Modeling Persuasive Discourse to Adaptively Support Students' Argumentative Writing

Thiemo Wambsganss
University of St.Gallen/ CH
Carnegie Mellon University/ US
thiemo.wambsganss@unisg.ch

Christina Niklaus
University of St.Gallen/ CH
christina.niklaus@unisg.ch

Abstract

We introduce an argumentation annotation approach to model the structure of argumentative discourse in student-written business model pitches. Additionally, the annotation scheme captures a series of persuasiveness scores such as the specificity, strength, evidence, and relevance of the pitch and the individual components. Based on this scheme, we annotated a corpus of 200 business model pitches in German. Moreover, we trained predictive models to detect argumentative discourse structures and embedded them in an adaptive writing support system for students that provides them with individual argumentation feedback independent of an instructor, time, and location. We evaluated our tool in a real-world writing exercise and found promising results for the measured self-efficacy and perceived ease-of-use. Finally, we present our freely available corpus of persuasive business model pitches with 3,207 annotated sentences in German language and our annotation guidelines.

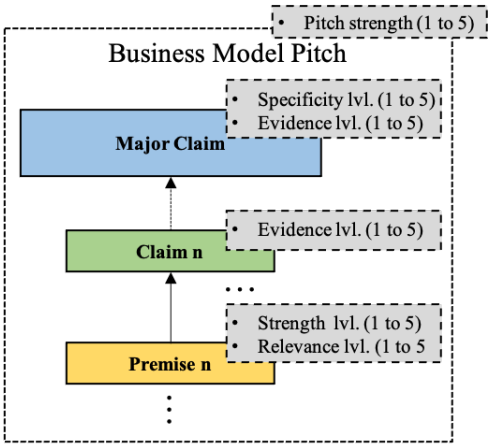

Figure 1: Argumentation annotation scheme. First, a text sentence is classified into an argumentative component (*claim, premise, major claim, none-argumentative*). Second, the same annotator captures the basic discourse structure between the components. Third, the components and the pitch are scored for the persuasiveness scores (specificity, evidence, strength, relevance) based on our annotation guideline on a 1-to-5 scale.

1 Introduction

Argumentation is an omnipresent rudiment of daily communication and thinking (Kuhn, 1992; Toulmin, 1984). The ability to form convincing arguments is not only fundamental to persuading an audience of novel ideas but also plays a major role in strategic decision-making, negotiation, and constructive civil discourse (Walton et al., 2008; Scheuer et al., 2010). However, humans often struggle to develop argumentation skills owing to a lack of individual and instant feedback in their learning process (Dillenbourg et al., 2009; Hattie and Timperley, 2007), since providing feedback on the individual argumentation skills of learners is time-consuming and not scalable if conducted manually by educators (OECD, 2018; Wambsganss et al., 2020b). Furthermore, novel distance learning scenarios such as massive open online courses (MOOCs) (Seaman et al., 2018) come with additional barriers related to individual feedback on a learner's argumentation.

One possible solution to this dilemma are adaptive argumentation support systems that enable individuals to train their argumentation skills, e.g., in collaborative learning settings (Dillenbourg et al., 2009) or by providing tailored argumentation feedback independent of an instructor, time and place (Wambsganss et al., 2020b, 2021). Such tools are increasingly utilizing recent developments in computational linguistics in the form of computer-assisted writing (Rosé et al., 2008) to provide tailored feedback about textual documents (Song et al., 2014; Stab and Gurevych, 2014a). In this context, Argumentation Mining (AM) research is a crucial field for the development of support systems that identify arguments in unstructured texts (Lippi and Torroni, 2015; Lawrence and Reed, 2019).

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8748 - 8760

However, corpora that are applicable for the design and development of adaptive argumentative writing systems in pedagogical scenarios are rather scarce. To the best of our knowledge, there are only two collections from the educational domain which are based on student-written texts and annotated for argumentative discourse structures (Stab and Gurevych, 2017a; Wambsganss et al., 2020c).

We propose a novel argumentation annotation scheme for persuasive student-written business model pitches. Therefore, we introduce a corpus of 200 student-written persuasive pitches with 3,207 sentences that are annotated for argument components, their relations, and persuasiveness scores to judge the argumentation quality of the single arguments. We trained different models and embedded them as feedback algorithms in a novel writing support tool that provides students with individual argumentation feedback and recommendations in a persuasive writing exercise. The design of our tool is based on the self-evaluation mechanism for students to improve self-efficacy and argumentation learning outcomes during a learning process (i.e., self-regulated learning theory Bandura (1991); Zimmerman and Schunk (2001)). We asked students to conduct a persuasive writing exercise and provided them with argumentation self-evaluation. The measured argumentation (Toulmin, 2003), the perceived self-efficacy (Bandura, 1991), and the perceived usefulness (Venkatesh and Bala, 2008) in an evaluation provided promising results for using our approach in different large-scale learning scenarios to offer quality education with individual feedback independent of an instructor, time, and location.

Hence, we contribute to research by (1) deriving an annotation scheme for a new data domain for AM based on argumentation theory and previous work on annotation schemes for student-written texts (Stab and Gurevych, 2017a; Carlile et al., 2018; Wambsganss et al., 2020c), (2) presenting an annotation study based on 50 persuasive business model pitches and two annotators to show that the annotation of student-written pitches is reliably possible, (3) offering our final and freely available corpus of 200 student business pitches consisting of 3,207 annotated sentences collected from a lecture about digital business models in German, and (4) embedding and evaluating our annotation approach as predictive models in a writing support system in a real-world writing exercise. We, therefore, hope to encourage future research on argumentation discourse and persuasiveness levels in student-written texts and on writing support systems for argumentation.

