The Specificity and Helpfulness of Peer-to-Peer Feedback in Higher Education

Roman Rietsche[1], Andrew Caines[2], Cornelius Schramm[1], Dominik Pfütze[1], Paula Buttery[2]
[1] Institute of Information Management, University of St Gallen, Switzerland
[2] ALTA Institute & Computer Laboratory, University of Cambridge, United Kingdom
roman.rietsche@unisg.ch, cornelius.l.schramm@gmail.com
{andrew.caines,paula.buttery}@cl.cam.ac.uk

Abstract

With the growth of online learning through MOOCs and other educational applications, it has become increasingly difficult for course providers to offer personalized feedback to students. Therefore asking students to provide feedback to each other has become one way to support learning. This peer-to-peer feedback has become increasingly important whether in MOOCs to provide feedback to thousands of students or in large-scale classes at universities. One of the challenges when allowing peer-to-peer feedback is that the feedback should be perceived as helpful, and an import factor determining helpfulness is how specific the feedback is. However, in classes including thousands of students, instructors do not have the resources to check the specificity of every piece of feedback between students. Therefore, we present an automatic classification model to measure sentence specificity in written feedback. The model was trained and tested on student feedback texts written in German where sentences have been labelled as general or specific. We find that we can automatically classify the sentences with an accuracy of 76.7% using a conventional feature-based approach, whereas transfer learning with BERT for German gives a classification accuracy of 81.1%. However, the feature-based approach comes with lower computational costs and preserves human interpretability of the coefficients. In addition we show that specificity of sentences in feedback texts has a weak positive correlation with perceptions of helpfulness. This indicates that specificity is one of the ingredients of good feedback, and invites further investigation.

1 Introduction

With thousands of students in MOOCs and hundreds of students in university classes, instructors increasingly apply the approach of peer-to-peer feedback (or, 'peer feedback'), where students provide feedback to their peers (van Popta et al., 2017; Lipnevich and Smith, 2018). Peer-feedback enables instructors to provide individual feedback on every piece of coursework by leveraging the potential of students to provide feedback to each other (Piech et al., 2013). Nevertheless, since students are often not experts in providing feedback, the instructors need to ensure that the feedback is helpful (Strijbos et al., 2010). Research has shown that one factor determining feedback helpfulness is whether the feedback points are generic or specific (Lipnevich and Smith, 2018; Shute, 2008; Hattie and Timperley, 2007). Generic feedback such as "improve your submission" are less helpful than detailed, targeted advice such as "add a timeline" or "change the caption in Figure 1".

However, the challenge is that instructors who do not have time for providing feedback themselves also do not have time for checking the specificity level of peer feedback (Mulryan-Kyne, 2010). One approach is to develop a model which automatically analyses feedback specificity using natural language processing. Recent work has been carried out into automatic classification of sentence specificity in newspaper articles; for instance by Li and Nenkova (2015), Louis and Nenkova (2011) and Ko et al. (2019).

Our work builds on this previous research and at the same time provides distinct contributions. Firstly, we apply the approach in the novel domain of education and peer-feedback, which is inherently different in its purpose and nature compared to the news domain which features in previous work. News articles are written for a general audience with the purpose to inform, whereas peer-feedback texts are written to reveal the strengths and weaknesses of written work and provide suggested improvements. Furthermore, in the peer-feedback scenario, each student has put effort into their assignments: thus they have a certain expectation as to the quality of feedback they ought to receive.

Secondly, we have developed a unique dataset

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 107 - 117

of peer-feedback containing more than 1000 sentences labelled for specificity. Thirdly, the data we work with are in the German language: to the best of our knowledge, all previous related work has been on English. Fourthly, we find that there is a correlation, albeit weak, between sentence specificity and the perceived helpfulness of peer-feedback.

We train and evaluate four classifiers based on a feature set which is determined by methods described in previous work and our own observations of specificity in peer-feedback texts. We also explore the relationship between sentence specificity and perceived helpfulness of peer-feedback, finding a weak positive correlation, which suggests that specific sentences are helpful but also that further work is needed to uncover the other ingredients of good feedback.

We contribute our collected corpus of sentences from peer-feedback texts in German for further analysis and hope to provide researchers and practitioners with a detailed analysis and discussion of sentence specificity. The code and annotated corpus can be accessed via github[1].

