Modeling Appropriate Language in Argumentation

Timon Ziegenbein
Leibniz University Hannover
t.ziegenbein@ai.uni-hannover.de

Shahbaz Syed
Leipzig University
shahbaz.syed@uni-leipzig.de

Felix Lange
Paderborn University
flange@mail.uni-paderborn.de

Martin Potthast
Leipzig University and ScaDS.AI
martin.potthast@uni-leipzig.de

Henning Wachsmuth
Leibniz University Hannover
h.wachsmuth@ai.uni-hannover.de

Abstract

Online discussion moderators must make ad-hoc decisions about whether the contributions of discussion participants are *appropriate* or should be removed to maintain civility. Existing research on offensive language and the resulting tools cover only one aspect among many involved in such decisions. The question of what is considered appropriate in a controversial discussion has not yet been systematically addressed. In this paper, we operationalize appropriate language in argumentation for the first time. In particular, we model appropriateness through the absence of flaws, grounded in research on argument quality assessment, especially in aspects from rhetoric. From these, we derive a new taxonomy of 14 dimensions that determine inappropriate language in online discussions. Building on three argument quality corpora, we then create a corpus of 2191 arguments annotated for the 14 dimensions. Empirical analyses support that the taxonomy covers the concept of appropriateness comprehensively, showing several plausible correlations with argument quality dimensions. Moreover, results of baseline approaches to assessing appropriateness suggest that all dimensions can be modeled computationally on the corpus.

1 Introduction

People have varying degrees of sensitivity to controversial issues and may be triggered by different emotional responses dependent on the issue and the opponents' arguments (Walton, 2010). This often makes it hard to maintain a constructive discussion. In competitive debates, a moderator ensures that participants argue *appropriately*. Debating culture, dating back to the 18th century, demands appropriate behavior, such as staying on topic and avoiding overly emotional language (Andrew, 1996). Accordingly, Wachsmuth et al. (2017b) define arguments to be appropriate if they support credibility and emotions and match the issue.

Appropriate Argument

"There is scientific evidence that shows having a mother and a father is the healthiest way for a child to progress physically and mentally. So a lousy father is better than none (that is of course assuming that he is not abusive in any way). Also, people change. Who says that he will be lousy forever. There are family therapy sessions you can attend to help."

[Issue: Is it better to have a lousy father or to be fatherless?]

Inappropriate Argument

"for everyone who is talking about RAPE in this subject let me ask you one thing!!!! if you got in a huge fight with someone and ended up breaking your hand or arm... would you cut it off just because it would REMIND you of that expirience??? if your actualy SANE you would say no and if you say yes you need to see a Physiatrist!!!!"

[Issue: Pro choice vs pro life]

Figure 1: Two arguments from the corpus introduced in this paper, one appropriate and one inappropriate. The used colors match the taxonomy concepts we present in Section 3: toxic intensity (dark red), unclear meaning (orange), and missing openness (light purple).

Similarly, in many online forums, moderators ensure a certain level of civility in the discussions. What arguments are considered civil may differ from community to community. The task of discussion moderation thus requires ad-hoc decisions about the appropriateness of any contributed argument, calling out the inappropriate ones—a challenging task to master. Moreover, the amount of moderation required on the web necessitates automation of this task, as the resources for manual moderation are usually insufficient.

Figure 1 shows two exemplary arguments, assessed by human annotators. The inappropriate

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4344–4363

argument appeals excessively to emotions, is not easily understandable, and shows little interest in the opinion of others. Note that the last sentence of the argument is also a personal attack, a special case of inappropriate emotional language. Hence, multiple inappropriateness aspects can occur at the same time. The appropriate argument, on the other hand, does not contain any of these issues.

Most previous work on automatic content moderation has focused on detecting offensive content (Schmidt and Wiegand, 2017; Poletto et al., 2021). However, to create a climate in which controversial issues can be discussed constructively, combating only offensive content is not enough, since there are also many other forms of inappropriate arguments (Habernal et al., 2018). While the notion of appropriateness is treated in argumentation theory as an important subdimension of argument quality (see Section 2), there has been no systematic study of appropriateness, let alone a clear definition or operationalization. These shortcomings hinder the development of automatic moderation tools.

In this paper, we present a taxonomy of 14 inappropriateness dimensions, systematically derived from rhetoric (Burkett, 2011) and argument quality theory (Wachsmuth et al., 2017b), along with a corpus annotated for the dimensions. Matching elements of the concept of reasonableness by van Eemeren (2015), we argue appropriateness to be a minimal quality property that is necessary for any argument to consider it valuable in a debate.

We motivate the 14 dimensions empirically in Section 3 by analyzing interactions of low appropriateness with other quality issues of arguments, and we further refine the dimensions on this basis. To operationalize the taxonomy, we create a new corpus of 2191 arguments from debates, question-answering forums, and reviews (Section 4). The arguments are compiled from three existing argument quality corpora (Habernal and Gurevych, 2016a; Wachsmuth et al., 2017b; Ng et al., 2020), such that they cover both a variety of topics and selected topics in depth. All arguments are manually labeled for the dimensions in a human annotation study.

Given the new corpus, we analyze correlations between the 14 dimensions and the argument quality dimensions in the source corpora in Section 5. Several plausible correlations support that our taxonomy successfully aligns with the theoretical and practical quality aspects modeled in previous work. To gain insights into how well the proposed dimensions can be predicted automatically, we also evaluate first baseline approaches to the computational assessment of appropriateness (Section 6). The results do not fully compete with the average human performance. However, they show large improvements over basic baselines on all dimensions while suggesting that a semantic understanding of arguments is required for the task.

Altogether, this paper's main contributions are:[1]

- A theory-based taxonomy that specifies inappropriate language in online discussions
- A corpus with 2191 arguments from three different genres, manually annotated for the 14 taxonomy dimensions
- Empirical insights into the relation of appropriateness to previously studied quality dimensions and into its computational predictability

2 Related Work

The notion of appropriateness has been explored in several sub-disciplines of linguistics. In communicative competence research, Hymes et al. (1972) considered the knowledge about cultural norms as a requirement to produce appropriate speech, which is a central part of acquiring communicative competence. Defining sociolinguistics, Ranney (1992) linked appropriateness to the notion of politeness that is required in various social settings. Later, Schneider (2012) argued that appropriateness is a more salient notion than politeness as it explicitly accounts for the context. Some of these cultural speech properties were identified as linguistic etiquette by Jdetawy and Hamzah (2020), including correct, accurate, logical, and pure language.