2 Related Work

Argumentation Mining AM aims to identify argument components in the form of claims and premises, along with support and attack relationships that model the discourse structure of arguments. In recent years, this has been done for several domains, including legal texts (Mochales Palau and Ieven, 2009), newswire articles (Deng and Wiebe, 2015; Sardianos et al., 2015), or user-generated content (Wachsmuth et al., 2014; Habernal and Gurevych, 2015). The objective is to automatically identify arguments in unstructured textual documents based on the classification of argumentative and non-argumentative text units and the extraction of argument components and their relations. Recently, researchers have built increasing interest in adaptive argumentation support tools based on AM (Song et al., 2014; Stab and Gurevych, 2014a,b; Wambsganss et al., 2020b), offering argumentative writing support to students by providing individual feedback about the argumentation discourse. However, utilizing this technology in a pedagogical scenario for educational purposes lacks a wider-scale adoption (Stab and Gurevych, 2017b; Lawrence and Reed, 2019; Rosé et al., 2008), as argumentation-annotated corpora with student-written texts are rather rare (Lawrence and Reed, 2019; Wambsganss et al., 2020c).

Annotation Schemes and Corpora Since the availability of annotated data sets is crucial for designing, training, and evaluating AM algorithms, several research groups have dealt with creating labeled corpora, such as the Araucaria corpus (Reed et al., 2008), the European Court of Human Rights (ECHR) corpus (Mochales and Moens, 2008), or the Debatepedia corpus (Cabrio and Villata, 2012). Creating gold standards and test collections requires a formal representation model as well as corresponding annotation guidelines. While a number of well-defined models exist in the field of AM (e.g., Freeman (2001); Walton (1996); Wambsganss et al. (2020a), there is no general argumentation annotation scheme across all domains and genres of texts. Instead, the proposed representations differ in granularity, expression power, and categorization (Lawrence and Reed, 2019). There-

fore, conducting annotation studies with several annotators when introducing new annotation schemes is crucial for the quality of argumentation corpora.

Annotated Corpora for Education With the exception of the corpora proposed in Stab and Gurevych (2014a, 2017a) and Wambsganss et al. (2020c), prior argument-annotated data sets are not easily applicable for the development of argumentative writing support systems for students in a real-world case. The reasons are twofold. First, the texts are not extracted from a pedagogical scenario in which the annotation allows for training a model that provides students with individual and reliable feedback on the texts. Second, the data is often not annotated at the level of discourse (Stab and Gurevych, 2017a; Lawrence and Reed, 2019), which is necessary, for example, to give students feedback on insufficiently supported claims. Stab and Gurevych (2014a) identified the lack of linguistic corpora in the domain of student-written texts for designing and developing argumentative writing support systems by leveraging AM (Stab and Gurevych, 2014a). Therefore, they introduced an annotation scheme for annotating argument components and their relationships in persuasive English student essays. Afterwards, several researchers built on their corpus, including, e.g., Carlile et al. (2018), who use a subset of the essays and annotate their persuasiveness, and Ke et al. (2018), who train a persuasiveness scoring model on them. Recently, Wambsganss et al. (2020c) published an argumentation annotation scheme to capture the discourse level of student-written peer reviews. This corpus was successfully embedded in a writing support tool to provide students with adaptive argumentation tutoring (Wambsganss et al., 2020b). Building on the potential of argumentation-annotated corpora for adaptive skill learning, we propose to further transfer argumentation corpora to other educational domains and student-written texts.

3 Corpus Construction

Our corpus consists of 200 student-written business model pitches in which students present an entrepreneurial idea of a digital business model. Business model pitches - also called entrepreneurial or business pitches (Sabaj et al., 2020) - are described as *"a brief description of the value proposition of an idea or company"* (Daly and Davy, 2016) with the objective to convince a group of stakeholders of the novelty of an idea.

The formulation of persuasive business model pitches is increasingly used in modern pedagogical scenarios, e.g., to train the entrepreneurship mindset or agile work (i.e., OECD (2019)). Students are asked to write a concise but persuasive summary of the *"what, why, and how"* of their (business) idea in order to convince a peer. This pedagogical scenario is domain-independent, easy to implement in different settings (e.g., in MOOCs), and can be utilized to train skills such as logical argumentation. In fact, in their study about entrepreneurial business pitches, Fernández-Vázquez and Álvarez-Delgado (2019) found out that *"the lack of rational arguments determines the failure of the entrepreneur's efforts to be persuasive, regardless of the emotional appeals that are introduced into the pitch"*. Therefore, Fernández-Vázquez and Álvarez-Delgado (2019) calls for more emphasis on logical argumentation chains in business pitches.

However, linguistic research on business model pitches is a growing but still small field (Ducasse, 2020). Therefore, it is not surprising that no pitch corpus exists that is annotated for argumentation discourse structures based on an appropriate argumentation scheme (Lawrence and Reed, 2019). We propose a new annotation scheme to model argument components, their relations as well as argumentation quality labels that reflect the argumentative discourse structures in persuasive business model pitches. We based our annotation scheme on the model of Toulmin (1984) and the studies of Stab and Gurevych (2014a, 2017a); Wambsganss et al. (2020c); Carlile et al. (2018); Ke et al. (2019).

Following a 4-step methodology to build a robust corpus, we (1) searched literature and scientific theory on argumentation discourse structures and argumentation models in different text domains; (2) randomly sampled 50 student-written business pitches and, based on our findings from step 1, developed a set of annotation guidelines consisting of rules and limitations on how to annotate argumentation discourse structures; (3) applied, evaluated and improved our guidelines with three native speakers in five consecutive workshops to resolve annotation ambiguities; and (4) applied the final annotation scheme based on our 26-page guideline to a corpus of 200 student-written business pitches with 3,207 annotated sentences.[1]

[1]The annotation guidelines as well as the entire corpus can be accessed at https://github.com/thiemowa/-argumentative_business_model_pitches.