2 Theoretical Background

2.1 Characteristics of Sentence Specificity

In general, definitions of sentence specificity are often related to the "quality of belonging or relating uniquely to a particular subject" (Lugini and Litman, 2017) as well as the amount of detail contained within a sentence. The example sentences (s) from newspapers and product reviews below include *S1* and *S2*, which are more specific than *S3* and *S4*.

S1 "90% of women wear Mascara making it the most commonly worn cosmetic, and women will spend an average of $4,000 on it in their lifetimes" (Ko et al., 2019, p. 1).

S2 "While American PC sales have averaged roughly 25% annual growth since 1984 and West European sales a whopping 40%, Japanese sales were flat for most of that time" (Louis and Nenkova, 2011, p. 1818).

S3 "This cosmetic is very popular and many people use it regularly" (Ko et al., 2019, p. 1).

S4 "Now, the personal-computer revolution is finally reaching Japan" (Louis and Nenkova, 2011, p. 1818).

General sentences are broad statements about a topic, while specific sentences contain details and can be used to support or explain the general sentences further (Louis and Nenkova, 2012). General sentences create expectations in the reader's mind of further evidence or examples from the author. Specific sentences can stand by themselves, since they provide detailed information (Li and Nenkova, 2015). This difference in the level of detail contained in general and specific sentences is often a matter of degree, rather than an entirely straightforward distinction. Therefore the linguistic realisation of sentence specificity and its automatic detection is a rather complex matter.

In the domain of online education platforms featuring peer-feedback systems, sentence specificity refers to the level of detail in the feedback text (Shute, 2008). The analysis of online forum dialogues has shown that argument quality is highly correlated with specificity of claims in the context of argument mining (Swanson et al., 2015). Specific feedback guides students directly to changes in their assignment by helping them to identify those parts of the text that the reviewer considers more or less conducive to successful performance (Goodman and Wood, 2004). A large body of evidence suggests that increasing the specificity of feedback has a positive relationship with immediate or short-term performance (Kluger and DeNisi, 1996; Ilgen et al., 1979).

2.2 Related Work on Sentence Specificity

Previous work on sentence level specificity prediction has mostly been focused on English texts and on domains starkly different from academic feedback texts such as news articles (Louis and Nenkova, 2011) or tweets (Ko et al., 2019). Sentence specificity prediction as a task is proposed by Louis and Nenkova (2011), who re-purposed discourse relation annotations from *Wall Street Journal* articles (Prasad et al., 2008) for sentence specificity training. Li and Nenkova (2015) incorporated more news sentences as unlabeled data and developed Speciteller, a tool for predicting the specificity score of sentences. They improved classification accuracy by using a semi-supervised co-training method on over 30K sentences from the Associated Press, *The New York Times*, and the *Wall Street*

[1] https://github.com/RomanRietsche/feedbackspecificity

	German (Original)	English
S1	Auf Seite 4 beim Modul 2 solltest du besser 'würde' statt 'könnte' geschrieben.	On page 4 in module 2, you should write 'would' instead of 'could'.
S2	Den ersten Schritt des Service Blueprints würde ich "Registrierung auf der Hotel Match Plattform" nennen → klar machen, dass es sich um eine Website/ ein online tool handelt.	I would call the first step of the service blueprint "Registration on the Hotel Match platform" → make it clear that this is a website/online tool.
G1	Deine Lösung gefällt mir insgesamt sehr gut.	Overall, I like your solution a lot.
G2	Der Service Blueprint ist extrem gut gemacht und strukturiert dargestellt.	The visualization of the service blueprint is extremely good and structured.

Table 1: Examples of specific (S) and general (G) feedback sentences from our dataset, originally in German with English translation.

Journal.

Li et al. (2016) developed the annotation scheme used in Louis and Nenkova (2011) and Li and Nenkova (2015) by considering contextual information, and by using a scale from 0 to 6 rather than binary specificity annotations. Lugini and Litman (2017) produced a system to predict sentence specificity for classroom discussions, though the dataset they use is not publicly available. All the above systems are classifiers trained with categorical data (2 or 3 classes). Ko et al. (2019) presented an unsupervised domain adaptation system for sentence specificity prediction, designed to output real-valued estimates from binary training labels to generalize predictions to domains where no labeled data are available.