Regarding the discussion of controversial issues, debating culture has required participants since its origins to stay on topic and to avoid offensive and overly emotional formulations (Andrew, 1996). Likewise, Blair (1988) differentiate between good and bad bias in argumentation, where the latter exhibits close-mindedness, distortion of the conversation, or an imbalance of pro and con arguments. Similarly, Walton (1999) introduced the concept of dialectical bias, explicitly addressing the context in which an argument is judged to be appropriate. This perspective on argumentation is also described by Burkett (2011) as "[...] making appropriate choices in light of situation and audience."

[1]The corpus and experiment code can be found under: https://github.com/webis-de/ACL-23

As a sub-dimension of argument quality, appropriateness was first studied in NLP by Wachsmuth et al. (2017b), a significant inspiration for our work. The authors derived appropriateness as one of the rhetorical argument quality dimensions based on the work of Aristotle (Aristotle, 2007). While several of the quality dimensions they proposed were addressed explicitly in previous work, the appropriateness dimension has not been systematically assessed until now. Wachsmuth et al. (2017b) only provided a relatively shallow definition of appropriateness that requires a simultaneous assessment of three properties, namely the creation of *credibility* and *emotions* as well as *proportionality* to the issue. In contrast, we model these properties individually (in addition to several other dimensions) to better understand what exactly impacts appropriateness.

Computationally, only Wachsmuth and Werner (2020) tried to predict appropriateness alongside all the other quality dimensions of Wachsmuth et al. (2017b). However, their models relied on a rather small sample of 304 arguments. In comparison, our corpus consists of 2191 arguments spanning three argumentative genres, providing deeper insights into the appropriateness of an argument. Related to this notion is the convincingness of arguments studied by Habernal and Gurevych (2016a,b) which correlates with appropriateness (Wachsmuth et al., 2017a), as well as the effectiveness of arguments (Ng et al., 2020; Lauscher et al., 2020).

In the context of appropriateness, Walton (2010) explored the notion of emotional fallacies in reasoning, some of which were later assessed computationally (Habernal et al., 2017; Alhindi et al., 2022; Jin et al., 2022; Goffredo et al., 2022). Although we consider some of these fallacies in our work, we also consider other dimensions and exclude some irrelevant to appropriateness (i.e., logical fallacies) because of their more technical nature.

We model *toxic emotions* based on the emotional fallacies identified by Walton (2010): ad populum, ad misericordiam, ad baculum, and ad hominem. We merged these four into a single sub-dimension called *emotional deception* based on the results of a pilot annotation study (Section 4). Additionally, we define a sub-dimension *excessive intensity* to address overly intense emotions. In particular, our analysis revealed the presence of a subset of propaganda errors, including loaded language, flag-waving, repetition, exaggeration, and minimization Da San Martino et al. (2020).

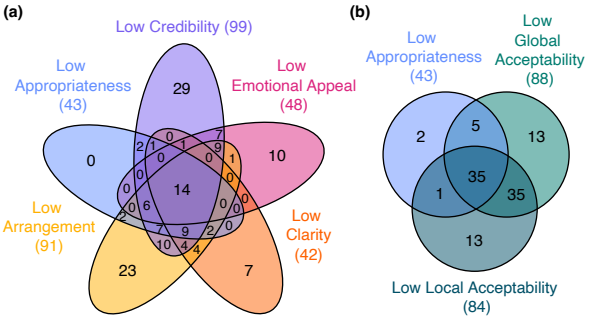


Figure 2: Venn diagrams showing the absolute counts of low-quality arguments in the corpus of Wachsmuth et al. (2017b) in terms of appropriateness and other dimensions: (a) The sub-dimensions of rhetorical effectiveness. (b) Local acceptability and global acceptability.

3 Modeling Appropriateness

This section explains how we established the relevant dimensions of appropriateness by systematically analyzing research on argument quality.

3.1 Appropriateness and Argument Quality

To learn what makes an argument *(in)appropriate*, we analyzed the interaction of appropriateness with other quality dimensions in the 304 arguments of Wachsmuth et al. (2017b). We selected the dimensions that correlated most with appropriateness according to Pearson's r. These include the four sub-dimensions of rhetorical effectiveness (besides appropriateness), namely, *credibility* (.49), *emotional appeal* (.30), *clarity* (.45), and *arrangement* (.48), as well as *local acceptability* (.54) (sub-dimension of logical cogency) and *global acceptability* (.59) (sub-dimension of dialectical reasonableness). We then counted the number of arguments with the lowest quality rating for both appropriateness and the other dimensions as we expected the most notable differences in those instances.

Figure 2 illustrates the absolute cooccurrence of flawed arguments for the selected dimensions. Uniquely, appropriateness flaws always occur with at least one other flawed rhetorical dimension in all 43 cases, and low acceptability in nearly all cases.

Consequently, we manually analyzed arguments by contrasting pairs of arguments with and without low appropriateness to find patterns that describe what drives the low appropriateness levels within these dimensions. For example, to model the overlap of appropriateness with credibility, we compared the 29 arguments with only low credibility in Figure 2 (a) to the 39 ($= 2 + 1 + 6 + 14 + 7 + 9$) arguments with low appropriateness and credibility.

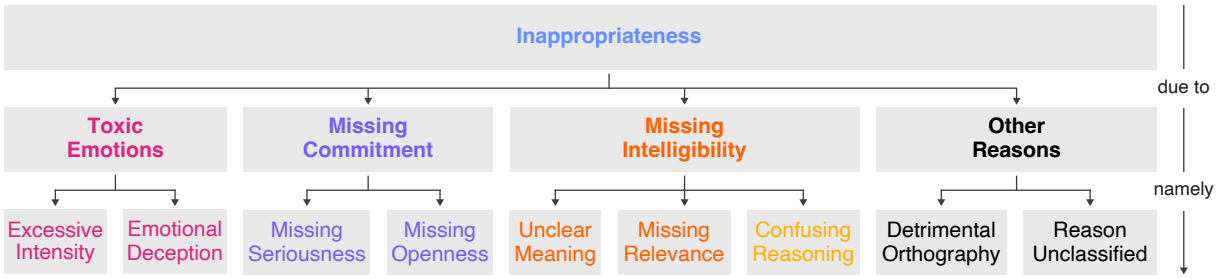


Figure 3: Proposed taxonomy of inappropriate language in argumentation, with 14 dimensions and sub-dimensions. The colors are aligned with the argument quality dimensions used to derive them (Figure 2).