3.1 Data Source

We gathered a corpus of 200 student-written business model pitches in German. The data was collected in a mandatory business model innovation lecture at a Western European university. In this lecture, around 200 students develop and present a new business model. Students are asked to write a concise but persuasive pitch about the *"what, why, and how"* of their novel business idea in order to convince peer students. Afterwards, the students receive peer feedback from three fellow students on the persuasiveness of their business model pitch. The business pitches were collected from 2019 to 2020 according to the ethical guidelines of our university and with approval from the students to utilize the writings for scientific purposes.

3.2 Annotation Scheme

Our objective is to model the argumentation discourse structures and the persuasiveness of student-written business model pitches by capturing argument components, their relations, and persuasiveness scores. The majority of the pitches in our corpus follow the same structure. They describe a novel business model and then provide convincing statements backed by examples, statistics, user-centered descriptions, quotes, or intuitions. However, we found that the specificity, the strength, the relevance, and the evidence level vary between the different components. Thus, we captured them with qualitative labels on a 1-to-5 scale. Our basic annotation scheme is illustrated in Figure 2.

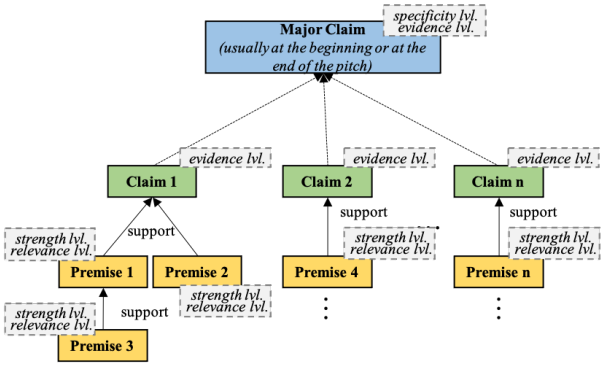

Figure 2: Overview of our argumentation annotation scheme for business model pitches, including argument components *(major claim, claim, premise)*, argumentative relations *(support)*, and persuasiveness scores *(specificity, evidence, strength, relevance)*.

Argument Components For argument components, we follow established models in argumentation theory which provide detailed definitions of argument components (e.g., Toulmin (1984); Stab and Gurevych (2017a)). These theories generally agree that a basic argument consists of multiple components and that it includes a *claim* that is supported or attacked by at least one *premise*. Also in student-written business model pitches, we found that a *claim* is the central component of an argument. It is a controversial statement (e.g., claiming a strength or novelty of a business model) that is either true or false and should not be accepted by the stakeholder without additional support or backing. In business model pitches, authors usually start or conclude with an overall idea and topic of the business model. Similar to the persuasive student essays corpus by Stab and Gurevych (2017a), we modeled this statement as a *major claim*. Usually, the major claim is present in the introduction or conclusion of the pitch - or in both. In the introduction, it often represents a general claim of the novelty of the business idea, whereas in the conclusion the major claim often summarizes or repeats the argumentation according to the author's business model idea. The major claim is then backed up by several other claims to manifest its validity. The *premise* supports the validity of the *claim* (e.g., by providing a statistic, analogy, user-centered example, or a value-based intuition). It is a reason given by the author to persuade the reader of their *claim*. Figure 3 illustrates a fully annotated example.[2]

Argumentative Relations The basic discourse structure in our data set of student-written business model pitches consists of one major claim and several claims, each independently supported by one or more premises. Since in our domain the writers aim to pitch their business idea as convincingly as possible, the texts generally do not include attack relations between the components, as is the case, for example, in student-written peer reviews (Wambsganss et al., 2020c). Therefore, we modeled and annotated only *support* relationships. Nevertheless, more complicated constellations of major claims, claims, and premises are possible. For example, a claim may be supported by several different premises or by a chain of premises in which each premise is in turn supported by another premise. In the same way, a claim can be supported by one premise. However, the simplest form consists of a major claim, backed up by a

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

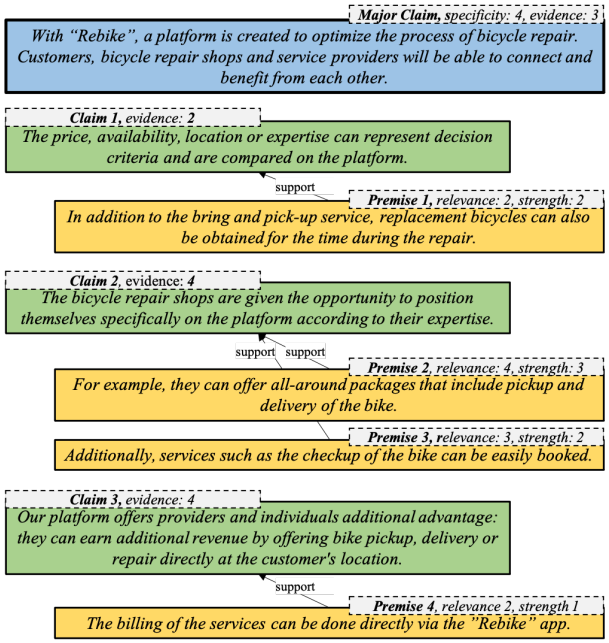

Figure 3: Annotated example of a pitch.

claim supported by a single premise. To provide an overview, we illustrated three basic examples of annotated relations in our corpus in the appendix.

Persuasiveness Scores To capture the differences in the persuasiveness levels of the components (i.e., the strength of a premise or the specificity of a major claim), we followed the approach of Carlile et al. (2018) and Ke et al. (2018) and defined five persuasiveness scores for the argumentative components (see Figure 1). Our objective was to capture the differences of a very persuasive major claim vs. a not very persuasive major claim accurately to provide students with more detailed writing support about *why* their argumentation is (un)persuasive. For the *major claim*, we found two attributes that differ in business model pitches: *specificity* and *evidence*. The *specificity* determines how detailed and specific the statement about the business model is, whereas the *evidence* ranks how well the major claim is backed up by supporting components. We found significant differences in both attributes throughout our corpus, which we aim to model with those scores. Tables 1 and 2 provide a more nuanced definition for the *specificity* and *evidence* in a 1-to-5 scale.

