Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale

Gabriella Skitalinskaya **Jonas Klaff**
Department of Computer Science
University of Bremen
Bremen, Germany
{gabski, joklaff}@uni-bremen.de

Henning Wachsmuth
Department of Computer Science
Paderborn University
Paderborn, Germany
henningw@upb.de

Abstract

Assessing the quality of arguments and of the claims the arguments are composed of has become a key task in computational argumentation. However, even if different claims share the same stance on the same topic, their assessment depends on the prior perception and weighting of the different aspects of the topic being discussed. This renders it difficult to learn topic-independent quality indicators. In this paper, we study claim quality assessment irrespective of discussed aspects by comparing different *revisions of the same claim*. We compile a large-scale corpus with over 377k claim revision pairs of various types from *kialo.com*, covering diverse topics from politics, ethics, entertainment, and others. We then propose two tasks: (a) assessing which claim of a revision pair is better, and (b) ranking all versions of a claim by quality. Our first experiments with embedding-based logistic regression and transformer-based neural networks show promising results, suggesting that learned indicators generalize well across topics. In a detailed error analysis, we give insights into what quality dimensions of claims can be assessed reliably. We provide the data and scripts needed to reproduce all results.[1]

1 Introduction

Assessing argument quality is as important as it is questionable in nature. On the one hand, identifying the good and the bad claims and reasons for arguing on a given topic is key to convincingly support or attack a stance in debating technologies (Rinott et al., 2015), argument search (Ajjour et al., 2019), and similar. On the other hand, argument quality can be considered on different granularity levels and from diverse perspectives, many of which are inherently subjective (Wachsmuth et al., 2017a); they depend on the prior beliefs and stance on a topic as well as on the personal weighting of different aspects of the topic (Kock, 2007).

Existing research largely ignores this limitation, by focusing on learning to predict argument quality based on subjective assessments of human annotators (see Section 2 for examples). In contrast, Habernal and Gurevych (2016) control for topic and stance to compare the convincingness of arguments. Wachsmuth et al. (2017b) abstract from an argument's text, assessing its relevance only structurally. Lukin et al. (2017) and El Baff et al. (2020) focus on personality-specific and ideology-specific quality perception, respectively, whereas Toledo et al. (2019a) asked annotators to disregard their own stance in judging length-restricted arguments. However, none of these approaches controls for the concrete aspects of a topic that the arguments claim and reason about. This renders it difficult to learn what makes an argument and its building blocks good or bad in general.

In this paper, we study quality in argumentation irrespective of the discussed topics, aspects, and stances by assessing different revisions of the basic building blocks of arguments, i.e., claims. Such revisions are found in large quantities on online debate platforms such as *kialo.com*, where users post claims, other users suggest revisions to improve claim quality (in terms of clarity, grammaticality, grounding, etc.), and moderators approve or disapprove them. By comparing the quality of different revisions of the same instance, we argue that we can learn general quality characteristics of argumentative text and, to a wide extent, abstract from prior perceptions and weightings.

To address the proposed problem, we present a new large-scale corpus, consisting of 124k unique claims from kialo.com spanning a diverse range of topics related to politics, ethics, and several others (Section 3). Using distant supervision, we derive a total number of 377k claim revision pairs from the

[1]Data and code:https://github.com/GabriellaSky/claimrev

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1718–1729
April 19 - 23, 2021.

Claim before Revision	Claim after Revision	Type
Dogs can help disabled people function better.	Dogs can help disabled people to navigate the world better.	Claim Clarification
African American soldiers joined unionists to fight for their freedom.	Black soldiers joined unionists to fight for their freedom.	Typo / Grammar Correction
Elections insure the independence of the judiciary.	Elections ensure the independence of the judiciary.	Typo / Grammar Correction
Israel has a track record of selling US arms to third countries without authorization.	Israel has a track record of selling US arms to third countries without authorization (*https://www.jstor.org/stable/1149008?seq=1#page_scan_tab_contents*).	Corrected / Added links

Table 1: Four examples of claims from Kialo before and after revision, along with the type of revision performed.

platform, each reflecting a quality improvement, often, with a specified revision type. Four examples are shown in Table 1. To the best of our knowledge, this is the first corpus to target quality assessment based on claim revisions. In a manual annotation study, we provide support for our underlying hypothesis that a revision improves a claim in most cases, and we test how much the revision types correlate with known argument quality dimensions.

Given the corpus, we study two tasks: (a) how to compare revisions of a claim by quality and (b) how to rank a set of claim revisions. As initial approaches to the first task, we select in Section 4 a "traditional" logistic regression model based on word embeddings as well as transformer-based neural networks (Vaswani et al., 2017), such as BERT (Devlin et al., 2019) and SBERT (Reimers and Gurevych, 2019). For the ranking task, we consider the Bradley-Terry-Luce model (Bradley and Terry, 1952; Luce, 2012) and SVMRank (Joachims, 2006). They achieve promising results, indicating that the compiled corpus allows learning topic-independent characteristics associated with the quality of claims (Section 5). To understand what claim quality improvements can be assessed reliably, we then carry out a detailed error analysis for different revision types and numbers of revisions.

