A Corpus for Suggestion Mining of German Peer Feedback

Roman Rietsche, Eva Ritz, Julius Janda, Dominik Pfütze
Institute of Information Management, University of St.Gallen
Müller-Friedberg-Str. 8, 9000 St.Gallen, Switzerland
{roman.rietsche, eva.ritz, julius.janda,dominik.pfuetze}@unisg.ch

Abstract

Peer feedback in online education becomes increasingly important to meet the demand for feedback in large scale classes, such as e.g. Massive Open Online Courses (MOOCs). However, students are often not experts in how to write helpful feedback to their fellow students. In this paper, we introduce a corpus compiled from university students' peer feedback to be able to detect suggestions on how to improve the students' work and therefore being able to capture peer feedback helpfulness. To the best of our knowledge, this corpus is the first student peer feedback corpus in German which additionally was labelled with a new annotation scheme. The corpus consists of more than 600 written feedback (about 7,500 sentences). The utilisation of the corpus is broadly ranged from Dependency Parsing to Sentiment Analysis to Suggestion Mining, etc. We applied the latter to empirically validate the utility of the new corpus. Suggestion Mining is the extraction of sentences that contain suggestions from unstructured text. In this paper, we present a new annotation scheme to label sentences for Suggestion Mining. Two independent annotators labelled the corpus and achieved an inter-annotator agreement of 0.71. With the help of an expert arbitrator a gold standard was created. An automatic classification using BERT achieved an accuracy of 75.3%.

Keywords: Peer Feedback, Suggestion Mining, Corpus, Annotation

1. Introduction

In the educational domain, large scale lectures often lack the possibility of providing personalised feedback, which is caused by scarce resources (Rietsche & Söllner, 2019). Further, the recent rise of technology-mediated learning environments, i.e. MOOCs impedes the student-per-educator ratio and educators struggle to give individual feedback to their students. Recently, one phenomenon used to overcome this issue are peer-review processes between students (Pastor & Baruffaldi, 2020). However, students are usually non experts in how to structure feedback, that it becomes most helpful for the feedback receivers to improve their work. Therefore, we introduced a feedback process for students within a master's university course to develop our corpus to detect suggestions on how to improve ones work. In a peer feedback setting, students give each other feedback on lecture-based tasks in text form. By using the corpus every semester, we can continuously expand the source texts and feedback. More detailed, we aim to extract information from the feedback and apply suggestion mining based on the corpus.

In general, the relevance of suggestion mining from unstructured text increased due to the large availability of reviews and feedback in different forms online (Viswanathan et al., 2011). Recent work has been conducted on mining texts to extract user sentiments or opinions and analyse sentiments to see positive or negative emotions in feedback. However, suggestion mining goes one step further by not only extracting sentiments but also detecting suggestions for improvements from the feedback (i.e. new ideas or solutions based on the task) (Verma & Ramamurthy, 2016). This is done by extracting sentences from unstructured feedback data.

Previously developed corpora often only incorporated a binary classification task, classifying each sentence in either a suggestion or a non-suggestion. We try to fill this gap with our research in which we created an annotation scheme including ten categories and in particular capture implicit and explicit suggestions. Two annotators labelled overall 7,488 sentences.

Our corpora can be used by researchers and practitioners as training data e.g. to develop a machine learning model for capturing suggestions. The model could than be used in downstream applications to predict peer feedback helpfulness and provide advice on which kind of suggestions need to be added in order to make the feedback more helpful.

In the following, we will provide more detailed information about the corpus within the following structure: The second section presents relevant work and previous corpora on suggestion mining form feedback. Thus, the following section presents our corpus: Section 3 describes the metadata used, section 4 specifies secondary data, section 5 provides detailed information on the corpus profile, section 6 the empirical evaluation is conducted and section 7 shows the results of machine learning experiments conducted on the corpus. After, section 8 provides discussion and section 9 a conclusion.

2. Related Work

Suggestion Mining is the extraction of sentences that contain suggestions from unstructured text (Negi et al., 2018). The Oxford dictionary defines a suggestion as *an idea or a plan that you mention for somebody else to think about*. Based on Negi (2019) two different types