For *claims*, we defined *evidence* as a qualitative variable. Some claims seem to be strong in their statement. However, they do not contribute to the strength and persuasiveness of the overall business model. Thus, we specified *evidence* for a claim as the level of how well the claim supports the business model and/or the major claim. Most differences in the persuasiveness level in business model pitches can be found in the premises that back up the claims and thus the overall idea. We found premises to differ in two qualitative labels: *strength* and *relevance*. *Strength* is defined as how well a single premise contributes to the persuasiveness of the argument, and *relevance* determines how relevant a premise is for the overarching business idea. We believe that with these two scores we can model the most significant differences in the persuasiveness level of premises. Tables 3 and 4 provide an overview of the two scores.

Moreover, we found the business model pitches to also differ in their argumentative power on a discourse level. Sometimes a major claim is well formulated and supported by several claims and premises, but the business model is not really strong or novel in the overall picture because the argumentative discourse structure is weak. Therefore, we defined a document level score termed *"pitch strength"* to capture the persuasiveness of the argumentation discourse level of a business model. More information on pitch strength can be found in the Table 5.

All qualitative attributes are measured on a 1-to-5 scale following Carlile et al. (2018), with every level being precisely defined in our annotation guidelines. A summary of the variables is illustrated in Table 6.

3.3 Annotation Process

Two native German speakers annotated the business pitches independently from each other for the *major claim*, *claims*, and *premises* as well as their *argumentative relationships*. Moreover, they labeled the *pitch strength*, the specificity and the evidence of the major claim, the evidence of claims, and the strength and relevance of premises according to the annotation guidelines we specified. Inspired by Stab and Gurevych (2017a); Wambsganss et al. (2020c), our guideline consisted of 26 pages, including definitions and rules for what is an argument, which annotation scheme is to be used, and how argument components, argumentative relations, and the qualitative attributes are to be judged. After constructing the annotation guidelines, the results were discussed and validated by two independent senior researchers concerning the criteria of robustness, conciseness, extensibility, and com-

Score	Description
5	The major claim summarizes the argument well and has an addendum that indicates the extent to which the claim applies. Claims that summarize the argument must refer to most or all of the supporting components.
4	The major claim summarizes the argument very well by mentioning most or all of the supporting components. However, there is no addendum that states the conditions under which the claim is true. Alternatively, the claim moderately summarizes the argument by referring to a minority of the supporting components and includes an addendum.
3	The major claim contains a supporting component or addendum that indicates whether the claim is true. However, it does not adequately summarize the argument.
2	The major claim does not summarize the idea and does not contain an addendum that indicates whether the claim is true.
1	The major claim does not summarize the idea and is not explained by supporting components. It remains unclear to what the business model idea refers.

Table 1: Description of the *specificity* score for major claims.

Score	Description
5	A very strong, very convincing argument. There are many supporting components that have high relevance scores.
4	A strong, persuasive reasoning pattern. There are enough supporting components with respectable relevance scores.
3	The reasoning pattern is present. However, the supporting components do not have high relevance scores.
2	A poor, only possibly persuasive reasoning pattern. There are few supporting components. The relevance scores of the existing supporting components are low.
1	An unconvincing reasoning pattern. There are few or no supporting components. The relevance scores of the existing supporting components are low.

Table 2: Description of the *evidence* score for major claims and claims.

Score	Description
5	The relationship between the premise and the claim is very clear. It is very easy to see how the premise contributes to the clarity and persuasiveness of the claim.
4	The relationship between premise and claim is clear. At least one of the components is very specific and clear, while the other component might be not specific.
3	The Relationship between premise and claim is only clear with imagination. It takes some thought to imagine how the components are related. Both statements refer to the same topic but have no related ideas within the domain of the referred content.
2	The connection between premise and claim is not clearly evident. Some important assumptions are needed to relate the two components. A component may also receive this rating if both components have a low clarity.
1	The relationship between premise and claim is not apparent and is disjointed. Few people can see how the claim and premise are related.

Table 3: Description of the *relevance* score for premises.

Score	Description
5	A strong premise. By itself, it contributes very well to the persuasiveness of the argument.
4	A reasonable premise. It is a fairly strong point, but it could be improved to increase its persuasiveness.
3	An inadequate premise. It is not a strong premise and may persuade only a few readers.
2	A weak premise. It can only help persuade a small number of readers.
1	The premise does not contribute to persuasiveness at all.

Table 4: Description of the *strength* score for premises.

Score	Description
5	Little improvement or no improvement needed. The pitch describes, without any doubt, a very persuasive and strong business model.
4	The business idea is generally well understood but can be expanded.
3	Poorly understandable idea due to errors or ambiguity.
2	It is unclear what idea the author wants to support argumentatively (no relevant idea, idea is incomprehensible).
1	The pitch does not introduce an idea. It remains totally unclear what the business model is about.

Table 5: Description of the *strength* score for the pitch.

Score	Level	Description
pitch strength	pitch	How argumentative and persuasive is the overall business model pitch?
specificity	major claim	How detailed and specific is the statement about the business model?
evidence	major claim	How well is the major claim backed up by supporting components?
evidence	claim	How well does the claim support the business model / the major claim?
strength	premise	How well does a single premise contribute to the persuasiveness of the argument?
relevance	premise	How relevant is a premise for the overarching business idea?