The main contributions of our work are: (1) A new corpus for topic-independent claim quality assessment, with distantly supervised quality improvement labels of claim revision pairs, (2) initial promising approaches to the tasks of claim quality classification and ranking, and (3) insights into what works well in claim quality assessment and what remains to be solved.

2 Related Work

In the recent years, there has been an increase of research on the quality of arguments and the claims and reasoning they are composed of. Wachsmuth et al. (2017a) describe argumentation quality as a multidimensional concept that can be considered from a logical, rhetorical, and dialectical perspectives. To achieve a common understanding, the authors suggest a unified framework with 15 quality dimensions, which together give a holistic quality evaluation at a certain abstraction level. They point out, that several dimensions may be perceived differently depending on the target audience. In recent follow-up work, Wachsmuth and Werner (2020) examined how well each dimension can be assessed only based on plain text only.

Most existing quality assessment approaches target a single dimension. On mixed-topic student essays, Persing and Ng (2013) learn to score the clarity of an argument's thesis, Persing and Ng (2015) do the same for argument strength, and Stab and Gurevych (2017) classify whether an argument's premises sufficiently support its conclusion. All these are trained on pointwise quality annotations in the form of scores or binary judgments. Gretz et al. (2019) provide a corpus with crowdsourced quality annotations for 30,497 arguments, the largest to date for pointwise argument quality. The authors studied how their annotations correlate with the 15 dimensions from the framework of Wachsmuth et al. (2017a), finding that only *global relevance* and *effectiveness* are captured. Similarly, Lauscher et al. (2020) built a new corpus based on the framework to then exploit interactions between the dimensions in a neural approach. We present a small related annotation study for our dataset below. However, we follow Habernal and Gurevych (2016) in that we cast argument quality assessment as a relation classification problem, where the goal is to identify the better among a pair of instances.

In particular, Habernal and Gurevych (2016) created a dataset with argument convincingness pairs

on 32 topics. To mitigate annotator bias, the arguments in a pair always have the same stance on the same topic. The more convincing argument is then predicted using a feature-rich SVM and a simple bidirectional LSTM. Other approaches to the same task map passage representations to real-valued scores using Gaussian Process Preference Learning (Simpson and Gurevych, 2018) or represent arguments by the sum of their token embeddings (Potash et al., 2017), later extended by a Feed Forward Neural Network (Potash et al., 2019). Recently, Gleize et al. (2019) employed a Siamese neural network to rank arguments by the convincingness of evidence. In our experiments below, we take on some of these ideas, but also explore the impact of transformer-based methods such as BERT (Devlin et al., 2019), which have been shown to predict argument quality well (Gretz et al., 2019).

Potash et al. (2017) observed that longer arguments tend to be judged better in existing corpora, a phenomenon we will also check for below. Toledo et al. (2019b) prevent such bias in their corpora for both pointwise and pairwise quality, by restricting the length of arguments to 8–36 words. The authors define quality as the level of preference for an argument over other arguments with the same stance, asking annotators to disregard their own stance. For a more objective assessment of argument relevance, Wachsmuth et al. (2017b) abstract from content, ranking arguments only based on structural relations, but they employ majority human assessments for evaluation. Lukin et al. (2017) take a different approach, including knowledge about the personality of the reader into the assessment, and El Baff et al. (2020) study the impact of argumentative texts on people depending on their political ideology.

As can be seen, several approaches aim to control for length, stance, audience, or similar. However, all of them still compare argumentative texts with different content and meaning in terms of the aspects of topics being discussed. In this work, we assess quality based on different revisions of the same text. In this setting, the quality is primarily focused on how a text is formulated, which will help to better understand what influences argument quality in general, irrespective of the topic. To be able to do so, we refer to online debate portals.

Debate portals give users the opportunity to discuss their views on a wide range of topics. Existing research has used the rich argumentative content and structure of different portals for argument mining, including *createdebate.com* (Habernal and Gurevych, 2015), *idebate.org* (Al-Khatib et al., 2016), and others. Also, large-scale debate portal datasets form the basis of applications such as argument search engines (Ajjour et al., 2019). Unlike these works, we exploit debate portals for studying *quality*. Tan et al. (2016) predicted argument persuasiveness in the discussion forum *ChangeMyView* from ground-truth labels given by opinion posters, and Wei et al. (2016) used user upvotes and downvotes for the same purpose. Here, we resort to *kialo.com*, where users cannot only state argumentative claims and vote on the impact of claims submitted by others, but they can also help improve claims by suggesting revisions, which are approved or disapproved by moderators. While Durmus et al. (2019) assessed quality based on the impact value of claims from kialo.com, we derive information on quality from the revision history of claims.