	Unigrams in German	English translation	Bigrams in German	English translation	Trigrams in German	English translation
EXP-1	einfügen, ergänzen, hinzufügen, weglassen, löschen, noch, ergänzt	insert, supplement, add, omit, delete, still, supplements	ergänzt werden, würde ich, hinzugefügt werden, noch einen, noch hinzufügen, noch ergänzen, eine ebene	be supplemented, I would, be added, another, still add, still supplement, a level	würde ich noch, ich würde dies, noch erwähnt werden, ich noch die, bei der abgrenzung, könntest du noch, würde ich eine	would i still, i would this, still be mentioned, i still the, at the demarcation, could you still, i would a
EXP-2	genauer, würde, noch, detaillierter, beschreiben, ausführen, etwas	more precise, would, still, detailed, describe, elaborate, something	noch etwas, noch genauer, würde ich, etwas detaillierter, etwas genauer, mehr auf, ich würde	a little more, more detailed, I would, a little more detailed, a little more precise, more on, I would	noch etwas detaillierter, noch mehr auf, könntest du noch, mehr auf die, genauer beschrieben werden, vielleicht könntest du, würde ich empfehlen	more detailed, more on the, could you still, more on the, more detailed description, maybe you could, I would recommend
EXP-3	überlegen, dir, überdenken, überlege, nochmals, schaue, überprüfe	consider, you, reconsider, think, again, look, review	überlegen ob, überlege dir, du dir, dir überlegen, ob du, noch einmal, gedanken machen	consider whether, you consider, you consider, whether you, once again, give thought to	würde ich mir, du dir überlegen, dir überlegen ob, könntest du dir, ich würde mir, kannst du dir, könnte man sich	I would consider, you consider, you consider whether, could you consider, I would myself, could you yourself, could one oneself
IMP-1	fehlt, fehlen, vergessen, zeilen, max, besucher, interface	misses, missing, forgotten, lines, max, visitor, interface	es fehlt, fehlt die, fehlt noch, fehlen die, fehlt mir, es fehlen, fehlt der	it is missing, misses the, still missing, missing the, missing me, missing, missing the	bei der qualitativen, beim service blueprint, pfeil von der, im e3 value, business modell ist, der qualitativen beschreibung, in der Datenbank	in the qualitative, in the service blueprint, arrow of the, in the e3 value, business model is, the qualitative description, in the database
IMP-2	ganz, verstehe, unklar, wer, nicht, warum, was	quite, understand, unclear, who, not, why, what	nicht ganz, ganz klar, was ist, ist mir, nicht klar, meinst du, verstehe nicht	not quite, quite clear, what is, is to me, not clear, do you mean, do not understand	nicht ganz klar, ist mir nicht, mir nicht ganz, ist nicht ganz, ich verstehe nicht, verstehe ich nicht, ganz klar was	not quite clear, is not to me, to me not quite, is not quite, I don't understand, don't I understand, quite clear what
IMP-3	falsch, ob, sehe, kritisch, herausforderung, wäre, schwäche	wrong, whether, see, critical, challenge, would, weakness	sehe ich, sicher ob, nicht sicher, ganz korrekt, bin mir, bin ich, stellt sich	I see, sure if, not sure, quite correct, I am, I am, turns out to be	nicht sicher ob, mir nicht sicher, nicht ganz korrekt, bin ich mir, schwäche der lösung, ich mir nicht, sicher ob die	not sure if, not sure to me, not quite correct, i am myaself, weakness of the solution, i am not, sure if the
ILLU	beispiel, beispielsweise, zum, beispielsweisew, etc, vs, rabatt	example, exemplarily, for, exampleew, etc, vs, discount	zum beispiel, beispiel der, ein beispiel, als beispiel, eine lösung, kooperationen mit, dies kann	for example, example of, an example, as an example, a solution, cooperations with, this may be	zum beispiel der, sich der kunde, die plattform die, auf der plattform, mit dem kunden, in deinem fall, zum beispiel bei	for example the, the customer itself, the platform the, on the platform, with the customer, in your case, for example with
JUST	denn, ja, dadurch, so, dies, somit, heute	because, yet, thereby, so, this, thus, nowadays	denn wenn, ja nicht, denn ich, so kann, dies würde, ist ja, ja auch	because if, yet not, because I, so can, this would, is, yes also	du in der, sich der kunde, der fall ist, der kunde nicht, auf dem markt, weiss ich nicht, in diesem fall	you in the, the customer itself, the case is, the customer not, on the market, i don't know, in this case
SUMM	oben, genannten, erwähnten, änderungen, anpassungen, umsetzen, überarbeitung	above, mentioned, changes, adaptations, implement, revision	die oben, bpmn noch, oben genannten, der überarbeitung, nur noch, unter punkt, wie im	the above, bpmn still, above mentioned, of the revision, only, under the point, as in the	stärker auf die, in der überarbeitung, überarbeitung der lösung, der überarbeitung sollte, auf die value, bei der überarbeitung, alles in allem	more on the, in the revision, revision of the solution, the revision should, on the value, in the revision, all in all
APPR	gut, sehr, finde, verständlich, idee, gute, ausführlich	good, very, find, understandable, idea, good, detailed	sehr gut, finde ich, ist sehr, ich finde, und gut, gute arbeit, die idee	very good, think i, is very, i think, and good, good work, the idea	finde ich sehr, ist sehr gut, ich finde die, ich finde deine, ich sehr gut, gut dass du, sehr ausführlich und	I find very, is very good, I find the, I find your, I very good, good that you, very detailed and

Table 1: Most frequent unigrams, bigrams and trigrams per label in the gold standard

of suggestions were identified. On the one hand, *explicit* suggestions, in which the reader receives a direct action from the reviewer. On the other hand, *implicit* suggestions, in which the reader has to imply the action suggested by the reviewer. We oriented on this suggestion scheme as well.