Concretely, we compared them incrementally, starting from arguments that do not have low values in any quality dimension except appropriateness and credibility, proceeding to those with exactly one other low value, and so forth until we reach the 14 arguments that have low values in all dimensions.

3.2 Defining Inappropriateness

The findings from our analysis led to four core inappropriateness dimensions in our taxonomy: We deem an argument *inappropriate* (in light of its discussion context) if it is *missing commitment* of its author to the discussion, uses *toxic emotions*, is *missing intelligibility*, or seems inappropriate for *other reasons*. We detailed each in the following:

Toxic Emotions We model *toxic emotions* based on the emotional fallacies identified by Walton (2010): ad populum, ad misericordiam, ad baculum, and ad hominem. We merged these four into a single sub-dimension called *emotional deception* based on the results of a pilot annotation study (Section 4). Additionally, we define a sub-dimension *excessive intensity* to address overly intense emotions. In particular, our analysis revealed the presence of a subset of propaganda errors, including loaded language, flag-waving, repetition, exaggeration, and minimization Da San Martino et al. (2020).

Missing Commitment This dimension resembles the *credibility* dimension of Wachsmuth et al. (2017b), but it differs in that we do not mandate arguments to come from or include a trusted source. Rather, the arguments should demonstrate the participant's general interest in participating in the debate. To formalize this concept, we drew on the five rules for "A Good Dialogue" (Walton, 1999) to create two sub-dimensions of commitment, *missing seriousness* and *missing openness*, by examining the extent to which they apply to the arguments identified in the overlap analysis.

Missing Intelligibility The core dimension *missing intelligibility* results from the overlap analysis of the *clarity* and *arrangement* dimensions of Wachsmuth et al. (2017b). We found that the main point of an argument was partly unclear either due to (un)intentional vagueness or overly (un)complex language, which we refer to in our taxonomy as the sub-dimension *unclear meaning*. Also, derailing a discussion to another issue is a common issue (represented by the sub-dimension *missing relevance*). Finally, in some cases the individual claims and premises were intelligible but not their connection. We refer to this as a *confusing reasoning*.

Other Reasons This dimension accounts for reasons that do not fit into the other core-dimensions. As part of this, we observed that some arguments have a *detrimental orthography*, limiting intelligibility in some cases (spelling or grammatical errors) or increasing emotions in others (capital letters, repeated exclamation points). We leave any other case of inappropriateness as *reason unclassified*.

Figure 3 depicts the final taxonomy of all 14 dimensions we propose. We hierarchically decompose *inappropriateness* into the four core dimensions and those further into the nine discussed sub-dimensions to obtain a nuanced understanding of inappropriateness. The argument-centric focus of our taxonomy allows annotators to quickly formulate reasons for inappropriateness in the form "a is inappropriate because of σ", where a is an argument and σ a specific sub-dimension from the taxonomy. We define each dimension below.

3.3 A Hierarchical Taxonomy

Since *appropriateness* itself is already discussed in the literature, we refrain from redefining it here. Instead, we build on Wachsmuth et al. (2017b) who state that an argument "has an appropriate style if the used language supports the creation of credibility and emotions as well as if it is proportional to

the issue." Their annotation guidelines further suggest that "the choice of words and the grammatical complexity should [...] appear suitable for the topic discussed within the given setting [...], matching the way credibility and emotions are created [...]".

While our goal is to model appropriate language in argumentation, we decided to define when an argument is *not* appropriate (as indicated above) to maintain freedom of speech as much as possible. Therefore, we define the four core dimensions and their sub-dimensions from Figure 3 in a "reverse" way, clarifying what is considered *in*appropriate:

Toxic Emotions (TE) An argument has toxic emotions if the emotions appealed to are deceptive or their intensities do not provide room for critical evaluation of the issue by the reader.

- *Excessive Intensity (EI).* The emotions appealed to by an argument are unnecessarily strong for the discussed issue.
- *Emotional Deception (ED).* The emotions appealed to are used as deceptive tricks to win, derail, or end the discussion.

Missing Commitment (MC) An argument is missing commitment if the issue is not taken seriously or openness other's arguments is absent.

- *Missing Seriousness (MS).* The argument is either trolling others by suggesting (explicitly or implicitly) that the issue is not worthy of being discussed or does not contribute meaningfully to the discussion.
- *Missing Openness (MO).* The argument displays an unwillingness to consider arguments with opposing viewpoints and does not assess the arguments on their merits but simply rejects them out of hand.

Missing Intelligibility (MI) An argument is not intelligible if its meaning is unclear or irrelevant to the issue or if its reasoning is not understandable.

- *Unclear Meaning (UM).* The argument's content is vague, ambiguous, or implicit, such that it remains unclear what is being said about the issue (it could also be an unrelated issue).
- *Missing Relevance (MR).* The argument does not discuss the issue, but derails the discussion implicitly towards a related issue or shifts completely towards a different issue.
- *Confusing Reasoning (CR).* The argument's components (claims and premises) seem not to be connected logically.

Other Reasons (OR) An argument is inappropriate if it contains severe orthographic errors or for reasons not covered by any other dimension.

- *Detrimental Orthography (DO).* The argument has serious spelling and/or grammatical errors, negatively affecting its readability.
- *Reason Unclassified (RU).* There are any other reasons than those above for why the argument should be considered inappropriate.

4 The Appropriateness Corpus

This section details the data acquisition and annotation process of our *Appropriateness Corpus* and provides statistics of the collected annotations. Statistics of our corpus split by argument source are found in Appendix F.