The only work we are aware of that analyzes revision quality of argumentative texts is the study of Afrin and Litman (2018). From the corpus of Zhang et al. (2017) containing 60 student essays with three draft versions each, 940 sentence writing revision pairs were annotated for whether the revision improves essay quality or not. The authors then trained a random forest classifier for automatic revision quality classification. In contrast, instead of sentences, we shift our focus to claims. Moreover, our dataset is orders of magnitude larger and includes notably longer revision chains, which enables deeper analyses and more reliable prediction of revision quality using data-intensive methods.

3 Data

Here, we present our corpus created based on claim revision histories collected from *kialo.com*.

3.1 A New Corpus based on Kialo

Kialo is a typical example of an online debate portal for collaborative argumentative discussions, where participants jointly develop complex pro/con debates on a variety of topics. The scope ranges from general topics (religion, fair trade, etc.) to very specific ones, for instance, on particular policy-making (e.g., whether wealthy countries should provide citizens with a universal basic income). Each debate consists of a set of claims and is associated with a list of related pre-defined generic categories, such as politics, ethics, education, and entertainment.

What differentiates Kialo from other portals is

Corpus	Type of Instances	Instances
ClaimRev$_{BASE}$	**Total claim pairs**	**210 222**
	Claim Clarification	63 729
	Typo/Grammar Correction	59 690
	Corrected/Added Links	17 882
	Changed Meaning of Claim	1 178
	Misc	10 464
	None	57 279
ClaimRev$_{EXT}$	**Total claim pairs**	**377 659**
	Revision distance 1	77 217
	Revision distance 2	27 819
	Revision distance 3	10 753
	Revision distance 4	4 460
	Revision distance 5	2 055
	Revision distance 6+	2 008
Both Corpora	**Claim revision chains**	**124 312**

Table 2: Statistics of the two provided corpus versions. ClaimRev$_{BASE}$: Number of claim pairs in total and of each revision type. ClaimRev$_{EXT}$: Number of claim pairs in total and of each revision distance. The bottom line shows the number of unique revision chains in the corpora.

that it allows editing claims and tracking changes made in a discussion. All users can help improve existing claims by suggesting edits, which are then accepted or rejected by the moderator team of the debate. As every suggested change is discussed by the community, this collaborative process should lead to a continuous improvement of claim quality and a diverse set of claims for each topic.

As a result of the editing process, claims in a debate have a version history in the format of claim pairs, forming a chain where one claim is the successor of another and is considered to be of higher quality (examples found in Table 1). In addition, claim pairs may have a revision type label assigned to them via a non-mandatory free form text field, where moderators explain the reason of revision.

Base Corpus To compile the corpus, we scraped all 1628 debates found on Kialo until June 26th, 2020, related to over 1120 categories. They contain 124,312 unique claims along with their revision histories, which comprise of 210,222 pairwise relations. The average number of revisions per claim is 1.7 and the maximum length of a revision chain is 36. 74% of all pairs have a revision type. Overall, there are 8105 unique revision type labels in the corpus. 92% of labeled claim pairs refer to three types only: *Claim Clarification*, *Typo/Grammar Correction*, and *Corrected/Added Links*. An overview of the distribution of revision labels is given in Table 2. We refer to the resulting corpus as *ClaimRev$_{BASE}$*.

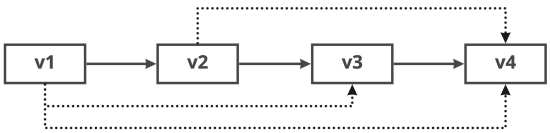

Figure 1: Visual representation of relations between revisions. Solid and dashed lines denote original and inferred non-consecutive relations respectively.

Data pre-processing included removing all claim pairs from debates carried out in languages other than English. Also, we considered claims with less than four characters as uninformative and left them out. As we seek to compare different versions of the *same* claim, claim version pairs with a general change of meaning do not satisfy this description. Thus, we removed such pairs from the corpus, too (inspecting the data revealed that such pairs were mostly generated due to debate restructuring). For this, we assessed the cosine similarity of a given claim pair using *spacy.io* and remove a pair if the score is lower than the threshold of 0.8.