3 Data

Our dataset consists of peer-feedback texts written by students on a Masters Course on Business Innovation at a German-speaking University, collected over the past five years. Students followed a peer-feedback process which is similar to the scientific paper review process in academia (Ziman, 1974). Students submitted their assignment to a learning management system. Each assignment was afterwards anonymously distributed to three reviewers who each wrote their feedback before then being sent back to the assignment author. There were no rules on how to write the feedback, students only received three guiding questions: *what was good, what was not so good and what possible improvements could be made?* Each feedback text is on average 250 words long.

Table 1 provides examples from our dataset taken from both ends of the specificity spectrum. Specific feedback gives the recipient a more direct indication of strengths, weaknesses, and suggested changes (e.g. [S1] and [S2]). General sentences such as [G1] and [G2], on the other hand, often refer to entire sections or the whole work and require further clarification-questions or interpretation by the feedback recipient. Note that peer-feedback has unique characteristics which differ from other domains. It is possible for sentences to contain generalized statements which would normally be classified as such, yet in the context of peer review feedback they are in fact specific suggestions. For instance the sentence, "young people are much less obsessed with their car's internal specs than older people", contains a rather generalized statement. Yet in the context of a reviewer critiquing the reviewee's business personas, it may appear to be more specific: "I do not think the persona of Anna would be interested in your service, because young people are much less obsessed with their car's internal specs than older people.". This more complex sentence becomes a more specific criticism than simply stating, "I don't think that the persona of Anna is realistic".

For the annotation process we randomly sampled 1000 feedback texts from our corpus and adopted two strategies for annotation. First, relying on many annotators who rated only a limited amount of sentences, whereby each sentence is annotated by 5 annotators and second, relying on two students who in several workshops receive training on how to annotate specificity and an expert in NLP as arbitrator for the two annotations. In both strategies the annotators rated the specificity on a scale of 1 (very general) to 5 (very specific) developed by Li et al. (2016) and Ko et al. (2019).

We chose the two strategies because, both have their advantages and disadvantages. For example, the first approach reduces systemic bias of one individual annotator on the whole dataset, since annotators only labelled a limited number of sentences. A downside of this strategy is that, there is no opportunity for annotators to learn over time and therefore reaching agreement on the level of specificity for one sentence is more difficult. The second strategy has the benefit of learning effects but the possible downside of systemic biases by two annotators labelling many sentences.

For the first strategy, we used Survey Circle[2]. The dataset was formed from a random sample of 1000 sentences from the 1000 feedback texts. We made the annotation job available to Survey Circle users based in Germany, Austria, Switzerland, specifying that they should be German speakers. The users on Survey Circle are typically students from a variety of disciplines. Overall, 1000 sentences where annotated by 200 users who each annotated 25 sentences. Each sentence was reviewed five times by five different annotators. For quality control, we removed ratings by users who chose the same label for every one of their sentences, and who did not complete at least 15 annotations. Since our focus was on high quality data we only chose sentences with an inter-annotator agreement (IAA) higher than 60% to further proceed with in our classification algorithm, leaving us with 331 sentences with an average IAA of 0.804. To create a final dataset, we took the mean of 5 annotations, which resulted in the final specificity score. The fact that we had to filter out so many sentences at this stage, due to low IAA, prompted us to try a different approach to annotation.

In the second strategy, we randomly selected 75 of the 1000 feedback texts and removed all sentences having a character length lower than 40 (since usually those sentences solely included bullet points, enumerations, or wrong sentence segmentations). This pre-processing resulted in a final dataset of 800 sentences. Two native German speakers annotated the sentences independently from each other in the same manner as done previously on Survey Circle, but this time using the decision tree shown in Figure 1. A team workshop and several calibration training sessions were performed to reach a common understanding of the annotation. 800 feedback texts were annotated by

[2] https://www.surveycircle.com

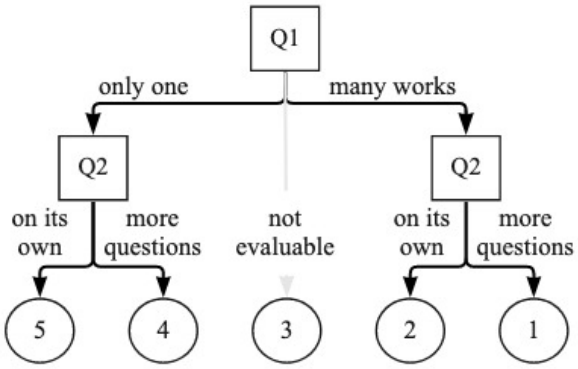

Figure 1: Specificity Annotator Prompt. Q1: "Is this feedback sentence only applicable to this individual work (eg. it references specific paragraphs, objects or people from the source text) or could it be written generically about many different works?" Q2: "Can this feedback sentence stand on its own, or does it require concretising questions or interpretation by the feedback recipient, in order to be implemented or understood?"

