the whole document. This helps to get a first impression of the review and get an overview of the single components and the structure of it.
2. **Labeling the components and elaborations:** After reading the entire student-written peer review, the annotator is asked to label the three different components (*strengths, weaknesses* and *suggestions for improvement*). Every supporting sentence (such as explanation, example, etc.) is annotated together with the referred component.
3. **Classification of the cognitive and emotional empathy levels:** Each component is assessed on its level of cognitive and emotional empathy by giving a number between 1-5. Each category is carefully defined and delimited according to Table 7 and Table 8.

C Disagreement Analysis

To analyze the disagreement between the three annotators, we created a confusion probability matrix (CPM) (Cinková et al., 2012) for the review components and the empathy level scores. A CPM contains the conditional probabilities that an annotator assigns to a certain category (column) given that another annotator has chosen the category in the row for a specific item. In contrast to traditional confusion matrices, a CPM also allows for the evaluation of confusions if more than two annotators are involved in an annotation study (Stab and Gurevych, 2014).

Table 9 shows that there is a broad agreement between the annotators in distinguishing between the different types of review components. The major disagreement is between suggestions and weaknesses, though with a score of 60%, the agreement is still fairly high. Consequently, the annotation of review components in terms of strengths, weaknesses, and suggestions for improvements yields highly reliable results.

The CPMs for the empathy levels (see Tables 10 and 11 reveal that there is a higher confusion between the scores assigned by the three reviewers, as compared to the annotation of the review components. However, when analyzed more closely, one can see that the scores mostly vary only within a small window of two or three neighboring scores. Therefore, we conclude that the annotation of cognitive and emotional empathy scores is reliably possible, too.

[9] Further examples and descriptions can be found in our annotation guideline.

Score	Description
5 = strong	The student fully understands the peer's thoughts. She completely steps outside her own perspective and thinks from the peer's perspective. She does that by carefully evaluating the peer's idea with rich explanations. Questions, personal pronouns, or direct addressing of the author can be used in order to better understand and elaborate on the peer's perspective. *Strengths:* The student fully grasps the idea of the peer. She elaborates on strengths that are important for the peer for her continuation of the task and adds explanations, thoughts, or examples to her statements and reasons why the strength is/strengths are important for the business idea. *Weaknesses:* The student thinks completely from the peer's perspective and what would help him/her to further succeed with the task. The student explains the weakness in a very detailed manner and describes why the weakness is important to consider. He can also give counterarguments or ask questions to illustrate the weakness. *Suggestions for improvement:* The student suggests improvements as if he were in the peer's position in creating the best possible solution. The student completes his suggestions with rich explanations on why he/she would do so and elaborates on the improvements in a very concrete and detailed way. Almost every suggestion is supported by further explanations.
4 = fairly strong	The student thinks from the perspective of the peer. She elaborates in a way that serves the peer best to further establish the idea or activity. Each component is affirmed with further explanations. *Strengths:* The student is able to recognize one or more strengths that are helpful for the peer to affirm their business idea and activity. He/She highlights contextual strengths rather than formal strengths. The student supports most statements with examples or further personal thoughts on the topic but might still be missing some reasonings. *Weaknesses:* The student thinks from the peer's perspective and what would help him/her to further succeed with the task. This could be demonstrated by stating various questions and establishing further thoughts. The student explains the weakness and adds examples, but he/she is still missing some reasonings. *Suggestions for improvement:* The student suggests one or more improvements that are relevant for the further establishment of the activity and idea from the perspective of the peer. Most suggestions are written concretely and, if applicable, supported by examples. In most cases, the student explains why he/she suggests a change.
3 = slightly weak / equal	The student tries to understand the perspective of the peer and adds further elaborations to her statements. However, her elaborations are not completely thought through and her feedback is missing some essential explanations, examples, or questions to make sure she understood everything correctly. *Strengths:* The student mentions one or more strengths and explains some of them with minor explanations or examples on why it is seen as a strength. However, most strengths focus on formal aspects rather than contextual aspects. *Weaknesses:* The student states one or more weaknesses and explains some of them with minor explanations or examples. The student could also just state questions to illustrate the weakness in the peer's business idea. Most weaknesses are not explained why they are such. *Suggestions from improvements:* The student suggests one or more improvements that are mostly relevant for the further establishment of the activity. The suggestions are written only on a high-level and most of them do not include further explanations or examples. The student explains only occasionally why he/she suggests a change or how it could be implemented.
2 = very weak	The student does not try to understand the peer's perspective. The student rather just tries to accomplish the task of giving feedback. *Strengths:* The student mentions one or more strengths. They could be relevant for the peer. However, he does not add any further explanation or details. *Weaknesses:* The student states one or more weaknesses without explaining why they are seen as such. They could be relevant for the peer. However, the statements do not include any further elaboration on the mentioned weakness. *Suggestions for improvement:* The student suggests one or more improvements that could be relevant for the peer. However, the student does not explain why he/she suggests the change or how the suggestions for improvement could be implemented.
1 = absolutely weak	The student's feedback is very short and does not include the peer's perspective. She does not add any further elaboration in her thoughts. *Strengths:* The student only mentions one strength. This might not be relevant at all and lacks any further explanation, detail, or example. *Weakness:* The student only mentions one weakness. This might not be relevant at all and lacks any further explanation, detail, or example. *Suggestions for improvement:* The student only mentions one suggestion. The suggestion is not followed by any explanation or example and might not be relevant for the further revision of the peer.