4.1 Data Acquisition

Studying the applicability of our taxonomy requires a set of arguments that is both diverse and sufficiently large. We rely on manually labeled examples of reasonable quality to ensure that our corpus only contains argumentative texts. In particular, we collected all 2191 arguments on 1154 unique issues from existing corpora (Habernal and Gurevych, 2016b; Wachsmuth et al., 2017b; Ng et al., 2020).[2] All corpora are used in research on argument quality assessment (Habernal and Gurevych, 2016a; Wachsmuth and Werner, 2020; Lauscher et al., 2020) and contain annotations that we identified as related to appropriateness:

- The Dagstuhl-15512 ArgQuality corpus (Wachsmuth et al., 2017b) covers appropriateness and its most correlated dimensions.
- The UKPConvArg2 (Habernal and Gurevych, 2016a) corpus has reason labels for why argument a is more convincing than argument b.
- The GAQCorpus (Ng et al., 2020) covers four argument quality dimensions, including effectiveness, the "parent" of appropriateness.

We carefully selected the source corpora such that about 50% of the arguments belong to only 16

[2]The arguments from Wachsmuth et al. (2017b) are a subset of those from Habernal and Gurevych (2016b) with additional annotations. We include each argument once only.

		(a) Count		(b) Agree.		(c) Kendall's τ Correlation													
Dimension		**Yes**	**No**	**Full**	α	**In**	**TE**	**EI**	**ED**	**MC**	**MS**	**MO**	**MI**	**UM**	**MR**	**CR**	**OR**	**DO**	**RU**
In	**Inappropriateness**	**1182**	1009	60%	.45		.56	.38	.44	.59	.35	.47	.62	.41	.42	.25	.21	.18	.10
TE	**Toxic Emotions**	594	1597	77%	.36	.56		**.66**	**.78**	.35	.11	.35	.13	.01	.12	.06	.00	.00	.00
EI	Excessive Intensity	402	1789	82%	.27	.38	.66		.22	.25	.06	.26	.11	.00	.09	.07	.00	.00	.01
ED	Emotional Deception	427	1764	82%	.36	.44	**.78**	.22		.28	.11	.25	.09	.01	.09	.03	.01	.00	.00
MC	**Missing Commitment**	735	1456	69%	.21	.59	.35	.25	.28		**.57**	**.81**	.22	.08	.20	.07	.02	.02	.01
MS	Missing Seriousness	183	2008	93%	.51	.35	.11	.06	.11	.57		.12	.15	.09	.17	.03	.02	.02	.00
MO	Missing Openness	658	1533	71%	.11	.47	.35	.26	.25	**.81**	.12		.18	.05	.16	.07	.01	.00	.01
MI	**Missing Intelligibility**	774	1417	68%	.25	**.62**	.13	.11	.09	.22	.15	.18		**.64**	**.66**	**.41**	.14	.15	.02
UM	Unclear Meaning	459	1732	80%	.16	.41	.01	.00	.01	.08	.09	.05	.64		.17	.21	.16	.20	0.1
MR	Missing Relevance	508	1683	78%	.19	.42	.12	.09	.09	.20	.17	.16	**.66**	.17		.07	.03	.02	.02
CR	Confusing Reasoning	174	2017	92%	.16	.25	.06	.07	.03	.07	.03	.07	.41	.21	.07		.14	.15	.01
OR	**Other Reasons**	108	2083	95%	.23	.21	.00	.00	.01	.02	.02	.01	.14	.16	.03	.14		**.88**	**.43**
DO	Detrimental Orthography	77	2114	97%	.31	.18	.00	.00	.00	.02	.02	.00	.15	.20	.02	.15	**.88**		.01
RU	Reason Unclassified	32	**2159**	**99%**	.00	.10	.00	.01	.00	.01	.00	.01	.02	.01	.02	.01	.43	.01	

Table 1: Corpus statistics of the 2191 annotated arguments: (a) Counts of annotations for each inappropriateness dimension, when being aggregated conservatively (i.e., at least one annotator chose *yes*). (b) *Full* agreement and Krippendorff's α agreement of all three annotators. (c) Kendall's τ correlation between the 14 inappropriateness dimensions, averaged over the correlations of all annotators. The highest value in each column is marked in bold.

issues while the rest covers the remaining 1138 issues, making our corpus valuable both vertically (issues with many arguments allow deeper analyses) and horizontally (large number of issues promotes generalizability). The average sentence length of arguments is 4.8. The corpus includes arguments of three genres, 1590 from debate portals, 500 from question answering forums, and 101 reviews.

4.2 Annotation Process

We designed a task-specific annotation interface that leverages the hierarchical structure of the taxonomy in Figure 3. Specifically, annotators needed to label sub-dimensions, only if the respective core dimension was labeled before as given for an argument. Following Wachsmuth et al. (2017b), we used an ordinal scale for the *inappropriateness* dimension described as (1) fully inappropriate, (2) partially (in)appropriate, and (3) fully appropriate.

Likewise, a binary yes/no scale was used for all the other dimensions, where *yes* means inappropriateness in terms of the respective dimension. Annotators were required to select a reason (core dimension) from the taxonomy only for partially or fully inappropriate arguments. We provided a coherent and self-descriptive interface (see Appendix D) to reduce the cognitive load on the annotators. The annotators also had the opportunity to provide their own reasons for the *reason unclassified* dimension.

We conducted two rounds of annotation to find qualified annotators. In the first round, eight native English speakers hired on *Upwork* and two authors of this paper (5 female, 5 male in total) each annotated 100 arguments, randomly sampled from our corpus. Based on the results and feedback on the annotation interface and the guidelines, we refined our taxonomy, most notably reducing the number of dimensions from 18 to 14. For the second round, we selected the three Upwork annotators with the highest expert correlations (2 female, 1 male). We paid $13 per hour for annotating all 2191 arguments, as we did in the first round. To mitigate the cognitive overload entailed by prolonged reading, we divided the annotation into 14 batches of roughly 150 arguments each and limited the number of batches to be annotated per day to one.

4.3 Corpus Statistics and Agreement

To combine the annotators' labels in our corpus, we first use MACE (Hovy et al., 2013) in order to consider the annotators' reliability. We then compute Krippendorff's α between the MACE labels and those obtained with either of three combination strategies: *Liberal* considers an argument appropriate if at least one annotator marked it as such. *Majority* considers the label for which at least two annotators agree. *Conservative*, finally, considers an argument inappropriate if at least one annotator marked it as such. Table 2 shows that the MACE labels correlate best with the conservative labels in all cases. Consequently, to obtain the final corpus annotations, we combined the three labels of each argument following the conservative strategy. This strategy also seems most consistent with the current belief system in many societies around the world, that is, to accommodate minorities in language.