Several corpora have been developed to automatically mine suggestions from reviews or feedbacks. In the last years, the mining of suggestions from unstructured data became a popular research interest of the computer linguistics community (Nicol & Macfarlane-Dick, 2006). However, previously developed corpora often only incorporated a binary classification task, classifying each sentence in either a suggestion or a non-suggestion. For

	German Example	English translation
EXP-1	Auch die Beschriftungen von Events würde ich ein bisschen verschieben, damit diese nicht über den Pfeilen liegen.	I would also move the event labels a bit so that they are not on top of the arrows.
EXP-2	Als Alternative versuchen die Personas neu zu definieren.	Alternatively, try to redefine the personas.
EXP-3	Was du dir überlgen kannst, den Service neben einer App auch als webbasierte Lösung anzubieten.	What you can think about is to offer the service as a web-based solution in addition to an app.
IMP-1	In dem Kundenpool fehlt das Endereignis.	The final event is missing in the customer pool.
IMP-2	(Auch ist mir unklar ob das die Tür aufschliesst oder öffnet im Sinn einer automatischen Schiebetür).	(It is also not clear to me whether this unlocks or opens the door in the sense of an automatic sliding door).
IMP-3	Wären die Anbieter/Künstler oder die Verkäufer von Konzertkarten nicht auch Akteure innerhalb eines derartigen Plattformgeschäfts?	Wouldn't the providers/artists or the sellers of concert tickets also be actors within such a platform business?
ILLU	Unmittelbares Melden von Beschädigung der Gegenstände etc.	Immediate reporting of damage to objects etc.
JUST	Module müssten folglich im Unternehmenspool, bzw. über dessen Lanes gebildet werden.	Modules would therefore have to be created in the company pool or via its lanes.
SUMM	Fokussiere dich am besten auf die Schwächen aus meinem Punkt 2.	It is best to focus on the weaknesses from my point 2.
APPR	Die Alternativen Geschäftsmodelle passen sehr gut zum Hauptangebot.	The alternative business models fit very well with the main offer.

Table 2: Sample sentences per label extracted from the gold standard

instance, Brun & Hagege (2013) annotated and created a detection system to mine suggestions from product reviews in English language, performed an in-domain evaluation of the classifier models and reached a 77% precision. Moreover, Wambsganss et al. (2020) developed a corpus for argumentative writing support based on student-written feedback on business models in the German language, but also included only two labels, leading to a 70% accuracy. Negi (2016) advanced these mining corpora by performing a suggestion classification task in a domain-independent setting. Based on Negi's previous work, we conducted the annotation. To the best of our knowledge, our corpus is the first one to automatically mine suggestions from unstructured feedback data with ten labels in German language.

3. Corpus Construction

3.1. Data Source

The corpus introduced here was compiled from university students' peer feedback the students gave to each other via an online learning platform which contained the peer feedback process. The corpus covers a time period from 2017 to 2018, in which the students attended a master's course of business information management. During the course, each student has to provide two separate *writing assignments* which belong to two tasks. The first task is to write a new business model or expand an existing business model of choice. Later in the course, the second task is to design a business process model and notation (BPMN) and to explain it. After the writing assignment each student becomes a reviewer and has to provide another assignment in the form of giving feedback to their peers. Three guiding questions were provided which the reviewer could use when writing the *feedback assignment*. First, which are the strengths of the solution of your peer? Second, what are the weaknesses of the solution and how can they be addressed? Third, what should be emphasised during the revision of the solution? The feedback is given back to the writer of the source text in order to improve the initial work. As a final third assignment, the student rewrites the initial solution with the feedback of her fellow student and submits the final solution as a *revised assignment*. Students participating in the peer feedback process received points for their final grade. The corpus presented here is built upon the feedback assignment of the students.

We introduce our corpus with the following metadata available:

- `year`: the year in which the student attended the course (here: 2017 or 2018)
- `task`: the task on which the source text was based on (here: "Write a new business model or expand an existing business model of your choice." and "Design a BPMN and explain it."
- `pos`: sentence segmented part-of-speech tags generated with the Python library spaCy using the German language model *de_core_news_sm* (Honnibal et al., 2020)
- `reader`: anonymised reader id
- `reviewer`: anonymised reviewer id
- `sents`: number of sentences
- `words`: number of words
- `characters`: number of characters including spaces