Table 7: Detailed description of the cognitive empathy scores.

Score	Description
5 = strong	The student is able to respond very emotionally to the peer's work and fully represents the affectional state in her entire review. She illustrates this by writing in a very emotional and personal manner and expresses her feelings (positive or negative) throughout the review. Strong expressions include exclamation marks (!). Typical feedback in this category includes phrases such as "brilliant!", "fantastic", "excellent", "I am totally on the same page as you", "I am very convinced", "Personally, I find this very important, too", "I am very unsure", "I find this critical", "I am very sure you feel", "This is compelling for me", etc.
4 = fairly strong	The student is able to respond emotionally to the peer's submitted activity with suitable emotions (positive or negative). She returns emotions in her feedback on various locations and expresses her feelings by using the personal pronoun ("I", "You"). Some sentences might include exclamations marks (!). Typical feedback in this category includes phrases such as "I am excited", "This is very good!", "I am impressed by your idea", "I feel concerned about", "I find this very...", "In my opinion", "Unfortunately, I do not understand", "I am very challenged by your submission", "I am missing", "You did a very good job", etc.
3 = slightly weak / equal	The student occasionally includes emotions or personal emotional statements in the peer review. They could be quite strong. However, the student's review is missing personal pronouns ("I", "You") and is mostly written in third person. Emotions can both be positive or negative. Negative emotions can be demonstrated with concern, missing understanding or insecurity (e. g., with modal verbs or words such as rather, perhaps). Typically, scale 3 includes phrases such as "it's important", "the idea is very good", "the idea is comprehensible", "it would make sense", "the task was done very nicely", "It could probably be that", etc.
2 = very weak	Mostly, the student does not respond emotionally to the peer's work. Only very minor and weak emotions or personal emotional statements are integrated. The student writes mostly objectively (e.g., "Okay", "This should be added", "The task was done correctly", etc.). In comparison to level 1, she might use modal verbs (might, could, etc.) or words to show insecurity in her feedback (rather, maybe, possibly).
1 = absolutely weak	The student does not respond emotionally to the peer's work at all. She does not show her feelings towards the peer and writes objectively (e.g., no "I feel", "personally" "I find this.." and no emotions, such as "good", "great", "fantastic", "concerned", etc.). Typical examples would be "Add a picture." or "The value gap XY is missing."

Table 8: Detailed description of the emotional empathy scores.

	Suggestions	Weakness	Strength	None
Suggestions	**0.6056**	0.2970	0.0214	0.0759
Weakness	0.2139	**0.7009**	0.0203	0.0648
Strength	0.0264	0.0347	**0.8340**	0.1049
None	0.0662	0.0784	0.0742	**0.7812**

Table 9: CPM for review component annotations.

	1	2	3	4	5
1	**.113**	.387	.175	.165	.160
2	.125	**.266**	.362	.211	.035
3	.025	.159	**.223**	.482	.112
4	.014	.054	.283	**.300**	.349
5	.021	.014	.105	.556	**.303**

Table 10: CPM for cognitive empathy level annotations.