Table 6: Description of the persuasiveness scores.

prehensibility. Several private training sessions and three team workshops were performed to resolve disagreements among the annotators and to reach a common understanding of the annotation guidelines. We used the *tagtog* annotation tool[3]. First, a text was classified into argumentative components (*major claim, claim, premise*) by the trained annotators. Second, the same annotators scored the argumentative relations and the qualitative attributes of the *major claim*, *premises*, and *claims* based on our annotation guideline on a 1-to-5 scale. After the first 50 pitches had been annotated by both annotators, we calculated the inter-annotator agreement (IAA) scores. As we obtained satisfying results, we proceeded with a single annotator who marked up the remaining 150 documents.

4 Corpus Analysis

4.1 Inter-Annotator Agreement

To evaluate the reliability of the argument component and argumentative relation annotations, we followed the approach of Stab and Gurevych (2014a).

Argument Components With regard to the argument components, two strategies were used. Since there were no predefined markables, the annotators not only had to identify the *type of argument component* but also its *boundaries*. In order to assess the latter, we use Krippendorff's α_U (Krippendorff, 2004), which allows for assessing 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 an argument component for a given sentence, we calculate percentage agreement and two chance-corrected measures, multi π (Fleiss, 1971) and Krippendorff's α (Krippendorf, 1980).

[3] https://tagtog.net/

	%	Multi-π	Krip. α	Krip. α_U
Major claim	0.9948	0.9673	0.9673	0.5186
Claim	0.8729	0.7087	0.7088	0.5002
Premise	0.8768	0.7454	0.7455	0.5356

Table 7: IAA of argument component annotations.

Table 7 displays the resulting IAA scores. We obtain an IAA of 87.3% for the claims and 87.7% for the premises. The corresponding multi-π scores are 0.71 and 0.75. Regarding Krippendorff's α, a score of 0.71 and 0.75 is obtained, indicating a substantial agreement for both categories. With a score of 0.50 and 0.54, the unitized α of both the claim and premise annotations is somewhat smaller compared to the sentence-level agreement. Thus, the boundaries of argument components are less precisely identified in comparison to the classification into argument types. Yet the scores still suggest that there is a moderate level of agreement between the annotators. Finally, with an IAA of 99.5% and a score of 0.97 for both multi-π and Krippendorff's α, we obtain an almost perfect agreement for the major claims. Hence, we conclude that the annotation of the argument components in student-written business model pitches is reliably possible.

Argumentative Relations To evaluate the reliability of the argumentative relations, we used the data set of all pairs of argument components that were possible during the annotation task according to our annotation scheme, i.e., all pairs of a major claim and a claim, a claim and a premise, and two premises. In total, the markables include 3,032 pairs of which 16.8% are annotated as support relations, while 83.2% of the possible pairs were left unidentified by an annotator. We obtained an IAA of 91.5% for the support relations. The corresponding multi-π and Krippendorff's α scores both amount to 0.61. Therefore, we conclude that argumentative relations can also be reliably annotated in business model pitches.

Persuasiveness Scores Finally, we determined the reliability of the qualitative argumentation labels based on Cohen's κ (Cohen, 1988). Considering the strength of the pitch, we obtained an almost perfect agreement between the two annotators (κ=0.88). With respect to the strength of the premise, we found moderate agreement (κ=0.47). The same applies to the specificity of the major

claim (κ=0.41), which allows the conclusion that the annotators' labels are reliable. Regarding the evidence for both the claim and the major claim, as well as the relevance of the premise, there is some room for improvement. However, with scores of κ=0.33, κ=0.30, and κ=0.28, the annotations still show a fair agreement between the labelers. Thus, qualitative argumentation labels can be reliably annotated in business model pitches, too.

4.2 Corpus Statistics

The final corpus consists of 200 student-written business pitches in German that are composed of 3,207 sentences with 61,964 tokens in total. Hence, on average, each document has 16 sentences and 305 tokens. A total of 262 major claims, 1,270 claims, and 1,481 premises were annotated. 1,069 textual spans were identified as not being an argument component ("None"). 2,018 support relationships were marked up by the annotators.[4]

5 Providing Students Adaptive Feedback

Modelling Argumentation Structures After constructing and analyzing our corpus, we leveraged the novel data to train a machine learning model. Our objective was to embed a classification algorithm in the back end of an argumentative writing support system to provide students with individual argumentation feedback in the writing process. The task is considered a sentence-based classification task, where each sentence can be either a *major claim*, a *claim*, a *premise*, or *non-argumentative*. Therefore, we trained and tuned a Long Short-Term Memory (LSTM) model (Hochreiter and Schmidhuber, 1997) to classify the argumentative components of a given text. We tokenized the texts and transformed them into word embeddings. The data set was split into training and test sets using an 80:20 split. For the component classification we received an accuracy of 54.12%, a precision of 55.90% and a recall of 54.12% on the test data. We benchmark our approach against a BERT model (Devlin et al., 2018). However, we received a rather unsatisfying accuracy of 47.50%, a precision of 46.66% and a recall of 47.50%. More information about the modeling can be found in Section B of the appendix.

Argumentation Writing Support System We designed and built an adaptive writing support system that provides students with individual feedback on their argumentation skill level based on our model. For the design of the tool, we followed the design principles of Wambsganss et al. (2020b) and self-regulated learning theory (Bandura, 1991; Zimmerman and Schunk, 2001). Our goal is to provide learners with adaptive self-evaluation opportunities based on logical argumentation errors irrespective of instructor, time, and location. Our system is illustrated in Figure 4.

Evaluation in a Writing Exercise We embedded the tool into a persuasive writing exercise where students were asked to write an argumentative pitch about a business idea. During this writing task, they received adaptive feedback on their argumentation level based on our model. The evaluation was conducted as a part of an exercise with students from a Western-European University, and thus designed and reviewed according to the ethical guidelines of the university. To keep data privacy standards, the students' data were additionally anonymized.