4. Annotation Scheme

For identifying suggestions we follow the work by (Negi et al., 2018). In which suggestions are categorised into the two main categories explicit and implicit suggestions. Figure 1 shows an overview of our annotation scheme and the process of labeling sentences into the categories. If the feedback reader should add, remove, or move something in her work, the suggestion is of type EXP-1. If she should clarify, restructure, or redefine something, the suggestion is of type EXP-2. And if the reader should consider a new

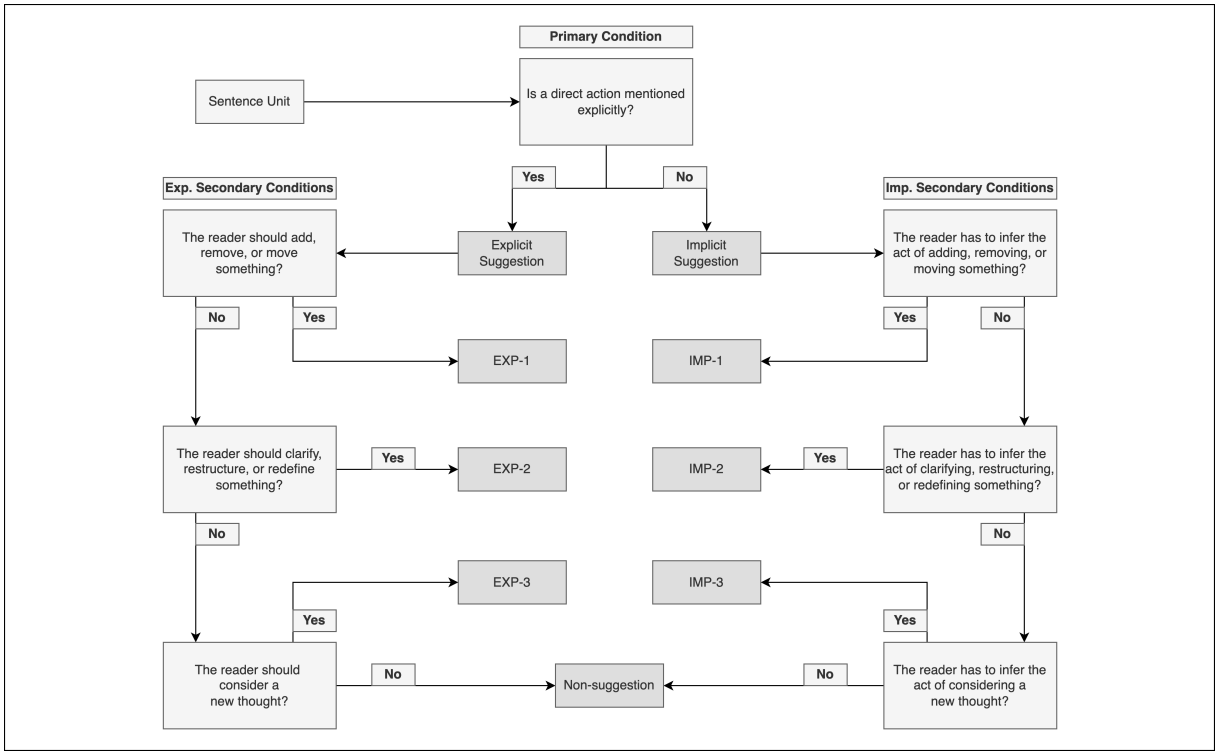

Figure 1: Annotation process of implicit vs. explicit suggestions

thought given by the reviewer, the suggestion is of type EXP-3. If the reader has to infer those actions, the suggestion type is IMP-1, IMP-2, and IMP-3, respectively. Is such an action mentioned neither explicitly nor implicitly, the sentence constitutes a non-suggestion. Non-suggestions may include additional information expanding on a suggestion. These categories are discussed subsequently under the term enrichment (ILLU, JUST, SUMM). Additionally, non-suggestions may constitute statements that simply approve of the work done by the recipient, subsequently referred to as approval (APPR). The breath of these categories enabled assigning every sentence in a given document one of the ten labels. In ambiguous cases, i.e. when definitions of two or more categories overlapped for a single sentence, the most salient label was chosen by the annotators and, given disagreements, resolved by an arbitrator. For a better understanding of which characteristics a sentence must satisfy in order to be annotated with a specific label, we analyse the gold standard data in more detail. Therefore, we determined the most frequently occurring n-grams of the respective annotated sentences. Table 1 shows the seven most frequent unigrams, bigrams and trigrams per label in its original German form and in a translated English version in a descending order of frequency. For the sake of clarity, the first column of Table 1 is further explained and Table 2 depicts a sample sentence per label in an original and translated version.