	Krippendorff's α		
Dimension	**Liberal**	**Majority**	**Conservat.**
Inappropriateness	0.16	0.54	**0.95**
Toxic Emotions	0.14	0.45	**1.00**
Excessive Intensity	-0.08	0.30	**1.00**
Emotional Deception	0.05	0.41	**1.00**
Missing Commitment	-0.03	0.30	**1.00**
Missing Seriousness	0.27	0.54	**1.00**
Missing Openness	-0.12	0.22	**0.96**
Missing Intelligibility	-0.03	0.41	**1.00**
Unclear Meaning	-0.07	0.19	**1.00**
Missing Relevance	-0.04	0.22	**1.00**
Confusing Reasoning	-0.04	0.19	**0.95**
Other Reasons	0.08	0.31	**1.00**
Detrimental Orthography	0.13	0.42	**1.00**
Reason Unclassified	-0.01	-0.01	**1.00**

Table 2: Krippendorff's α agreement between MACE labels and the manual labels obtained by each evaluated combination strategy (liberal, majority, conservative).

Table 1(a) presents the corpus distribution of the annotations aggregated conservatively. For readability, we binarized the overall inappropriateness in the table, considering both fully and partially inappropriate arguments as inappropriate. 1182 arguments were considered at least partially (in)appropriate (540 of them fully inappropriate).

Among the reasons given, *missing intelligibility* is the most frequent core dimension (774 arguments) and *missing openness* the most frequent sub-dimension (658), matching the intuition that a missing openness to others' opinions is a key problem in online discussions. The least frequent core dimension is *other reasons* (108), and the least frequent sub-dimension *reason unclassified* (32). That is, our annotators rarely saw additional reasons, indicating the completeness of our taxonomy.

Table 1(b) shows inter-annotator agreement. For *inappropriateness*, the annotators had full agreement in 60% of all cases, suggesting that stricter settings than our conservative strategy can also be applied without limiting the number of annotations too much. The Krippendorff's α agreement is limited but reasonable given the subjectiveness of the task. It ranges from .11 to .51 among the dimensions (not considering *reason unclassified*), with .45 for overall *inappropriateness*. These values are similar to those of Wachsmuth et al. (2017b).

5 Analysis

Building on existing corpora on theoretical and practical argument quality, we now report the correlations of our proposed dimensions and the quality dimensions of Wachsmuth et al. (2017b) and Habernal and Gurevych (2016a). Correlations with Ng et al. (2020) are found in Appendix E (only one dimension is directly related to appropriateness).

5.1 Relations between Corpus Dimensions

Table 1(c) presents the Kendall's τ correlations between all inappropriateness dimensions. Among the core dimensions, we find *missing intelligibility* to be most (.62) and *other reasons* to be least (.21) correlated with *inappropriateness (In)*. In case of the sub-dimensions, *missing openness* is most (.47) and *not classified* least (.10) correlated with it.

The sub-dimensions are mostly correlated with their direct parent, with values between .41 and .88, which is expected due to our annotation study setup. However, there are clear differences between sub-dimensions of the same parent; for example, *excessive intensity* and *emotional deception* are highly correlated with *toxic emotions* (.66 and .78) but have low correlation with each other (.22). Cross-dimensional correlations among the core- and sub-dimensions are highest between *toxic emotions* and *missing intelligibility* (.35) and *excessive intensity* and *missing openness* (.28) respecitvely. This suggests that overly intense emotions sometimes signify a rejection of others' opinions and vice versa.

5.2 Relation to Theory of Argument Quality

Table 3 shows the Kendall's τ correlations between our dimensions and the theoretical quality dimensions of Wachsmuth et al. (2017b). We observe the highest correlation for the two *(in)appropriateness* dimensions (.41), showing that our annotation guideline indeed captures the intended information for the annotated arguments. Furthermore, seven of our dimensions correlate most strongly with *appropriateness* in the Dagstuhl-15512 ArgQuality corpus, and all 14 dimensions have the highest correlation with one of the seven argument quality dimensions that we used to derive the taxonomy.

The values of *reason unclassified (RU)* are low (between .02 and .14), speaking for the completeness of our taxonomy. However, its most correlated quality dimension is *cogency*, possibly indicating a minor logical component of appropriateness.

5.3 Relation to Practice of Argument Quality

Table 4 shows the correlations between our dimensions and the convincingness comparison reasons of Habernal and Gurevych (2016a). We see that

Quality Dimension	Description	Kendall's τ Correlation with Inappropriateness Dimensions														
		In	**TE**	**EI**	**ED**	**MC**	**MS**	**MO**	**MI**	**UM**	**MR**	**CR**	**OR**	**DO**	**RU**
Cogency	P acceptable / relevant / sufficient	.27	.18	.19	.13	.22	.22	.15	.27	.20	.21	.20	.12	.11	**.14**
Local acceptability	P rationally believable	.36	**.29**	**.27**	**.26**	.28	.23	.21	.32	.20	.26	.23	.13	.12	.07
Local relevance	P contribute to acceptance / rejection	.30	.16	.16	.14	.22	.26	.13	.31	.21	.25	.21	.13	.14	.06
Local sufficiency	P give enough support	.25	.15	.15	.10	.20	.19	.14	.27	.20	.22	.17	.10	.07	.07
Effectiveness	a persuades target audience	.27	.18	.18	.13	.23	.21	.17	.26	.19	.20	.21	.12	.11	.04
Credibility	a makes author worthy of credence	.32	.22	.16	.19	.29	.24	.22	.29	.22	.21	**.25**	.15	.12	.07
Emotional appeal	a makes target audience more open	.17	.13	.10	.13	.16	.15	.12	.14	.15	.14	.16	.10	.04	.08
Clarity	a uses correct/unambiguous language	.20	.09	.08	.07	.10	.19	.04	.25	.18	.22	.21	.14	**.21**	.08
Appropriateness	a's credibility/emotions are proportional	**.41**	.25	.24	.21	**.35**	**.28**	**.27**	**.36**	**.30**	.25	.21	**.20**	.20	.08
Arrangement	a has components in the right order	.26	.13	.15	.10	.16	.19	.10	.31	.25	.21	.24	.13	.15	.02
Reasonableness	A acceptable / relevant / sufficient	.34	.23	.23	.16	.27	.23	.20	.32	.24	.26	.20	.16	.14	.11
Global acceptability	A worthy to be considered	.38	.28	**.27**	.22	.31	.25	.24	.35	.24	**.27**	**.25**	.16	.16	.07
Global relevance	A contribute to issue resolution	.24	.16	.21	.11	.17	.23	.10	.26	.19	.21	.18	.12	.11	.08
Global sufficiency	A adequately rebuts counterarguments	.20	.15	.19	.09	.21	.16	.16	.19	.17	.12	.15	.06	.04	.08
Overall quality	a/A is of high quality	.30	.19	.20	.15	.24	.22	.17	.30	.24	.21	.21	.14	.12	.10