We conducted a field experiment to see if and how individual argumentation self-evaluation with adaptive feedback can assist students in writing more persuasive writings. We created a pedagogical scenario in which participants had to write a 300-word persuasive business pitch. The declared goal was to write a convincing pitch to persuade potential investors. Students were not required to participate in the assignment in order to pass the class; nevertheless, by successfully completing the assignment, they may increase their final mark by 2.2 percent. The persuasiveness of the business presentation had no influence on the assignment's grading and, as a result, no impact on final marks.

After the treatment, we measured the perceived ease-of-use according to Venkatesh and Bala (2008) by asking the following three items: "It would be easy for me to become adept at using the reasoning tool", or "I find the reasoning tool easy to interact with", and "Learning how to use the reasoning tool would be easy for me". Moreover, we measured the self-efficacy of students for the task of argumentation skill learning based on three items following Bandura (1991) to control for self-regulated learning. The items included, "In comparison to other users, I will write a good argumentative text", "I am sure that I could write a very good argumentative text", and "I think I now know quite a bit about argumentative writing." Both con-

[4]For detailed statistics see Section A of the appendix.

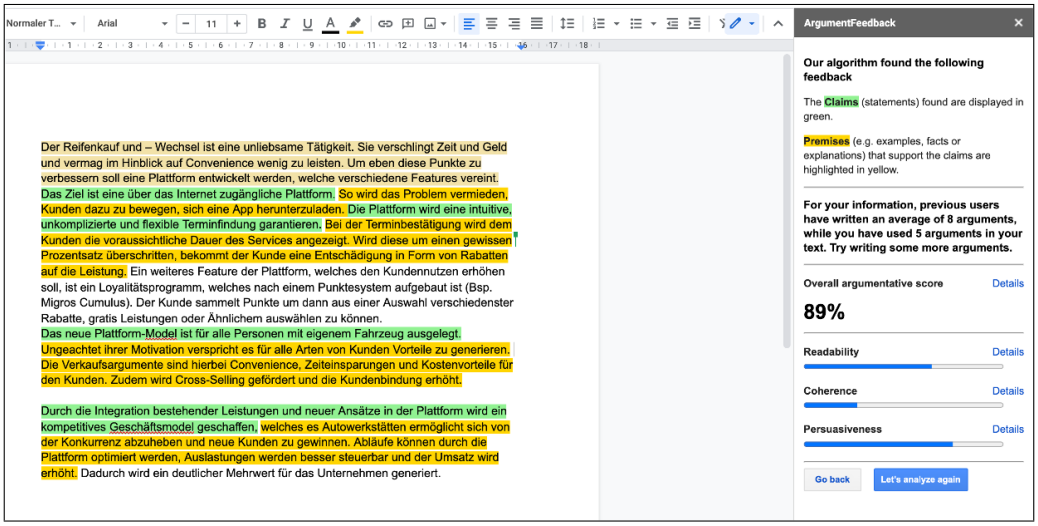

Figure 4: Screenshot of a trained model on our corpus as an adaptive writing support system.

structs were measured with a 1-to-7 point Likert scale (1: totally disagree to 7: totally agree, with 4 being a neutral statement). 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.

Results We received 25 valid results where participants successfully finished the writing exercises and the post-survey. Participants had an average age of 24.24 (SD= 3.83, 13 males, 12 females). The persuasive writing task took an average of 30 to 45 minutes. We calculated the mean for both constructs and compared them to the midpoints. All results were greater than the neutral value of 4, indicating a positive value for the design and the pedagogical scenario. A high perceived ease-of-use (mean= 4.94, SD= 0.98, normalized = 0.71) is especially important for learning tools to ensure students are experiencing the usage of the tool as a benefit and that they find it easy to interact with. This will foster the motivation, engagement, and adoption of the learning application. Moreover, positive effects for self-regulated learning can be also seen by comparing the means of the measured self-efficacy against the midpoints (Bandura, 1991). The average self-efficacy was 4.98 (SD= 0.98, normalized = 0.71) on a 1-7 Likert scale. Compared to the neutral value of 4, this is a positive indication that argumentation self-monitoring and self-evaluation help students learn in a self-regulated way.

Moreover, we evaluated the students' qualitative perception in order to check for the validity of our tool and model instantiation. The general attitude for our tool was positive. Participants positively mentioned the intelligent self-evaluation, the embedding in Google Docs, and the in-text highlighting several times. However, participants also asked for the tool to provide concrete argument suggestions on how to improve the argumentativeness.[5]

6 Conclusion

We propose an argumentation annotation scheme and introduce an annotated corpus of persuasive student-written business model pitches extracted from a pedagogical scenario. We offer a corpus of 200 student-written business model pitches with 3,207 sentences annotated for argument components, their relations, and six persuasiveness scores on different levels. By presenting an annotation study based on 50 persuasive pitches, we demonstrate that the annotation of student-written business model pitches is reliably possible. Finally, we embed and evaluated a trained model based on our corpus in an argumentation writing support tool for students. We thus aim to encourage fellow researchers to leverage our annotation scheme and corpus to design and develop argumentation support systems for students in large-scale scenarios.

Acknowledgments

We thank the Swiss National Science Foundation for supporting this research (grant 200207).

[5]More details on the application, and the evaluation can be found in Section B of the appendix.

References

Albert Bandura. 1991. Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, 50(2):248–287.

Elena Cabrio and Serena Villata. 2012. Natural language arguments: A combined approach. *Frontiers in Artificial Intelligence and Applications*, 242:205–210.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:621–631.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*.

Peter Daly and Dennis Davy. 2016. Structural, linguistic and rhetorical features of the entrepreneurial pitch: Lessons from Dragons' Den. *Journal of Management Development*, 35(1):120–132.

Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An Entity/Event-Level Sentiment Corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado. Association for Computational Linguistics.

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.

Ana Maria Ducasse. 2020. Evidence-based persuasion: A cross-cultural analysis of entrepreneurial pitch in English and Spanish. *Journal of International Entrepreneurship*, 18(4):492–510.

José Santiago Fernández-Vázquez and Roberto Carlos Álvarez-Delgado. 2019. The interaction between rational arguments and emotional appeals in the entrepreneurial pitch. *International Journal of Entrepreneurial Behaviour and Research*, 26(3):503–520.

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

James B Freeman. 2001. Argument structure and disciplinary perspective. *Argumentation*, 15(4):397–423.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse. Technical report.

John Hattie and Helen Timperley. 2007. The power of feedback.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to give feedback: Modeling attributes affecting Argument persuasiveness in student essays. *IJCAI International Joint Conference on Artificial Intelligence*, 2018-July:4130–4136.

Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. 2019. Give Me More Feedback II: Annotating Thesis Strength and Related Attributes in Student Essays. pages 3994–4004.

Klaus Krippendorf. 1980. Content analysis : an introduction to its methodology.

Klaus Krippendorff. 2004. Measuring the Reliability of Qualitative Text Analysis Data. *Departmental Papers (ASC)*, 38.

Deanna Kuhn. 1992. Thinking as Argument. *Harvard Educational Review*, 62(2):155–179.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Marco Lippi and Paolo Torroni. 2015. Argumentation Mining: State of the Art and Emerging Trends. *IJCAI International Joint Conference on Artificial Intelligence*, 2015-Janua(2):4207–4211.

Raquel Mochales and Marie Francine Moens. 2008. Study on the structure of argumentation in case law. *Frontiers in Artificial Intelligence and Applications*, 189(1):11–20.

Raquel Mochales Palau and Aagje Ieven. 2009. Creating an argumentation corpus: do theories apply to real arguments? {A} case study on the legal argumentation of the {ECHR}. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Law (ICAIL 2009), Twelfth international conference on artificial intelligence and law (ICAIL 2009)., Barcelona, Spain, 8-12 June 2009*, pages 21–30. ACM.

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

OECD. 2019. OECD learning compass 2030: In brief.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie Francine Moens. 2008. Language resources for studying argument. *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 2613–2618.

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.

Omar Sabaj, Paula Cabezas, Germán Varas, Carlos González-Vergara, and Álvaro Pina-Stranger. 2020. Empirical Literature on the Business Pitch: Classes, Critiques and Future Trends. *Journal of technology management & innovation*, 15(1):55–63.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument Extraction from News. *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66.

Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce M. McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, 5(1):43–102.

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

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78.

Christian Stab and Iryna Gurevych. 2014a. 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. 2014b. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)(Oct. 2014), Association for Computational Linguistics, p.(to appear)*, pages 46–56.

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

Christian Stab and Iryna Gurevych. 2017b. Recognizing Insufficiently Supported Arguments in Argumentative Essays. Technical report.

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

Stephen E. Toulmin. 2003. *The uses of argument: Updated edition*.

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

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A Review Corpus for Argumentation Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8404*, CICLing 2014, pages 115–127, New York, NY, USA. Springer-Verlag New York, Inc.

D N Walton. 1996. *Argumentation Schemes for Presumptive Reasoning*. Argumentation Schemes for Presumptive Reasoning. L. Erlbaum Associates.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, Cambridge.

Thiemo Wambsganss, Tobias Kung, Matthias Sollner, and Jan Marco Leimeister. 2021. Arguetutor: An adaptive dialog-based learning system for argumentation skills. In *Conference on Human Factors in Computing Systems - Proceedings*.

Thiemo Wambsganss, Nikolaos Molyndris, and Matthias Söllner. 2020a. Unlocking Transfer Learning in Argumentation Mining: A Domain-Independent Modelling Approach. In *15th International Conference on Wirtschaftsinformatik (WI 2020)*, pages 341–356, Potsdam, Germany.

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Jan Marco Leimeister, and Siegfried Handschuh. 2020b. 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. 2020c. A Corpus for Argumentative Writing Support in German. In *28th International Conference on Computational Linguistics (Coling)*.

Barry J Zimmerman and Dale H Schunk. 2001. *Self-regulated learning and academic achievement: Theoretical perspectives*. Routledge.

A Corpus Statistics

Tables 8, 9 and 10 present some statistics of the final corpus.

Figure 5 illustrates three basic examples of annotated relation in our corpus.

B Application

Benchmark of model We benchmark our LSTM approach against a BERT model (Devlin et al., 2018). However, we received a rather unsatisfying accuracy of 47.50%, a precision of 46.66%, and a recall of 47.50%. More details about the modelling results can be found in the confusion matrices for the argumentation compound classification task in Figure 6.

	total	mean	stdev.	min	max	median
Sentences	3,207	15.80	4.51	4	31	16
Tokens	61,964	305.24	67.12	104	473	317

Table 8: Distribution of *sentences* and *tokens* in the created corpus. Mean, standard deviation, minimum, maximum and median refer to the number of sentences and tokens, respectively, per document.

	total	mean	stdev.	min / max	median	%
Major cl.	262	1.29	0.70	0 / 4	1	6.42
Claim	1,270	6.26	2.05	1 / 13	6	31.11
Premise	1,481	7.30	2.46	2 / 14	7	36.28
None	1,069	5.27	3.24	0 / 17	5	26.19
All	4,082	17.87	5.21	3 / 34	18	100

Table 9: Types of *argument components*. Total describes the number of occurrences of the component in the document set. Mean refers to the average number of respective argument components per document. Standard deviation describes the corresponding amount of variation of the number of argumentative discourse units. Min denotes the minimum number of respective component found in a document, while max refers to the corresponding maximum number. Median signifies represents the value for which 50% of observations a lower and 50% are higher. Percent represents the percentage of the corresponding argumentative discourse units in the total set of documents.