4.0.1. Explicit Suggestions

The label EXP-1 is given when the reader is prompted to add, remove, or move something in her work. As can be seen in the first field of Table 1 the most frequently used words here are *insert, supplement, add, omit, delete, still, and supplements*. All words are directly addressed verbs to the reader as a request to add or remove something and thus this label is well reflected in the data. However, to "move" is not represented in the unigrams, since in our dataset a reviewer wants something added and removed more frequently than moved. If we look at the bigrams, though, one might suspect that the last entry indicates a "move", since *a level* is mentioned as in "I would move this paragraph down a level". The adverb *still* also has no direct relation to type EXP-1, yet, it can be put in front of any aforementioned verb in order to accentuate it. Suggestion type EXP-2 shows *more precise, would, still, detailed, describe, elaborate, something* as most frequently used words. The annotator gives label EXP-2 if the reader is asked to clarify, restructure, or redefine something. *More precise, detailed, describe, elaborate* are all words in which the reviewer requests for a clarification. A specific word connected to "restructure" or "redefine" cannot be found in these most common n-grams, since those sub-labels are underrepresented in our dataset. The underrepresentation means that the reviewers' focus in their feedback texts was more on "clarifications" than on "restructures" and "redefinitions". The words *would, still, something* are again filler words which can be put in front or after the

aforementioned words. EXP-3 is annotated if the reviewer requests the reader to consider a new thought. In Table 1 *consider, you, reconsider, think, again, look, review* are the words connected to this label. All verbs give a good reflection of suggestion type EXP-3. However, *you* and *again* are more reasonable when considering the next column of bigrams in which both words appear once more.

4.0.2. Implicit Suggestions

The first implicit label is IMP-1, which is given when the reader has to infer an act of adding, removing, or moving some part in her text. As can be seen in the table, the words *misses, missing, forgotten* are indirectly addressed verbs to the reader to infer the act of adding more content. The high frequency of the word *lines* can be explained by a threshold given in the initial task of the writing process as in "The task did not allow to write more than 10 lines", whereby the reader has to infer the act of removing content. Words *max, visitor, interface* have no connection to suggestion type IMP-1, however seem to be highly represented in our dataset. The second implicit label is IMP-2, which is annotated if the reader has to infer the act of clarifying, restructuring, or redefining. The most frequent word connected to the label is *quite* which is more reasonable in the context of IMP-2 when considering the trigrams column. *Understand, unclear, who, not, why, what* are all expressions of a need for clarification on the reviewer side and thus give a good reflection of the label inside the data. The third implicit label, IMP-3, is given by the annotator when the reviewer asks the reader to infer the act of considering a new thought. IMP-3 is represented by the most frequent words of *wrong, whether, see, critical, challenge, would, weakness*. The annotators agreed to give this label also in case there were differences to common knowledge. The reviewer has therefore noticed that something has been understood wrong and asks the feedback reader to reconsider. Thus, the most frequent words in this category are more reasonable considering their general direction of meaning pointing all to a critical reconsideration of the reader's ideas.

4.0.3. Enrichments

Since a reviewer can enrich his core suggestion with as much information as pleased, we have added an additional category to the existing implicit and explicit category, which we named *enrichments*. We separated this new category into the three following labels: Illustration (ILLU), Justification (JUST), and Summarisation (SUMM). The label ILLU is given by the annotator if the reviewer enriched the core suggestion by examples. Table 1 shows the most frequent words of the ILLU label including "for example" and "etc", which we expected . The second enrichment, JUST, is annotated when the reviewer enriched the core suggestion by a justification. Words like *because, yet, thereby, so, this, thus* are all conjunctions introducing a justification, however the high frequency of *nowadays* might be explained by a justification made to accentuate the difference between the past and today as in "Nowadays, latest technology is used which outperforms your idea". The SUMM label is used by the annotator if the reviewer summarises the suggestions that she already posed. SUMM is reflected very well by all most frequent words (*above, mentioned, changes, adaptations, implement, revision*) connected to the label in the data.

Since the reviewer not only gives suggestions, but also approves ideas of the source text, we added another label category we called Approval (APPR), which is of special importance when using our dataset for the task of Sentiment Analysis. The annotator gives the label APPR if the reviewer approves the ideas of the feedback reader. As can be seen in the last field of Table 1, the most frequently used words are *good, very, find, understandable, idea, good, detailed*. Most of these words are positive adjectives and thus this label is well reflected in the gold standard data.

4.1. Annotation Process

Two native German speakers annotated the students' peer feedback independently from each other according to the annotation scheme specified in section 4. A team workshop and several private training sessions were performed to reach a common understanding of the annotation scheme. We used the TagTog annotation tool to label the texts since it provides a graphical interface and coloured text mark ups in order to make a clear demarcation of each of the ten labels (Campos Prieto & Cejuela, 2021). More than 600 feedback texts were annotated by two annotators, however, in case of disagreement an expert arbitrator was consulted in order to discuss the specific cases in detail and to reach an agreement between the two. The objective of these *consultations* were, on the one hand, that the annotators address questions and ambiguities to the arbitrator and, on the other hand, to jointly agree on a reasonable and proper annotation. Those consultations were held after 10, 20, 50, 100, and 200 labelled feedback texts. The remaining texts were annotated without consultation, resulting in an inter-annotator agreement (IAA) of 0.71. To create a single version of the gold standard, the arbitrator took the final decision in cases where the two annotators still disagreed. The specific links in-between labels, e.g. especially between suggestions and enrichments, have not been annotated in this study, however, this could be a useful addition in future research.