the two annotators – 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 annotations resulted in an IAA of 0.746. To create a single version of the gold standard, the arbitrator took the final decision in cases where the two annotators still disagreed. Finally, we merged both datasets to give a total of 1131 sentences.

For our machine learning experiments we additionally obtained a binary label for each sentence, therefore we aggregated the ratings 4 and 5 to be specific (meaning a label of 1) whereas ratings 1 and 2 were deemed to be general (i.e. a label of 0). We chose to remove from the dataset the 170 sentences that received a rating of 3 "unrateable". This gave us an almost balanced dataset containing 48% general sentences and 52% specific sentences.

4 Classification Experiments

Using the data described in the previous section we undertake experiments to automatically classify sentences as specific or general. Being able to do so accurately will allow us to identify when a reviewer's text contains no specific feedback, and potentially encourage them to be more specific in downstream applications. We compare 'classic' feature-based classifiers with a BERT-based model (Devlin et al., 2019) fine-tuned on our dataset.

4.1 Feature-based classification

We use the features described below for binary classification of sentence specificity based on those used in previous work and based on an intuition of what it means for a sentence to be specific or general in the context of peer feedback. We sample from the set of commonly used classifiers and train support vector machine (RBF kernel), logistic regression, and random forest models using the following features.

Sentence length: General sentences are expected to be shorter than specific ones (Louis and Nenkova, 2011). There are three features to capture this observation: the number of words in the sentence, the number of nouns, and the number of noun chunks as identified by spaCy[3]. Noun chunks are 'base noun phrases' – phrases with a noun as their head.

Word length: We compute the average length of words in each sentence, in characters, expecting long words to be indicative of more complex vocabulary and therefore more specific feedback (Ko et al., 2019).

Qualitative words: General sentences feature the frequent usage of qualitative words such as adjectives and adverbs (Louis and Nenkova, 2011). To capture this word-class based information we take counts of adjectives and adverbs in the texts.

Word specificity: We use three sets of features to capture specificity of words in the sentence. The first of these is based on GermaNet (Henrich and Hinrichs, 2010; Hamp and Feldweg, 1997), the German language adaptation of WordNet (Miller, 1995). We compute a specificity measure using the hypernym relations in GermaNet. For each noun and verb in our example sentences, we record the length of the path from the word to the root of the GermaNet hierarchy through the hypernym relations (Louis and Nenkova, 2011). The longer this path, the more specific we expect the word to be. The average, minimum and maximum values of these distances are taken for nouns and verbs found in GermaNet.

IDF: Another set of features is based on the inverse document frequency (IDF) of a word w (Sparck Jones, 1972), defined as $log\frac{N}{n}$, where N is the number of documents in a corpus, and n is the number of documents that contain the word w. We used 3 million German sentences taken from newspaper texts in 2015[4] from the Leipzig Corpus Collection (Goldhahn et al., 2012) to compute the idf (excluding punctuation and stop words). The features for a sentence are the average, minimum and maximum IDF scores for words in the sentence (Louis and Nenkova, 2011) – the intuition being that words in general sentences are more common whereas specific sentences contain words seen less often.

Sentiment: We noticed that general sentences were regularly found in positive feedback – often praising a section or even the entire work (recall examples [G1] and [G2] in Table 1). Therefore, we record the number of positive, negative, neutral and polar (not neutral) words per sentence using two lexicons – SentiWS, a publicly available German-language resource for sentiment analysis (Remus et al., 2010) and TextblobDE[5]. We add another set of features where each of these counts is normalized by the sentence length (Louis and Nenkova, 2011). In addition we obtain a count of **polar words** (non-neutral words) and a normalized **sentiment score** per sentence.