Extended Corpus To increase the diversity of data available for training models, without actually collecting new data, we applied data augmentation. ClaimRev$_{BASE}$ consists of consecutive claim version pairs, i.e., if a claim v has four versions, it will be represented by three three pairs: (v_1, v_2), (v_2, v_3), and (v_3, v_4), where v_1 is the original claim and v_4 is the latest version. We extend this data by adding all pairs between non-consecutive versions that are inferrable transitively. Considering the previous example, this means we add (v_1, v_3), (v_1, v_4), and $(v2, v4)$. This is based on our hypothesis that every argument version is of higher quality than its predecessors, which we come back to below. Figure 1 illustrates the data augmentation. We call the augmented corpus *ClaimRev$_{EXT}$*.

For this corpus, we introduce the concept of *revision distance*, by which we mean the number of revisions between two versions. For example, the distance between v_1 and v_2 would be 1, whereas the distance between v_1 and v_3 would be 2. The distribution of the revision distances across ClaimRev$_{EXT}$ is summarized in Table 2.

The number of claim pairs of the 20 most frequent categories in both corpus versions are presented in Figure 2. We will restrict our view to the topics in these categories in our experiments.

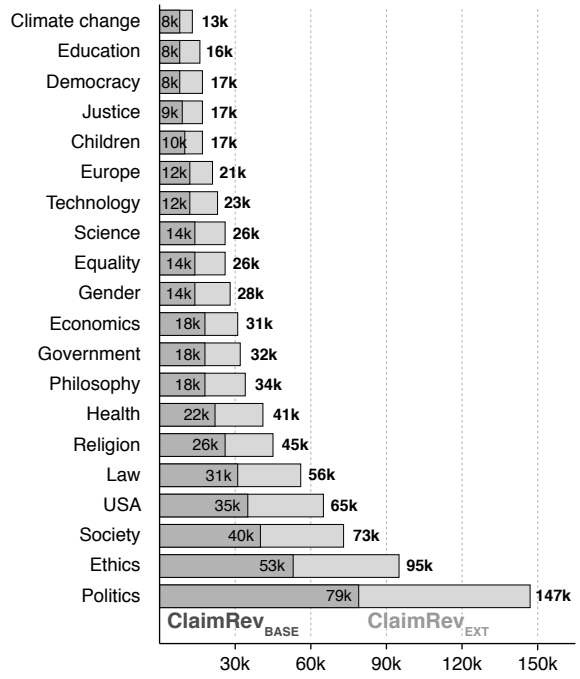

Figure 2: Number of claim revision pairs in each debate category of the two provided versions of our corpus (ClaimRev$_{\text{BASE}}$, ClaimRev$_{\text{EXT}}$).

3.2 Data Consistency on Kialo

While collaborative content creation enables leveraging the wisdom of large groups of individuals toward solving problems, it also poses challenges in terms of quality control, because it relies on varying perceptions of quality, backgrounds, expertise, and personal objectives of the moderators. To assess the consistency of the distantly-supervised corpus annotations, we carried out two annotation studies on samples of our corpus.

Consistency of Relative Quality In this study, we aimed to capture the general perception of claim quality on a meta-level, by deriving a data-driven quality assessment based on the revision histories. This was based on our hypothesis that every claim version is better than its predecessor. To test the validity of this hypothesis, two authors of this paper annotated whether a revision increases, decreases, or does not affect the overall claim quality. For this purpose, we randomly sampled 315 claim revision pairs, found in the supplementary material.

The results clearly support our hypothesis, showing an increase in quality in 292 (93%) of the annotated cases at a Cohen's κ agreement of 0.75, while 8 (3%) of the revisions had no effect on quality and only 6 (2%) led to a decrease. On the remaining 2%, the annotators did not reach an agreement.

Consistency of Revision Type Labels Our second annotation study focused on the reliability of the revision type labels. We restricted our view to the top three revision labels, which cover 96% of all revisions. We randomly sampled 140–150 claim pairs per each revision type, 440 in total. For each claim pair, the same annotators as above provided a label for the revision type from the following set: *Claim Clarification*, *Typo/Grammar Correction*, *Corrected/Added Links*, and *Other*.

Comparing the results to the original labels in the corpus revealed that the annotators strongly agreed with the labels, namely, with Cohen's κ of 0.82 and 0.76 respectively. The level of agreement between the annotators was even higher ($\kappa = 0.84$). In further analysis, we observed that most confusion happened between the revision types *Typo/Grammar correction* and *Claim Clarification*. This may be due to the non-strict nature of the revision type labels, which leaves space for different interpretations on a case-to-case basis. Still, we conclude that the revision type labels seem reliable in general.

3.3 Quality Dimensions on Kialo

To explore the relationship between the revision types on Kialo and argument quality in general, we conducted a third annotation study. In particular, for each of the 315 claim pairs from Section 3.2, one of the authors of this paper provided a label indicating whether the revision improved for each of the 15 quality dimensions defined by Wachsmuth et al. (2017a) or not. It should be noted that the annotators reached an agreement on the revision type for all these pairs.