Table 3: Kendall's τ correlation of the mean ratings of the argument *quality dimensions* of Wachsmuth et al. (2017b) with the 14 proposed *inappropriateness dimensions* (see Table 1 for the meaning of the acronyms). P are the premises of an argument a that is used within argumentation A. The highest value in each column is marked in bold.

Comparison	Reason	Kendall's τ Correlation with Inappropriateness Dimensions														
		In	**TE**	**EI**	**ED**	**MC**	**MS**	**MO**	**MI**	**UM**	**MR**	**CR**	**OR**	**DO**	**RU**
b is less convincing than a, since...	b is attacking / abusive	**.86**	**.70**	**.54**	**.70**	**.70**	.65	**.44**	.49	.34	.50	.01	.05	.02	.13
	b has language issues / humour / sarcasm	.77	.35	.21	.35	.61	**.69**	.14	.44	.41	.35	.16	.34	.33	.15
	b is unclear / hard to follow	.61	.16	.03	.17	.30	.36	.13	.52	.54	.33	**.36**	**.48**	**.47**	.11
	b has no credible evidence / no facts	.56	.20	.20	.12	.37	.46	.21	.47	.41	.29	.24	.15	.11	.12
	b has less or insufficient reasoning	.82	.32	.29	.24	.59	.64	.34	.63	**.57**	.46	.25	.08	.04	.16
	b uses irrelevant reasons	.69	.29	.21	.25	.45	.49	.25	.60	.44	.54	.20	.20	.09	.02
	b is only an opinion / a rant	.72	.21	.18	.19	.45	.55	.22	.57	.42	.44	.16	.19	.09	.14
	b is non-sense / confusing	.69	.22	.09	.25	.48	.57	.14	.58	.49	.43	.32	.40	.37	.06
	b does not address the topic	.82	.04	.03	.07	.30	.57	.09	**.75**	.48	**.72**	.05	.19	.00	.10
	b is generally weak / vague	.59	.18	.19	.11	.35	.42	.18	.53	.48	.33	.28	.08	.07	.15
a is more convincing than b, since...	a is more detailed / better reasoned / deeper	.50	.14	.11	.12	.32	.42	.16	.45	.40	.30	.21	.19	.15	.11
	a is objective / discusses other views	.44	.15	.09	.14	.30	.39	.18	.37	.36	.25	.17	.21	.15	.18
	a is more credible / confident	.40	.13	.05	.18	.20	.34	.10	.42	.38	.30	.18	.17	.12	.20
	a is clear / crisp / well-written	.55	.30	.22	.27	.35	.37	.24	.47	.41	.31	.32	.27	.25	.20
	a sticks to the topic	.65	.23	.17	.22	.34	.49	.07	.55	.39	.47	.15	.19	.13	.18
	a makes you think	.38	.13	.15	.08	.22	.34	.07	.27	.19	.20	.21	.11	.14	**.26**
	a is well thought through / smart	.56	.26	.14	.27	.36	.40	.23	.46	.34	.27	.30	.28	.24	.05
Overall	a is more convincing than b	.53	.17	.13	.15	.32	.43	.14	.44	.37	.32	.19	.19	.14	.13

Table 4: Kendall's τ correlation of the convincingness *comparison reasons* of argument pairs (a, b) of Habernal and Gurevych (2016a) with differences in the mean ratings of dimensions of the proposed *inappropriateness dimensions* (see Table 1 for the meaning of the acronyms). The highest value in each column is marked in bold.

attacking/abusive behavior is most correlated with our *inappropriateness* (In, .86), *missing commitment* (MC, .70) and *toxic emotions* (TE, .70) dimensions. *Missing seriousness (MS)* and *missing intelligibility (MI)* are mostly correlated with humor/sarcasm (.69) and not addressing (derailing) the topic (.75) respecitvely. *Confusing reasoning (CR)* is most correlated with an argument being hard to follow (.36), and *unclear meaning (UM)* with insufficient reasoning (.57).

We find that *detrimental orthography (DO)* renders an argument unclear and difficult to follow (.47). Finally, the *reason unclassified (RU)* dimension is most correlated with making a reader think about an argument. Manual inspection of the reasons for these annotations reveals that annotators chose *reason unclassified*, if they were unsure which of the other dimensions they should assign.

6 Experiments

The corpus from Section 4 is meant to enable the computational treatment of inappropriate language in argumentation. As an initial endeavor, this section reports baselines for classifying all 14 dimensions in the taxonomy from Section 3.

Approach	In	TE	EI	ED	MC	MS	MO	MI	UM	MR	CR	OR	DO	RU	Macro
Random baseline	.49	.47	.45	.45	.49	.39	.47	.48	.45	.47	.39	.37	.37	.34	.43
Majority baseline	.32	.42	.45	.45	.40	.48	.41	.39	.44	.43	.48	.49	.49	.50	.44
DeBERTaV3-large	**.75**	**.74**	**.69**	**.70**	**.75**	**.73**	**.72**	**.72**	**.69**	.68	**.62**	**.65**	**.67**	**.52**	**.69**†‡
DeBERTaV3-w/o-issue	**.75**	.73	.68	**.70**	**.75**	**.73**	.71	**.72**	.68	**.69**	.61	.63	.66	.51	.68‡
DeBERTaV3-shuffle	.72	.69	.64	.64	.71	.65	.68	.70	.66	.65	.57	.59	.57	.50	.64
Human performance	.78	.79	.73	.77	.73	.82	.70	.76	.73	.72	.74	.78	.80	.70	.75

Table 5: Evaluation of appropriateness classification: F_1-score of each approach in 5-times repeated 5-fold cross validation on all 14 proposed dimensions. The best value in each column is marked bold. We marked significant macro F_1-score gains over *DeBERTaV3-w/o-issue* (†) and *DeBERTaV3-shuffle* (‡) at $p < .05$.