5. Structure of the Final Corpus

The resulting corpus profile can be seen in Table 3. The corpus consists of 617 documents, which corresponds to 7,488 sentences and 747,680 characters. The minimum number of sentences, words and characters are

1, 27 and 187, respectively. The maximum number of sentences, words and characters are 102, 1,542 and 10,545, respectively. The average number of sentences per document, words per document, words per sentence and characters per word are 12.14, 175.84, 14.49 and 6.89, respectively.

Number of documents	617
Number of sentences	7,488
Number of words	108,496
Number of characters	747,680
Avg. no. of sentences per document	12.14
Avg. no. of words per document	175.84
Avg. no. of words per sentence	14.49
Avg. no. of characters per word	6.89

Table 3: Corpus profile

To provide a better overview of the textual data, the part-of-speech tags (POS) were generated using the spaCy language model *de_core_news_sm* and their relative frequencies were calculated and depicted in Figure 2. Punctuation, numerals, particles, determiners, and interjections are neglected in the pie chart since their frequency in the corpus is not representative and the informational relevance is low. Proper nouns and nouns are consolidated, as well as verbs and auxiliary verbs and conjunctions and subordinating conjunctions. As can be seen in Figure 2 the two detached frequencies of nouns (NOUN) and verbs (VERB) show the highest values of all seven POS analysed, followed by adjectives (ADJ), adverbs (ADV), adpositions (ADP), pronouns (PRON), and conjunctions (CONJ) in descending order.

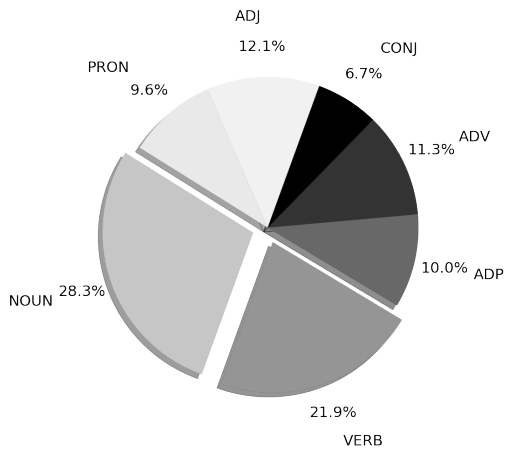

Figure 2: Distribution of part-of-speech tags in the gold standard corpus

6. Empirical Experiments

The annotated data provides two primary benefits for suggestion mining of peer feedback. First, previous suggestion mining datasets have been primarily confined to customer reviews and product feedback. This corpus provides labelled data specific to student peer feedback and thus enables new analyses and insights in the domain of peer assessment research. Second, suggestions have been conventionally analyzed through binary classification (Dong et al., 2013; Jia et al., 2021; Li, 2019; Negi et al., 2019; Ramanand et al., 2010; Zingle et al., 2019). The comprehensive annotation scheme used for this corpus allows framing suggestion mining as a multi-class classification problem, facilitating more detailed analysis of suggestions in peer feedback. To empirically validate the utility of the corpus in providing these benefits to suggestion mining, we have applied a deep-learning-based text classification model. This initial classification may be used to extend and improve the labelled data further by including the students giving and receiving the feedback in the annotation process, such as to flag and improve incorrect predictions provided by a continuously trained machine learning model. Expanding the corpus as part of the university's course repeating by semester offers a unique opportunity to incorporate such a continuous improvement of predictions into the learning management system originating the data.

6.1. Data

The training data for the model was directly derived from the corpus presented previously. Using all ten labelled class types, the following experiment validates the text classification accuracy using the annotated data, presupposing a comparatively fine-grained classification task. However, the annotation scheme also allows analyzing suggestions in more condensed setups. For example, depending on the use case, classes may be merged by implicit and explicit suggestions or by suggestions to add/remove/move, clarify/restructure/redefine, and consider/act. Thus, the following metrics provide a baseline assuming the most detailed analysis.