	1	2	3	4	5
1	**.106**	.459	.286	.086	.063
2	.154	**.234**	.455	.128	.029
3	.059	.282	**.350**	.240	.068
4	.026	.115	.347	**.295**	.218
5	.043	.061	.227	.501	**.168**

Table 11: CPM for emotional empathy level annotations.

	precision	recall	f1-score	support
non-empathic	0.5746	0.5662	0.5704	136
empathic	0.6364	0.5625	0.5972	112
neutral	0.5240	0.5707	0.5464	191
None	0.9863	0.9729	0.9795	295
micro avg	0.7322	0.7302	0.7482	734
macro avg	0.6803	0.6681	0.6734	734
weighted avg	0.7363	0.7302	0.7327	734
samples avg	0.7248	0.7302	0.7266	734

Table 12: BERT model results for emotional empathy.

	precision	recall	f1-score	support
non-empathic	0.5739	0.3587	0.4415	184
empathic	0.6434	0.5490	0.5925	286
neutral	0.3062	0.4747	0.3723	198
None	0.9841	0.9802	0.9822	506
micro avg	0.6949	0.6925	0.6937	1174
macro avg	0.6269	0.5907	0.5971	1174
weighted avg	0.7225	0.6925	0.6996	1174
samples avg	0.6861	0.6925	0.6882	1174

Table 13: BERT model results for cognitive empathy.

D Details on Application and Evaluation of Writing Support Tool

To ensure the validity of our BERT model, we benchmarked against bidirectional Long-Short-

	precision	recall	f1-score	support
non-empathic	0.5739	0.3587	0.4415	184
neutral	0.3062	0.4747	0.3723	198
empathic	0.6434	0.5490	0.5925	286
None	0.9841	0.9802	0.9822	506
f1 avg	0.64	0.64	0.64	368
weighted avg	0.73	0.73	0.73	368

Table 14: Results for the LSTM for emotional empathy.

	precision	recall	f1-score	support
non-empathic	0.74	0.28	0.40	83
neutral	0.43	0.55	0.49	60
empathic	0.35	0.63	0.45	57
None	0.99	0.94	0.97	168
f1 avg	0.63	0.60	0.58	368
weighted avg	0.75	0.68	0.68	368

Table 15: Results for the LSTM for cognitive empathy.

Term-Memory-Conditional-Random-Fields classifiers (BiLSTM-CRF). In combination with the corresponding embeddings vocabulary (GloVe) (Pennington et al., 2014), our LSTM reached an unsatisfying f1-score of 61% for detecting the emotional empathy level and 51% for detecting the cognitive empathy level.

More information on the results of our BERT model and the LSTM for emotional and cognitive empathy detection can be found in the Tables 12, 13, 15, and 15.

In the post-survey, we measured perceived usefulness following the technology acceptance model (Venkatesh and Bala, 2008). The items for the constructs were: ”*Imagine the tool was available in your next course, would you use it?*”, ”*Assuming the learning tool would be available at a next course, I would plan to use it.*”, or ”*Using the learning tool helps me to write more emotional and cognitive empathic reviews.* ” Moreover, we asked the participants to judge their perceived empathy skill learning (PESL) by asking two items that cover cognitive and emotional empathy skills (Spreng et al., 2009; Davis, 1983): *“I assume that the tool would help me improve my ability to give appropriate emotional feedback.”* and *“I assume that the tool would help me improve my ability to empathize with others when writing reviews.”* Finally, we surveyed the perceived feedback accuracy (PFA) (Podsakoff and Farh, 1989) of both learning tools by asking three items: *“The feedback I received reflected my true performance.”*, *“The tool accurately evaluated my performance.”*, and *“The feedback I received from the tool was an accurate evaluation of my performance”*. All constructs were measured with a 1- to 7-point Likert scale (1: totally disagree to 7: totally agree, with 4 being a neutral statement).

Cluster	**Feature**
On empathy feedback reaction	*”I think that this tool could help me not only to put myself in the position of a person in terms of content and make suggestions but also to communicate to them better”*
On the feedback for skill learning	*”The empathy feedback was clear and could be easily implemented. I had the feeling I learned something.Would use it again!”*
On cognitive and emotional empathy	*”It was helpful that a distinction was made between the two categories of empathy. This again clearly showed me that I do not show emotional empathy enough. It was also useful that the tool said how to show emotional empathy (feelings when reading the business idea etc.).”*
Improvements on feedback granularity	*”It would be better if the feedback was more s elective or with detailed categories about empathy.”*
Improvements on feedback recommendations	*”Even more detailed information on how I can improve my empathy writing would be helpful, e.g., with review examples.”*

Table 16: Representative examples of qualitative user responses after the usage of our empathy support tool.