Table 3 shows Pearson's r rank correlation for each quality dimension for the three main revision types. We observe a strong correlation between the revision type *Corrected/Added Links* and the logical quality dimensions *Cogency* (0.65) and *Local Sufficiency* (0.62), which matches the main purpose of such revisions: to add supporting information to a claim. The high negative correlation of this revision type with *Global Acceptability* (-0.82) indicates that improvements regarding the dimension in question are more prominent in other types. Complementarily, *Claim Clarification* mainly improves the other logical dimensions (*Local Acceptability* 0.38, *Local Relevance* 0.44), matching the intuition that a clarification helps to ensure a correct understanding of the meaning. *Typo/Grammar corrections*, finally, rather seem to support an accept-

	Clarification	Grammar	Links
Cogency	**-0.31**	**-0.31**	**0.65**
Local Acceptability	**0.38**	-0.20	-0.19
Local Relevance	**0.44**	-0.25	-0.22
Local Sufficiency	-0.28	**-0.33**	**0.62**
Effectiveness	0.02	**-0.35**	**0.34**
Credibility	0.06	-0.16	0.10
Emotional Appeal	0.00	0.00	0.00
Clarity	-0.16	**0.35**	-0.18
Appropriateness	0.01	0.02	-0.04
Arrangement	0.00	0.00	0.00
Reasonableness	0.07	-0.04	-0.04
Global Acceptability	**0.37**	**0.42**	**-0.82**
Global Relevance	0.02	**-0.43**	**0.42**
Global Sufficiency	0.00	0.00	0.00
Overall	-0.05	0.00	0.05
Pairs with revision type	120	100	95

Table 3: Pearson's r correlation in our annotation study between increases in the 15 quality dimensions of Wachsmuth et al. (2017a) and the main revision types: Claim *Clarification*, Typo/*Grammar* Correction, Corrected/Added *Links*. Moderate and high correlations are shown in bold ($r \geq 0.3$).

able linguistic shape, improving *Clarity* (0.35) and *Global Acceptability* (0.42).

Finding only low correlations for many rhetorical dimensions (credibility, emotional appeal, etc.) as well as for overall quality, we conclude that the revisions on Kialo seem to target primarily the general form a well-phrased claim should have.

4 Approaches

To study the two proposed tasks, claim quality classification and claim quality ranking, on the given corpus, we consider the following approaches.

4.1 Claim Quality Classification

We cast this task as a pairwise classification task, where the objective is to compare two versions of the same claim and determine which one is better. To solve this task, we compare four methods:

Length To check whether there is a bias towards longer claims in the data, we use a trivial method which assumes that claims with more characters are better.

S-BOW As a "traditional" method, we employ the siamese bag-of-words embedding (S-BOW) as described by Potash et al. (2017). We concatenate two bag-of-words matrices, each representing a claim version from a pair, and input the concatenated matrix to a logistic regression. We also test whether information on length improves S-BOW.

	v_1	v_2	v_3
v_1	0	0.018	0.002
v_2	0.982	0	0.428
v_3	0.998	0.572	0

Table 4: Example of a pairwise score matrix for ranking of three claim revisions, v_1–v_3, given the following pairwise scores: $(v_1, v_2) = (0.018, 0.982)$, $(v_2, v_3) = (0.428, 0.572)$, and $(v_1, v_3) = (0.002, 0.998)$.

BERT We select the BERT model, as it has become the standard neural baseline. BERT is a pre-trained deep bidirectional transformer language model (Devlin et al., 2019). For our experiments we use the pre-trained version *bert-base-cased*, as implemented in the *huggingface* library.[2] We fine-tune the model for two epochs using the Adam optimizer with learning rate 1e-5. [3]

SBERT We also use Sentence-BERT (SBERT) to learn to represent each claim version as a sentence embedding (Reimers and Gurevych, 2019), opposed to the token-level embeddings of standard BERT models. We fine-tune SBERT based on *bert-base-cased* using a siamese network structure, as implemented in the *sentence-transformers* library.[4] We set the numbers of epochs to one which is recommended by the authors (Reimers and Gurevych, 2019), and we use a batch-size of 16, Adam optimizer with learning rate 1e-5, and a linear learning rate warm-up over 10% of the training data. Our default pooling strategy is MEAN.

4.2 Claim Quality Ranking

In contrast to the previous task, we cast this problem as a sequence-pair regression task. After obtaining all pairwise scores using S-BOW, BERT, and SBERT respectively, we map the pairwise labels to real-valued scores and rank them using the following models, once for each method.

BTL For mapping, we use the well-established Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 2012), in which items are ranked according to the probability that a given item beats an item chosen randomly. We feed the BTL model a pairwise-comparison matrix for all revisions related to a claim, generated as follows: Each row

[2]Huggingface library, https://huggingface.co/transformers/pretrained_models.html

[3]We chose the number of epochs empirically, picking the best learning rate out of {5e-7, 5e-6,1e-5,2e-5,3e-5}.