	total	mean	stdev.	min / max	median	%
Support	2,018	5.47	1.76	1 / 11	5	12.75
None	13,814	72.51	41.06	2 / 242	72	87.25
All	15,832	77.99	42.66	3 / 253	78	100

Table 10: Types of *argumentative relations*.

	mean	stdev.	min	max	median
Strength of the pitch	3.56	0.68	2	5	4
Strength of the premise	3.25	0.45	2	5	3
Evidence of the major cl.	3.71	0.45	3	4	4
Evidence of the claim	3.53	0.51	2	5	4
Specificity of the major cl.	3.58	0.50	2	4	4
Relevance of the premise	3.49	0.51	2	4	3

Table 11: Statistics on the qualitative labels.

Hyperparameters of models The LSTM architecture consisted of eight layers and a dropout rate of 0. For the BERT model, we found a learning rate of 1e-5, a batch size of 16, and a training of the model over 25 epochs to provide the best results.

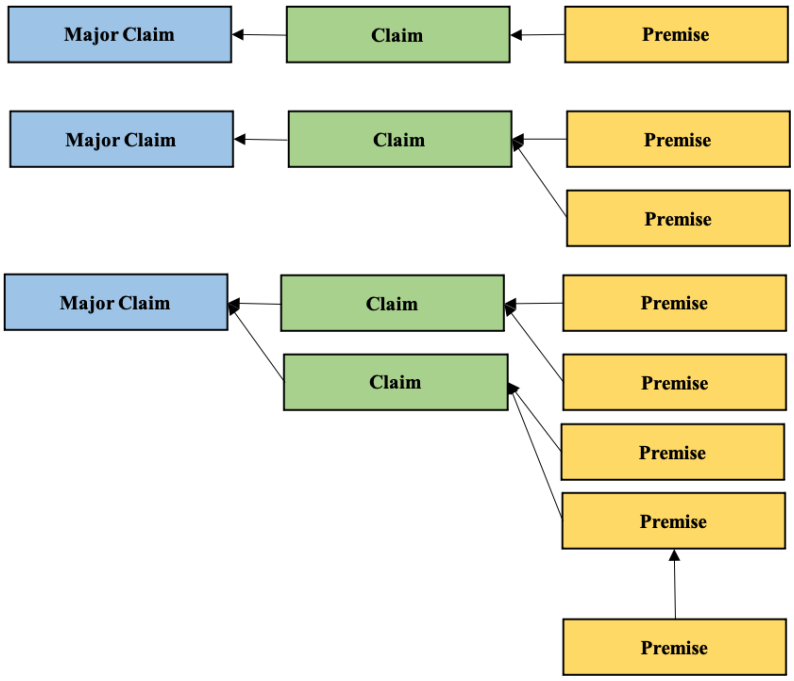

Figure 5: Examples of possible argumentation relations. The arrow signifies a support relation. The rectangle denotes an argument component in the form of a major claim, a claim or a premise.

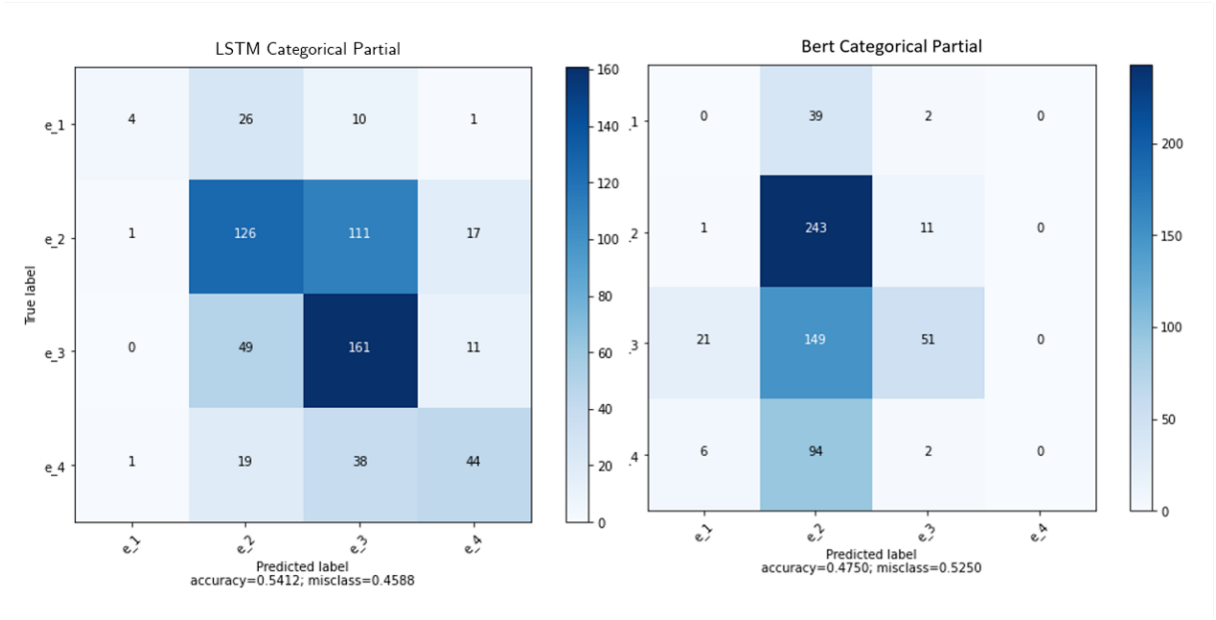

Figure 6: Confusion matrices for the argumentation compound classification task based on the LSTM and BERT models (e_1: major claim, e_2: claim, e_3: premise, e_4: non-argumentative).