6.1 Experimental Setup

In line with Table 1, we treat all annotations as binary labels. We performed five repetitions of 5-fold cross-validation (25 folds in total) and ensured a similar distribution of the labels in each fold. For each folding, we used 70% for training, 10% for selecting the best-performing approach in terms of the mean macro-F_1 score, and 20% for testing.

Models For classification, we employed the recent model *DeBERTaV3-large* (He et al., 2021), with an argument prepended by the discussion issue as input. Besides, we tested two "ablations": *DeBERTaV3-w/o-issue* receives only the argument to gain insight into how effective it is to provide the issue as context. *DeBERTaV3-shuffle* receives the argument and the issue with all words shuffled, to analyze the impact of proper syntactic and semantic formulations. We trained our models to predict all 14 dimensions via a multi-label prediction loss, accounting for data imbalance by assigning weights to all dimensions (more details in Appendix A).

Lower and Upper Bounds To quantify the impact of learning, we compare against a *random baseline* that chooses a label pseudo-randomly and a *majority baseline* that takes the majority label for each dimension. As an upper bound, we measure *human performance* in terms of the average of each human annotator in isolation on the dataset.

6.2 Results

Table 5 presents the mean F_1-score for all 14 inappropriateness dimensions averaged over all folds. *DeBERTaV3-large* performs best in terms of macro F_1-score (.69), significantly beating both *DeBERTaV3-w/o-issue* (.68) and *DeBERTaV3-shuffle* (.65) in a Wilcoxon signed-rank test ($p < .05$) . The gain over *DeBERTaV3-w/o-issue* is small though, suggesting that the context of a discussion (here, the issue) may be of limited importance for predicting inappropriateness. Plausible reasons are that (1) most arguments are (in)appropriate regardless of their context, or (2) the context of the argument is explicitly or implicitly contained within most arguments. *DeBERTaV3-w/-issue* clearly outperforms the random baseline and majority baseline on all dimensions, and it achieves about 92% of human performance in terms of macro F_1 (.75). These results suggest the possibility of automating the task of predicting appropriateness, however, encouraging further improvements.

7 Conclusion

Online discussions of controversial topics mostly turn out fruitful only, when the participants argue *appropriately*, a dimension of argumentative language that has received no systematic investigation so far. Therefore, we have presented a taxonomy of 14 dimensions to model inappropriate language in argumentation, derived from rhetoric and argumentation theory. To enable computational research on appropriateness, we compiled a corpus of 2191 arguments from three genres, carefully annotated for all dimensions.

Our extensive corpus analyses confirm correlations with both theoretical and practical dimensions of argument quality from the literature. The taxonomy covers inappropriateness comprehensively according to human annotators. While a DeBERTa-based baseline already comes rather close to human performance in classifying inappropriate language, our corpus allows for developing more sophisticated models in future work that may serve an automatic (or semi-automatic) content moderation.

To make content moderation successful and accepted, we think that providing clear reasons supporting the moderation is important, so the participants can better frame their arguments in online discussions. The defined taxonomy dimensions lay out how such reasons may look like.

8 Acknowledgments

This project has been partially funded by the German Research Foundation (DFG) within the project OASiS, project number 455913891, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999). We would like to thank the participants of our study and the anonymous reviewers for the feedback and their time.

9 Limitations

Aside from the still-improvable performance of the classification models we evaluated, our work is limited in two ways: the nature of what is considered appropriate as well as the difficulties that arise during corpus creation in NLP in general.

We point to the subjectivity in perception regarding appropriateness, which is also displayed and discussed in the paper by the inter-annotator agreement. Many sociocultural factors can influence this perception within cultures, such as age, gender, education, or ethnicity. We sought to account at least for gender by including both male and female annotators for all arguments. However, we encourage further studies that focus on other factors, as we expect appropriateness to be seen differently, primarily across cultures with varying styles of debates. Since our corpus contains only arguments written in English and is annotated by native English speakers, it may also be insufficient to generalize across languages.

Moreover, appropriateness perception is likely subject to change over time. Although we collected arguments from different years, we see long-time limitations to our corpus. In general, it also depends on the expectations of the discussion participants, which are to some extent predetermined by the context (e.g., a sales pitch vs. a discussion with friends). In that regard, the context of our corpus is solely that of discussing controversial issues with strangers on the web. Finally, the size of the created corpus we propose in the paper may limit the generalizability of approaches that build on it and should be investigated further in future work.

10 Ethical Considerations

The corpus and the computational baselines presented in this paper target a sensitive issue: what is considered appropriate to say in a discussion. We suggest differentiating between freedom of speech, hate speech, and inappropriate speech. We believe inappropriate speech is an extension of hate speech that leads to a less free but more healthy climate in speech exchange. While freedom of speech in many countries is limited by hate speech in law, the extension to inappropriate speech is not. Consequently, automating the detection of inappropriateness and dealing with it in the same way hate speech is addressed (often by removal) may be perceived as hurting individuals' freedom of speech and, thus, must be handled with great care.

However, we see no strong immediate ethical concerns regarding the computational methods specific to our work, as they only detect inappropriateness and do not recommend any actions. We stress, though, that they are not meant yet for real-life applications. Apart from the outlined limitations, we also do not see notable ethical concerns regarding our taxonomy, as we derived it systematically from existing literature and always encouraged our annotators to add their own reasons.

Finally, we aimed to ensure fair payment. As discussed in the paper, our annotators were paid about $13 per hour, which exceeds the minimum wage in most US states and is also conform to the standards in the regions of our host institutions.

References

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8172–8187, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Donna T Andrew. 1996. Popular culture and public debate: London 1780. *The Historical Journal*, 39(2):405–423.

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, Translator). Clarendon Aristotle series. Oxford University Press.

J Anthony Blair. 1988. What is bias? In Trudy Govier, editor, *Selected issues in logic and communication*, pages 93–104. Wadsworth Publishing Company.