[4]Sentence-transformers library, https://www.sbert.net/

	Test set: ClaimRev$_{BASE}$				Test set: ClaimRev$_{EXT}$			
	Random-Split		Cross-Category		Random-Split		Cross-Category	
Model	**Accuracy**	**MCC**	**Accuracy**	**MCC**	**Accuracy**	**MCC**	**Accuracy**	**MCC**
Length	61.3 / 61.3	0.23 / 0.23	60.7 / 60.7	0.21 / 0.21	60.8 / 60.8	0.22 / 0.22	60.0 / 60.0	0.20 / 0.20
SBOW	62.0 / 62.6	0.24 / 0.25	61.4 / 61.4	0.23 / 0.23	64.9 / 65.4	0.30 / 0.31	63.9 / 64.1	0.28 / 0.28
SBOW + Length	65.1 / 65.5	0.30 / 0.31	64.8 / 64.4	0.29 / 0.29	67.1 / 67.5	0.34 / 0.35	66.1 / 66.2	0.32 / 0.32
BERT	75.5 / 75.2	0.51 / 0.51	75.1 / 74.1	**0.51** / 0.49	76.4 / 76.5	0.53 / 0.53	76.2 / 75.4	0.53 / 0.51
SBERT	**76.2 / 76.2**	**0.53 / 0.52**	**75.5 / 75.4**	**0.51 / 0.51**	**77.4 / 77.7**	**0.55 / 0.55**	**76.8 / 76.8**	**0.54 / 0.54**
Random baseline	50.0 / 50.0	0.00 / 0.00	50.0 / 50.0	0.00 / 0.00	50.0 / 50.0	0.00 / 0.00	50.0 / 50.0	0.00 / 0.00
Single claim baseline	57.7 / 58.1	0.17 / 0.17	57.7 / 57.3	0.17 / 0.16	58.8 / 59.8	0.20 / 0.20	58.9 / 58.9	0.20 / 0.20

Table 5: Claim quality classification results: Accuracy and Matthew Correlation Coefficient (MCC) for all tested approaches in the random-split and the cross-category setting on the two corpus versions. The first value in each value pair is obtained by a model trained on ClaimRev$_{BASE}$, the second by a model trained on ClaimRev$_{EXT}$. All improvements from one row to the next are significant at $p < 0.001$ according to a two-sided Student's t-test.

represents the probability of the revision being better than other revisions. All diagonal values are set to zero. Table 4 illustrates an example for a set of three argument revisions.

SVMRank Additionally, we employ SVMRank (Joachims, 2006), which views the ranking problem as a pairwise classification task. First, we change the input data, provided as a ranked list, into a set of ordered pairs, where the (binary) class label for every pair is the order in which the elements of the pair should be ranked. Then, SVMRank learns by minimizing the error of the order relation when comparing all possible combinations of candidate pairs. Given the nature of the algorithm we cannot work with token embeddings obtained from BERT directly. Thus, we utilize one of most commonly used approaches to transform token embeddings to a sentence embedding: extracting the special [CLS] token vector (Reimers and Gurevych, 2019; May et al., 2019). In our experiments we select a linear kernel for the SVM and use PySVMRank,[5] a python API to the SVM^{rank} library written in C.[6]

5 Experiments and Discussion

We now present empirical experiments with the approaches from Section 4. The goal is to evaluate how hard it is to compare and rank the claim revisions in our corpus from Section 3 by quality.

5.1 Experimental Setup

We carry out experiments in two settings. The first considers creating *random splits* over revision histories, ensuring that all versions of the same claim are in a single split in order to avoid data leakage. We assign 80% of the revision histories to the training set and the remaining 20% to the test set. A drawback of this setup is that it is not clear how well models generalize to unseen debate categories. In the second setting, we therefore evaluate the methods also in a *cross-category* setup using a leave-one-category-out paradigm, which ensures that all claims from the same debate category are confined to a single split. We split the data in this way to evaluate if our models learn independent features that are applicable across the diverse set of categories. To assess the effect of adding augmented data, we evaluate all models on both ClaimRev$_{BASE}$ and ClaimRev$_{EXT}$.

For quality *classification*, we report accuracy and the Matthews correlation coefficient (Matthews, 1975). We report the mean results over five runs in the random setting and the mean results across all test categories in the cross-category setting. To ensure balanced class labels, we create one false claim pair for each true claim pair by shuffling the order of the claims: $(v_1, v_2, true) \rightarrow (v_2, v_1, false)$, where the label denotes whether the second claim in the pair is of higher quality. We report results obtained by models trained on ClaimRev$_{BASE}$ and ClaimRev$_{EXT}$ as score pairs in Table 5.