Discourse connectives: A count of the most common discourse connectives – "because", "furthermore", "either or", "on the other hand", *etc* – as these were often indicative of a point argued in greater detail which usually entailed a more specific sentence. Furthermore, we noticed that certain phrases were characteristic of general sentences ("in general", "overall", "all in all", *etc*) and count the occurrence of such words and phrases.

Non-alphanumeric characters: Another feature is the normalized count of non-alphanumeric or special characters (such as }%"§-'→) (Li and Nenkova, 2015). Due to the digital and conversational nature of the peer feedback we collected, symbols such as → were frequently used as substitutes for discourse connectives. Quotation marks, percentage or section signs were also often indicative of references to specific sections of the business plan.

URLs: Specific suggestions were sometimes accompanied with reference material in the form of internet links which is why we also count the number of URLs per sentence.

Named entities: These are generally regarded to be suggestive of specific sentences (Louis and Nenkova, 2011). In addition to counting all named entities using spaCy, we additionally count all mentions of *personas*, as they often appeared in contexts of the reviewer critiquing the recipient's pro-

[3] https://spacy.io

[4] https://www.kaggle.com/rtatman/3-million-german-sentences

[5] https://textblob-de.readthedocs.io

Model	Accuracy	Precision	Recall	F-measure
support vector	75.0	76.1	75.0	75.1
random forests	76.7	76.8	76.7	76.8
logistic regression	74.7	75.6	74.7	74.8
BERT_BASE cased	**81.1**	**81.5**	**81.1**	**81.0**

Table 2: Performance of sentence specificity classifiers on German sentences – accuracy, precision, recall, F-measure; mean of 10-fold cross-validation.

posed business personas.

Numbers: This is the count of numeric tokens or number words, since they are often associated with references to specific pages or other specifics of the student assignments.

Currency: In the context of business plans, currencies and currency symbols were often found in sentences criticising specific monetization or revenue schemes and therefore we count their occurrence in each text.

Morpho-syntactic labels: We use the spaCy dependency parser for German to extract a number of morpho-syntactic features from each sentence. We obtain counts of dependency relations, part-of-speech tags, and a concatenation of these for each token in a sentence. For instance, the sentence *Ich mag deine Arbeit* ('I like your work') would produce the following concatenated labels combining part-of-speech tags and dependency relations: PRON_sb, VERB_ROOT, DET_nk, NOUN_oa (subject, root, noun kernel element, accusative object in the TIGER treebank scheme (Rehbein and van Genabith, 2007)).

Word counts: We count the frequency of all non stop-words, as well as the sum of stop words both raw and normalized by sentence length.

Word vectors: We compute the average of the word vectors obtained from spaCy's `de_core_news_lg` model for German for each sentence, with L2 normalisation (Horn and Johnson, 2013). We also compute the vector average without the vectors of stop words.

4.2 BERT-based classification

It has become a common and successful practice in empirical NLP work in recent years to make use of large transformer language models for text classification in *transfer learning* scenarios (Rogers et al., 2020). Accordingly, we use the Hugging Face Transformers library to fine-tune the BERT_BASE cased model for German which was pre-trained and made available by deepset[6] (Wolf et al., 2020). We fine-tune to the training set in each of ten folds in our dataset in a cross-validation set-up.

4.3 Evaluation

Following Li and Nenkova (2015) we report four performance metrics for our experiments, where the *specific* label is viewed as the 'positive' one: *accuracy*, the proportion of correctly predicted sentence specificity labels; *precision*, the proportion of positive predictions which are correct; *recall*, the proportion of positive labels in the test set which are correctly identified; and the *F-measure*, the harmonic mean of precision and recall.

5 Results

In Table 2 we show performance metrics for the classification of sentence specificity in our German peer-feedback dataset. We report mean scores from ten-fold cross-validation, and we compare three feature-based classifiers with a fine-tuned BERT-based model.

To summarise, we find that the BERT-based fine-tuned classifier performs best. Not unexpectedly, the superior performance of BERT comes at a computational cost, as the fine-tuning of the transformer takes significantly longer than fitting the other models (>5mins as opposed to a few seconds), and requires GPU. Furthermore, BERT offers little in the way of interpretability. In this regard, algorithms such as logistic regression and random forests are advantageous due to their human understandable coefficients. We would therefore opt for a feature-based classifier if putting a sentence specificity detection system into production: the efficiency gains and advantage with respect to explainability in our view outweigh the performance boost provided by a BERT-based model.