6.2. Methods

Neural networks have generally surpassed traditional machine learning methods for text classification tasks using architectures such as recurrent neural networks and Transformers (Minaee et al., 2021). The Transformer architecture solely relies on self-attention mechanisms and thus allows for parallel computation and more efficient training (Vaswani et al., 2017). This results in a wide variety of available pretrained language models (PLMs) that can be used for transfer learning. Transfer learning allows reusing the knowledge acquired by the model of a particular task, such as language modeling, on a different task, such as text classification (Ruder et al., 2017). In utilizing transfer learning, we have compared the performance of multiple transformer-based pretrained language models for German text data fine-tuned to our classification task, including BERT, DistilBERT, and XLM. BERT, which

PLM	Architecture	Details	Corpora	Accuracy	Weighted F1-Score
German BERT (bert-base-german-cased)	BERT	12-layer, 768-hidden states, 12-attention heads, 110M parameters	German Wikipedia, OpenLegalData, News Articles	66.36%	65.88%
German BERT large (gbert-large)	BERT	24-layer, 1024 hidden states, 16 attention heads, 335M parameters	German OSCAR corpus, OPUS project, German Wikipedia, OpenLegal-Data	**73.56**%	**73.07**%
distilbert-base-german-cased	DistilBERT	6-layer, 768-hidden, 12-heads, 66M parameters	Wikipedia, EU Bookshop corpus, Open Subtitles, CommonCrawl, ParaCrawl, News Crawl	62.08%	60.13%
xlm-mlm-ende-1024	XLM	6-layer, 1024-hidden, 8-heads	English and German Wikipedia	50.60%	46.88%

Table 4: Comparison of different transformer-based pretrained language models for German and their accuracy on test data

stands for Bidirectional Encoder Representations from Transformers, is a transformer model developed at Google (Devlin et al., 2018). DistilBERT is a reduced version of BERT, retaining 97% of its language understanding capacity but decreasing its size by 40% (Sanh et al., 2019). Finally, XLM is a cross-lingual language model developed for pretraining in different languages (Lample & Conneau, 2019). We found the BERT architecture in connection with the "German BERT large" pretraining initialized through the HuggingFace library to provide the best results (Chan et al., 2020; Devlin et al., 2018; Wolf et al., 2020). Thus, this setup, consisting of 24 layers and appended with additional linear layers to enable fine-tuning for sequence classification, was chosen as the architecture for the subsequent experiment. The model was further customized to allow for hyperparameter optimization of the number of hidden layers, layer size, dropout rate, and sequence length. Additionally, the batch size, number of epochs, and learning rate were optimized.

7. Results

Table 4 compares the performance of the different PLMs for German language data mentioned previously. Based on these results, the BERT architecture using German BERT large was selected and further optimized using automatic hyperparameter optimization, with the following configuration receiving the best results:

- Linear Layers: 1
- Learning Rate: 0.00002
- Sequence Length: 50
- Epochs: 12
- Batch Size: 28
- Optimizer: AdamW

	Accuracy	Weighted F1-Score
SVM	56.61%	55.86%
BERT Baseline	73.56%	73.07%
BERT Optimized	75.30%	75.33%

Table 5: Comparison of multi-class classification setups on test data

As the configuration using no additional hidden linear layers performed best, the dropout rate and layer size could be neglected for the final model. Table 5 compares the initial BERT configuration, the optimized BERT configuration, and a baseline support vector machine (SVM) setup. The features used by the SVM are extracted by counting unigrams and bigrams and weighing the terms using Term Frequency-Inverse Document Frequency. N-grams appearing in over 70% of texts are disregarded. Thus, the SVM results provide a baseline for classifying the corpus sentences using only key unigrams and bigrams weighted by importance. Table 6 evaluates the precision and recall of all single classes, giving insight into how the proposed feedback classes differ in the model's capability to classify them.

This experiment provided baseline metrics for multi-class text classification using the presented corpus and feedback classes. To achieve accuracies comparable to state-of-the-art approaches, the experiment focused on neural network and transformer-based classifiers with pretrained language models for German text, substantially outperforming an approach based on SVMs. Fur-

	Precision	Recall	F1
EXP-1	75%	78%	77%
EXP-2	64%	77%	70%
EXP-3	69%	67%	68%
IMP-1	52%	46%	49%
IMP-2	78%	74%	76%
IMP-3	59%	58%	58%
ILLU	74%	68%	71%
JUST	60%	67%	64%
SUMM	56%	45%	50%
APPR	96%	94%	95%

Table 6: Precision, recall, and F1-Score per label

thermore, the experiment compared different PLMs that can be used with German text, of which BERT-based PLMs performed best. The final optimized BERT architecture achieved a 75.30% classification accuracy, validating the annotated data as per the ten proposed feedback classes. Future experiments may test neural network architectures beyond transformer-based approaches. Additionally, more PLMs may be assessed. In improving the transfer learning approach, unlabelled feedback data may be used to continue the pretraining of a chosen PLM, increasing the language understanding for the current domain and, thus, improving downstream classification performance. Finally, the comparison of precision and recall metrics per class may be used to inform future annotation initiatives of feedback classes.