To measure *ranking* performance, we calculate Pearson's r and Spearman's ρ correlation, as well as NDCG and MRR. We also compute the Top-1 accuracy, i.e. the proportion of claim sets, where the latest version has been ranked best. We average the results on each claim set across the test set for each metric. Afterwards we average the results across five runs or across all categories, depending on the chosen setting.

[5] PySVMRank, https://github.com/ds4dm/PySVMRank

[6] SVM^{rank}, www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

Model	Random-Split					Cross-Category				
	r	ρ	**Top-1**	NDCG	MRR	r	ρ	**Top-1**	NDCG	MRR
BTL + SBOW+L	0.38	0.37	0.62	0.94	0.79	0.36	0.35	0.60	0.94	0.78
BTL + BERT	0.60	0.59	0.74	0.96	0.86	0.58	0.57	0.72	0.96	0.85
BTL + SBERT	0.63	0.62	0.77	**0.97**	0.87	0.62	0.61	0.75	0.97	0.86
SVMRank + SBOW+L	0.18	0.18	0.50	0.93	0.73	0.24	0.23	0.52	0.93	0.75
SVMRank + BERT CLS	0.50	0.49	0.67	0.95	0.84	0.51	0.51	0.67	0.96	0.84
SVMRank + SBERT	**0.70**	**0.70**	**0.79**	**0.97**	**0.90**	**0.73**	**0.72**	**0.80**	**0.98**	**0.91**
Random baseline	0.00	0.00	0.42	0.91	0.68	0.00	0.00	0.42	0.91	0.67

Table 6: Claim quality ranking results: Pearson's r and Spearman's ρ correlation as well as top-1 accuracy for all tested approaches in the random-split and the cross-category setting on ClaimRev$_{\text{EXT}}$.In all cases, SVMRank + SBERT is significantly better than all others at $p < 0.001$ according to a two-sided Student's t-test.

5.2 Claim Quality Classification

The results in Table 5 show that a claim's *length* is a weak indicator of quality (up to 61.3 accuracy). An intuitive explanation is that, even though claims with more information may be better, it is also important to keep them readable and concise.

Despite *SBOW*'s good performance on predicting convincingness (Potash et al., 2017), the claim quality in our corpus cannot be captured by a model of such simplicity (maximum accuracy of 65.4). We point out that adding other linguistic features (for example, part-of-speech tags or sentiment scores) may further improve SBOW. Exemplarily, we equip SBOW with length features and observe a significant improvement (up to 67.5).

As for the transformer-based methods, we see that *BERT* and *SBERT* consistently outperform SBOW in all settings on both corpus versions, with SBERT's accuracy of up to 77.7 being best.[7]

A comparison of the performance of the methods depending on the corpus used for training in Table 5 shows the effect of augmenting the original Kialo data. In most cases, the results obtained by models trained on ClaimRev$_{\text{EXT}}$ are comparable (slightly higher/lower) than results obtained by models trained on ClaimRev$_{\text{BASE}}$. This means that adding relations between non-consecutive claim versions does not improve the reliability of methods. Given that the performance scores obtained on the ClaimRev$_{\text{EXT}}$ test set are evidently higher than on the ClaimRev$_{\text{BASE}}$ test set, we can conclude that the augmented cases are easier to classify and the cumulative difference in quality is more evident.

[7]Additionally, we have experimented with an adversarial training algorithm, ELECTRA (Clark et al., 2020), and obtained results slightly better than BERT, yet inferior to SBERT. We omit to report these results here, since they did not provide any further notable insights.

We can also see in Table 5 that the trained models are able to generalize across categories; the accuracy and MCC scores in the random split and cross-category settings for each method are very similar, with only a slight drop in the cross-category setting. This indicates that the nature of the revisions is relatively consistent among all categories, yet reveals the existence of some category-dependent features.

To find out whether BERT really captures the relative revision quality and not only lexical features present in the original claim, we introduced a *Single claim* baseline, analogous to the *hypothesis-only* baseline in natural language inference(Poliak et al., 2018). It can be seen that the accuracy and MCC scores are low across all settings (maximum accuracy of 59.8), which indicates that BERT indeed captures relative revision quality mostly.

5.3 Claim Quality Ranking

Table 6 lists the results of our ranking experiments, which show patterns similar to the results achieved in the classification task.

We can observe similar patterns in both of the selected ranking approaches: SBERT consistently outperforms all other considered approaches across all settings (up to 0.73 and 0.72 in Pearson's r and Spearman's ρ accordingly). BERT and SBERT outperform SBOW, indicating that transformer-based methods are more capable of capturing the relative quality of revisions. While BTL + BERT obtains results comparable to BTL + SBERT, we find that using the CLS-vector as a sentence embedding representation leads to lower results. We point out, though, that using other sentence embeddings and/or pooling strategies (for example, averaged BERT embeddings) may further improve results.