We analysed which of our features were the best predictors of sentence specificity. To that end we

[6] https://deepset.ai

Feature	Ratio
numbers	1.98
noun chunks	1.59
non-alphanumeric characters	1.53
SentiWS negative words	1.42
named entities	1.42
discourse connectives	1.08
adjectives	1.05
discourse chunks	1.02
currency	1.00
SentiWS positive words	0.99
adverbs	0.91
TextblobDE negative words	0.90
minimum GermaNet hypernym path	0.86
TextblobDE sentiment score	0.84
TextblobDE polar words	0.80

Table 3: Top 15 features from the logistic regression model ranked by coefficient representing odds ratios.

rank the features from the logistic regression classifier by coefficient. The coefficients represent log odds that an observation is in the target class ('specific'), and thus we take the exponent of the coefficients to obtain odds ratios. Table 3 shows the top 15 features ranked by coefficient, where the latter indicate that for every one unit increase in the value of the feature the odds that the sentence is specific are n times greater than the odds that the sentence is not specific, with all other features held constant.

We find that features relating to *numbers and currency*, *non-alphanumeric characters*, and *named entities* are the most likely to occur in specific feedback. This reflects the fact that the subject domain is business but also that such features are associated with specific references to locations in the text, and the *non-alphanumeric characters* featuring in specific feedback formatting such as bullet points, section markers and parentheses, or punctuation used as connectives (e.g. right arrows and dashes). We find that other highly weighted features are representative of specific feedback texts in general, such as a high number of *noun chunks*, *named entities*, *words with clear polarity*, *adjectives* and *discourse connectives*. Finally, we note that a longer minimum *hypernym path* in GermaNet for words in a sentence is associated with more specific feedback, as we hypothesised (section 4.1).

6 Feedback Specificity and Helpfulness

We examined the interplay between feedback specificity and helpfulness to evaluate the hypothesis that more specific feedback is more helpful (Strijbos et al., 2010). We sampled 500 feedback texts from the business masters course previously referred to, presented them to Survey Circle annotators (students and PhDs), and asked them to score the strength of their agreement with the following four statements on a scale of 1 to 10 for each text: *"The feedback from the reviewer was helpful"*,*"The reviewer was able to provide constructive suggestions on their stated critical aspects"*, *"The reviewer was able to identify critical aspects in the assignment"*, or *"The feedback from the reviewer was of high helpfulness"*. The mean of these Likert scores was taken from 5 annotators per text and across all 4 statements to give an overall feedback helpfulness score for each text between 1 and 10.

To derive a specificity score for a feedback text, we made per-sentence specificity predictions using the BERT-based model trained on the annotated peer-feedback set of 1000 sentences described above. The score per text was then the average sentence specificity prediction, a value between 0 and 1. The correlation between text specificity scores and helpfulness ratings showed a correlation of 0.21 with a statistically significant p-value <.001. This finding helps to corroborate the hypothesized relationship between specificity and feedback helpfulness, while reminding us that the relationship is not straightforwardly linear. A strongly helpful feedback text should not contain entirely general sentences or entirely specific ones, but some combination of the two. In Figure 2 we show a scatter plot of the feedback specificity per text (per cent of sentences in a text classified as specific by the model) against the feedback helpfulness score per text calculated in the way described above, and both the weak correlation and variation in the relationship are apparent.

7 Discussion

We show that sentence specificity can be classified successfully in German peer-feedback texts. This can be a useful first step for various education technology applications. For instance we can provide students with automated advice on how to improve their written peer-feedback. It can potentially help with feedback to students on their written assignments as well, in cases where students have not

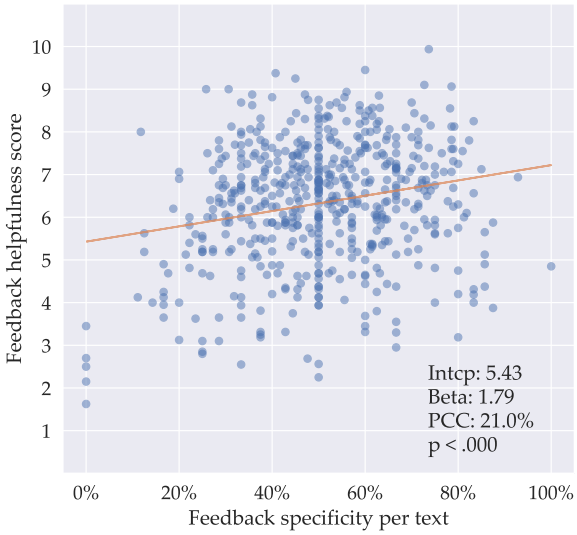

Figure 2: Correlation between text specificity (% specific sentences per text) and feedback helpfulness score (average human ratings of 4 criteria) for 500 peer feedback texts.

made sufficiently specific statements. For this reason, explainability and low computational cost are important factors in weighing up the performance of our feature-based and BERT-based specificity classification models.