8. Conclusion

In this paper we introduced a corpus for Suggestion Mining of more than 600 German peer feedback texts (about 7,500 sentences) to be able to detect suggestions on how to improve the students' work and to be able to capture peer feedback helpfulness. To the best of our knowledge, this corpus is the first student peer feedback corpus in German which additionally was labelled with a new annotation scheme. The annotation scheme involved ten annotation labels which are separated into explicit and implicit suggestions, enrichments and approvals. Two native German speakers annotated the corpus in consultation with an expert arbitrator and achieved an IAA of 0.71. To create a single version of the gold standard, the arbitrator took the final decision in cases where the two annotators disagreed. The labelled corpus was empirically validated by automatic multi-class classification using BERT. The best model of our automatic classification approach yielded an accuracy of 75.3%. The neural approach outperformed the feature-engineered baseline model, which was trained on a SVM.

Future work may address both elaborated and distilled annotation schemes. Specifically, sentences that fall into the implicit suggestions and enrichment categories may be subject to change, motivated by this paper's investigation of their class precision and recall scores. The resulting training sets may be used in conducting further suggestion mining experiments.

As the corpus encompasses more than 10,000 feedback texts in total, it provides a high amount of unlabelled data. This unlabelled data may be used to improve future models through continuing the pre-training of the language model on tasks such as masked language modeling and next sentence prediction as used by BERT and thus improve down-stream classification tasks.

For researchers interested in the data, we will provide the annotated corpus and Python scripts for the purpose of implementation and progress on further experiments.

9. Bibliographical References

Brun, C., & Hagege, C. (2013). Suggestion mining: Detecting suggestions for improvement in users' comments. *Res. Comput. Sci.*, *70*(79), 171–181.

Campos Prieto, J., & Cejuela, J. M. (2021). *TagTog: The Text Annotation Tool to Train AI [Software].* Available from: https://tagtog.net/.

Chan, B., Schweter, S., & Möller, T. (2020). German's next language model. *arXiv preprint arXiv:2010.10906*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, L., Wei, F., Duan, Y., Liu, X., Zhou, M., & Xu, K. (2013). The automated acquisition of suggestions from tweets. In *Twenty-seventh aaai conference on artificial intelligence.*

Honnibal, M., Montani, I., Landeghem, S. V., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python.* Zenodo. doi: 10.5281/zenodo.1212303

Jia, Q., Cui, J., Xiao, Y., Liu, C., Rashid, P., & Gehringer, E. F. (2021). All-in-one: Multi-task learning bert models for evaluating peer assessments. *arXiv preprint arXiv:2110.03895*.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Li, J. (2019). Lijunyi at semeval-2019 task 9: An attention-based lstm and ensemble of different models for suggestion mining from online reviews and forums. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 1208–1212).

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, *54*(3), 1–40.

Negi, S. (2016). Suggestion mining from opinionated text. In *Proceedings of the acl 2016 student research workshop* (pp. 119–125).

Negi, S. (2019). *Suggestion mining from text* (Unpublished doctoral dissertation). National University of Ireland–Galway.

Negi, S., Daudert, T., & Buitelaar, P. (2019). Semeval-2019 task 9: Suggestion mining from online reviews and forums. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 877–887).

Negi, S., De Rijke, M., & Buitelaar, P. (2018). Open domain suggestion mining: Problem definition and datasets. *arXiv preprint arXiv:1806.02179*.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, *31*(2), 199–218.

Pastor, J.-C., & Baruffaldi, L. (2020). The role of regulatory focus on a peer-feedback process: A longitudinal study with mba students. *Academy of Management Learning & Education*(ja).

Ramanand, J., Bhavsar, K., & Pedanekar, N. (2010). Wishful thinking-finding suggestions and'buy'wishes from product reviews. In *Proceedings of the naacl hlt 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 54–61).

Rietsche, R., & Söllner, M. (2019). Insights into using it-based peer feedback to practice the students providing feedback skill. In *Proceedings of the 52nd hawaii international conference on system sciences* (pp. 63–72).

Ruder, S., et al. (2017). Transfer learning-machine learning's next frontier. *Accessed: April*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Verma, S., & Ramamurthy, A. (2016). Analysis of users' comments on political portal for extraction of suggestions and opinion mining. In *Proceedings of the international conference on advances in information communication technology & computing* (pp. 1–4).

Viswanathan, A., Venkatesh, P., Vasudevan, B., Balakrishnan, R., & Shastri, L. (2011). Suggestion mining from customer reviews. In *Proceedings of the seventeenth americas conference on information system.*

Wambsganss, T., Niklaus, C., Söllner, M., Handschuh, S., & Leimeister, J. M. (2020). A corpus for argumentative writing support in german. *arXiv preprint arXiv:2010.13674*.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., ... others (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45).

Zingle, G., Radhakrishnan, B., Xiao, Y., Gehringer, E., Xiao, Z., Pramudianto, F., ... Arnav, A. (2019). Detecting suggestions in peer assessments. *International Educational Data Mining Society*.