Similar to the results of the classification task, we observe only a slight performance drop in the

Task	Label	Accuracy	Instances
Type	Claim Clarification	69.7	12 856
	Typo/Grammar Correction	83.6	12 125
	Corrected/Added Links	89.3	3 660
	Changed Meaning of Claim	57.3	232
	Misc	67.2	2 130
	None	78.3	45 842
Distance	Revision distance 1	76.2	42 341
	Revision distance 2	79.6	17 478
	Revision distance 3	80.6	8 023
	Revision distance 4	81.0	3 979
	Revision distance 5	79.5	2 103
	Revision distance 6+	74.9	2 921
	All	**77.7**	**76 845**

Table 7: Accuracy of the best model, SBERT, on each single revision type and distance in ClaimRev$_{\text{EXT}}$, along with the number of instances per each case.

cross-category setting when using BTL for ranking, yet an increase when using SVMRank, again emphasizing the topic-independent nature of claim quality in our corpus.

5.4 Error Analysis

To further explore the capabilities and limitations of the best model, SBERT, we analyzed its performance on each revision type and distance.

As the upper part of Table 7 shows, SBERT is highly capable of assessing revisions related to the correction and addition of links and supporting information. This revision type also obtained the highest correlations between quality dimensions and type of revision (see Table 3), which indicates that the patterns of changes performed within this type are more consistent. In contrast, we observe that the model fails to address revisions related to the changed meaning of a claim. On the one hand, this may be due to the fact that such examples are underrepresented in the data. On the other hand, the consideration of such examples in the selected tasks is questionable, since changing the meaning of claim is usually considered as the creation of a *new claim* and not a *new version* of a claim.

An insight from the lower part of Table 7 is that the accuracy of predictions increases from revision distance 1 to 4. We obtain better results when comparing non-consecutive claims than when comparing claim pairs with distance of 1. An intuitive explanation is that, since each single revision should ideally improve the quality of a claim, the more revisions a claim undergoes, the more evident the quality improvement should be. For distances > 5, the accuracy starts to decrease again, but this may be due to the limited number of cases given.

6 Conclusion and Future Work

In this paper, we have proposed a new way of assessing quality in argumentation by considering different revisions of the same claim. This allows us to focus on characteristics of quality regardless of the discussed topics, aspects, and stances in argumentation. We provide a new corpus of web claims, which is the first large-scale corpus to target quality assessment and revision processes on a claim level. We have carried out initial experiments on this corpus using traditional and transformer-based models, yielding promising results but also pointing to limitations. In a detailed analysis we have studied different kinds of claim revisions and provided insights into the aspects of a claim that influence the users' perception of quality. Such insights could help improve writing support in educational settings, or identify the best claims for debating technologies and argument search.

We seek to encourage further research on how to help online debate platforms automate the process of quality control and design automatic quality assessment systems. Such systems can be used to indicate if the suggested revisions increase the quality of an argument or recommend the type of revision needed. We leave it for future work to investigate whether the learned concepts of quality are transferable to content from other collaborative online platforms (such as idebate.org or Wikipedia), or to data from other domains, such as student essays and forum discussions.

Acknowledgments

We thank Andreas Breiter for feedback on early drafts, and the anonymous reviewers for their helpful comments. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number 374666841, SFB 1342.

References

Tazin Afrin and Diane Litman. 2018. Annotation and classification of sentence-level revision improvement. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–246, New Orleans, Louisiana. Association for Computational Linguistics.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein.

2019. Data acquisition for argument search: The args.me corpus. In *KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings*, pages 48–59.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. The role of pragmatic and discourse context in determining argument impact. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5668–5678, Hong Kong, China. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the persuasive effect of style in news editorial argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis. *arXiv preprint arXiv:1911.11408*.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 217–226, New York, NY, USA. Association for Computing Machinery.

Christian Kock. 2007. Dialectical obligations in political debate. *Informal Logic*, 27(3):233–247.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

R Duncan Luce. 2012. *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753. Association for Computational Linguistics.

B.W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442 – 451.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Peter Potash, Robin Bhattacharya, and Anna Rumshisky. 2017. Length, interchangeability, and external knowledge: Observations from predicting argument convincingness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 342–351, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Peter Potash, Adam Ferguson, and Timothy J. Hazen. 2019. Ranking passages for argument convincingness. In *Proceedings of the 6th Workshop on Argument Mining*, pages 146–155, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019a. Automatic argument quality assessment - New datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635. Association for Computational Linguistics.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019b. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017b. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745, Barcelona,

Spain (Online). International Committee on Computational Linguistics.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Berlin, Germany. Association for Computational Linguistics.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578, Vancouver, Canada. Association for Computational Linguistics.