One limitation of the current sentence level approach is that it fails to deal with dependent sentences where a feedback point is argued for over multiple sentences. To accurately rate the specificity in such cases, it can be crucial to take into account the context in which a sentence appears. Consider, for instance, the following example:

> **[1]** *Regarding your business processes on page 10 - does it really need a chatbot that asks for targets here?* **[2]** *One input line would be enough for that.* **[3]** *Chatbots only make sense when customers actually interact with them.*

Sentence [**3**], taken on its own, contains a rather general statement and the logistic regression model assigns a probability of around 0.06 that it is specific (less than 0.5, thus 'general'). When taking its context into account, it becomes clear that the entire section is referencing a specific element of the business plan and calling into question a specific piece of the business process with a concrete argument. Consider a reformulation of the three previous sentences like so:

> **[4]** *I consider the chatbot that asks for destinations (page 10) to be superfluous, as chatbots only make sense when customers interact with them — one input line would be enough for that.*

Now the model assigns a probability of around 0.73 that the sentence is specific, thereby classifying it as 'specific'. Naturally, sentence **[4]** is more likely to have features associated with specificity since it is longer than sentence **[3]**, but the change in regression scores does illustrate how specificity of feedback can develop in context. Since we model specificity only at the sentence level in this work, the application of our model to feedback texts is determined by the author's punctuation choices and the sentence tokenization that results.

To address this issue in future work, we can attempt to segment texts into 'argumentation chunks' rather than sentences. Such an approach requires a combination of information density extraction, argumentation mining and specificity prediction. This observation is congruent with previous work which concluded that context information should be considered in the annotation procedure to mitigate the effect of anaphoric and topical references that may otherwise be inadequately dealt with (Louis and Nenkova, 2012; Li et al., 2016). In addition, it is apparent that any downstream application should be tuned so that recommendations on feedback specificity at a per-sentence level take the whole text into account, so that the student is encouraged to write a well structured mix of general and specific feedback.

Finally, we note that specificity could be just one of multiple components that determine the helpfulness of feedback. In truth, feedback helpfulness is difficult to measure objectively since in large part it is driven by how helpful a student *perceives* it to be. O'Donovan et al. (2019) state that, "what a student considers good assessment and feedback is shaped by the assumptions they hold as to the nature and certainty of knowledge (Baxter Magolda, 1992), their prior learning experiences (O'Donovan, 2017) as well as the timing of their consideration (Carless and Boud, 2018)". Just getting technical factors right will not ensure student satisfaction with feedback (p. 8)". In the long run, the sole focus on the feedback itself and its language is too narrow as it is only part of the complexity of providing good feedback (Evans, 2013). To make a holistic improvement to feedback procedures at large as well as enhance student engagement and satisfaction, peer assessment process design, pre-feedback con-

ditions, and predictability need to be considered as well (O'Donovan et al., 2019). It is likely that perceptions of feedback helpfulness are influenced by a number of contributing factors, some of which are in the text – e.g. lexical content, pragmatic implication and argumentation – while others are external and concern the wider educational context of the assignment. For instance, the feedback should be relevant to the task, on topic, and consistent with the curriculum. We expect that specific sentences should also be used with more generic 'big picture' and bridging sentences, and that feedback providers could be prompted to provide a mixture of both. There is also the pedagogical question of timing: when more specific feedback is beneficial for the student and when it is not. These issues represent opportunities for future investigation.

8 Conclusion

We have presented experiments in automatic classification of the specificity of German sentences in peer feedback written by students in an online assignment reviewing system. We derived features based on previous work and the qualitative analysis of our dataset, and performed multiple experiments using machine learning models compared to a transfer learning approach with BERT (Devlin et al., 2019). We found that our classifiers were able to successfully predict sentence specificity with an accuracy of at least 70% for all models. The BERT model mostly outperforms the feature-based classifiers, but it has the highest computational cost and does not have human interpretable coefficients. SVM performs best on the peer-feedback texts for feature-based models, is computationally more efficient and provides per-feature coefficients which enable downstream explainability for any user-facing system.

In addition, in the analysis of our logistic regression model we report which features are most likely to indicate feedback specificity, and find that *numbers*, *noun chunks* and *non-alphanumeric characters* are at the top of the list. We found a weak correlation between crowdsourced assessments of feedback helpfulness and feedback specificity, underlining that texts containing relatively high proportions of specific sentences are more likely to represent good quality feedback.

Acknowledgements

The second and fifth authors are supported by Cambridge University Press & Assessment, University of Cambridge.

References

Marcia B. Baxter Magolda. 1992. *Knowing and reasoning in college: Gender-related patterns in students' intellectual development*, 1st ed. edition. The Jossey-Bass social and behavioral science series. Jossey-Bass, San Francisco.

David Carless and David Boud. 2018. The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8):1315–1325.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Stroudsburg, PA, USA. Association for Computational Linguistics.

C. Evans. 2013. Making Sense of Assessment Feedback in Higher Education. *Review of Educational Research*, 83(1):70–120.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *LREC*, volume 29, pages 31–43.

Jodi S. Goodman and Robert E. Wood. 2004. Feedback Specificity, Learning Opportunities, and Learning. *Journal of Applied Psychology*, 89(5):809–821.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. In *ACL workshop on Automatic information extraction and building of lexical semantic resources for NLP applications*.

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

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT-the GermaNet editing tool. In *Proceedings of the ACL 2010 System Demonstrations*, pages 19–24.

Roger A. Horn and Charles R. Johnson. 2013. *Matrix analysis*, second edition edition. Cambridge University Press, New York, NY.

Daniel R. Ilgen, Cynthia D. Fisher, and M. Susan Taylor. 1979. Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64(4):349–371.

Avraham N. Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2):254–284.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain agnostic real-valued specificity prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6610–6617.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2281–2287.

Junyi Jessy Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016. Improving the Annotation of Sentence Specificity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3921–3927, Portorož, Slovenia. European Language Resources Association (ELRA).

Anastasiya A. Lipnevich and Jeffrey K. Smith, editors. 2018. *The Cambridge handbook of instructional feedback*. Cambridge University Press, Cambridge, United Kingdom.

Annie Louis and Ani Nenkova. 2011. General versus specific sentences: automatic identification and application to analysis of news summaries. *Technical Reports (CIS)*.

Annie Louis and Ani Nenkova. 2012. A corpus of general and specific sentences from news. In *The International Conference on Language Resources and Evaluation (LREC)*, pages 1818–1821.

Luca Lugini and Diane Litman. 2017. Predicting Specificity in Classroom Discussion. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Catherine Mulryan-Kyne. 2010. Teaching large classes at college and university level: Challenges and opportunities. *Teaching in Higher Education*, 15(2):175–185.

Berry O'Donovan. 2017. How student beliefs about knowledge and knowing influence their satisfaction with assessment and feedback. *Higher Education*, 74(4):617–633.

Berry M. O'Donovan, Birgit den Outer, Margaret Price, and Andy Lloyd. 2019. What makes good feedback good? Studies in Higher Education, 1-12. *Studies in Higher Education*, pages 1–12.

Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. 2013. Tuned models of peer assessment in MOOCs. *arXiv*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ines Rehbein and Josef van Genabith. 2007. Treebank Annotation Schemes and Parser Evaluation for German. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189.

Karen Sparck Jones. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval: Journal of Documentation, 28(1), 11-21. *Journal of Documentation*, 28(1):11–21.

Jan-Willem Strijbos, Susanne Narciss, and Katrin Dünnebier. 2010. Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction*, 20(4):291–303.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument Mining: Extracting Arguments from Online Dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Stroudsburg, PA, USA. Association for Computational Linguistics.

Esther van Popta, Marijke Kral, Gino Camp, Rob L. Martens, and P. Robert-Jan Simons. 2017. Exploring the value of peer feedback in online learning for the provider. *Educational Research Review*, 20:24–34.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

John Michael Ziman. 1974. *Public knowledge: An essay concerning the social dimension of science*. Cambridge University Press, London.