John Walt Burkett. 2011. *Aristotle,"Rhetoric" III: A commentary*. Texas Christian University.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016b. 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.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Dell Hymes et al. 1972. On communicative competence. *sociolinguistics*, 269293:269–293.

Loae Fakhri Jdetawy and Modh Hilmi Hamzah. 2020. Linguistic etiquette: a review from a pragmatic perspective. *Technium Soc. Sci. J.*, 14:695.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Susan Ranney. 1992. Learning a new script: An exploration of sociolinguistic competence. *Applied Linguistics*, 13(1):25–50.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Klaus P Schneider. 2012. Appropriate behaviour across varieties of english. *Journal of Pragmatics*, 44(9):1022–1037.

Frans H. van Eemeren. 2015. *Reasonableness and Effectiveness in Argumentative Discourse: Fifty Contributions to the Development of Pragma-Dialectics*. Argumentation Library. Springer International Publishing.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. 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 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.

Douglas Walton. 1999. *One-sided arguments: A dialectical analysis of bias*. SUNY Press.

Douglas Walton. 2010. *The place of emotion in argument*. Penn State Press.

A Training Hyperparameters

We did a single initial round of hyperparameter optimization and sticked to the best values for all of our DeBERTaV3 experiments: a polynomial learning rate with a warmup ratio of .10, a batch size of 10, and an initial learning rate of $3 \cdot 10^{-6}$, trained for 10 epochs in all cases.

B Computational Infrastructure

Our experiments were done on Ubuntu 20.04 with Python version 3.7.12, CUDA version 11.3 and one A100-SXM4-40GB GPU. We used the following main libraries in our experiments (we include a full list of packages and their versions in the requirements.txt in the supplementary material):

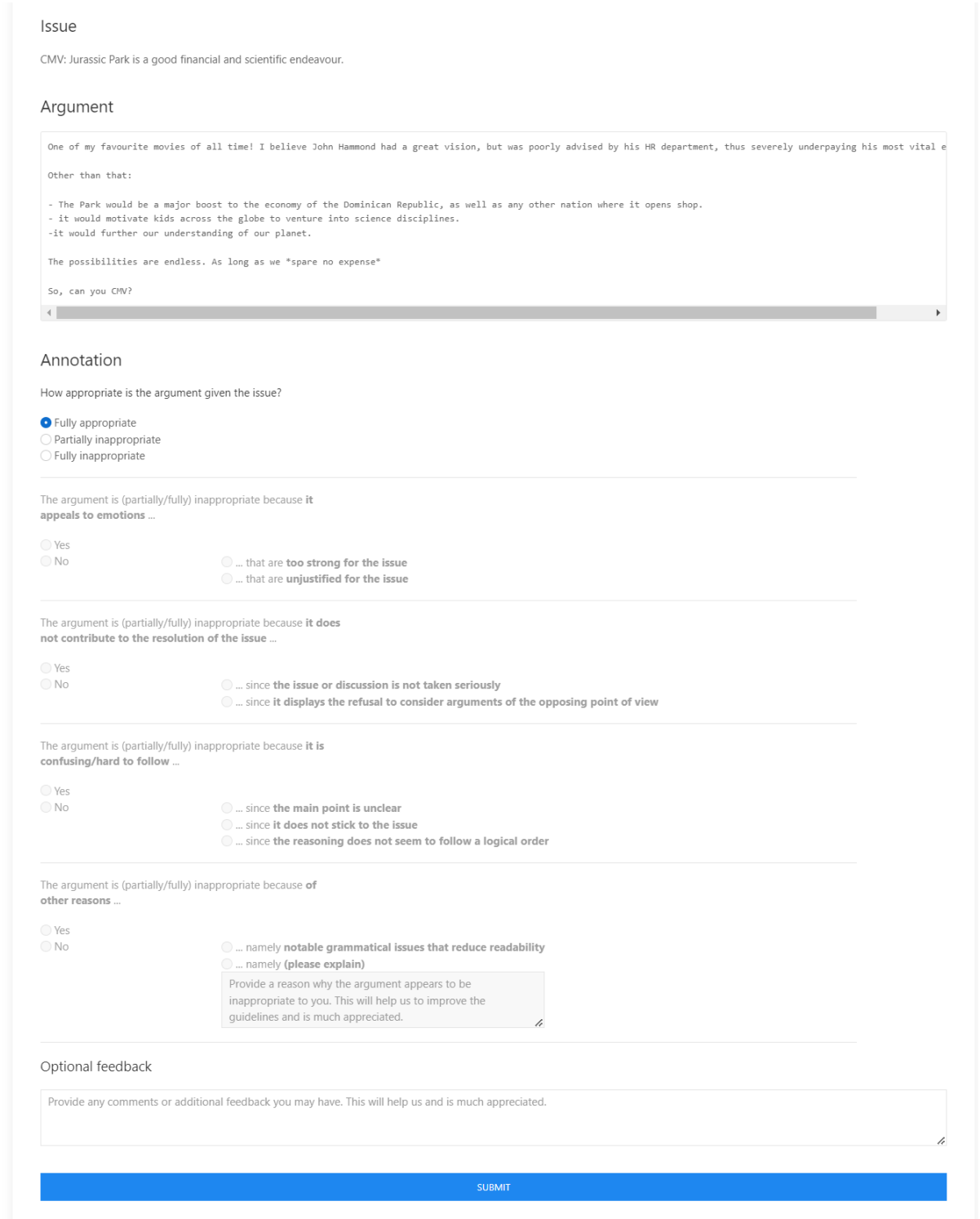
- torch==1.10.2+cu113
- transformers==4.21.0.dev0
- spacy==3.3.1

C Examples

Type	Issue	Argument
Toxic Intensity	If given the choice what way would you vote for the death penalty to be brought back in Britain?	Yes i am completely for it. People are arguing that it is barbaric and inhumane but who can stand up and say that some perv who has raped and killed a child still has human rights and the right to live. We would put down a dangerous dog why not do the same to some of the scum that lives in our country. The justice system in britain at the moment is hopeless. Far to many people are gettin away with all sorts and something needs to be done!!
Emotional Deception	Can the police keep your car permantly if you have 3rd suspended license?	Towed three times and impounded for 30 days each time? Man, you're just not getting the message, are you?If you are in California, you bet the police can forfeit your vehicle and it doesn't take three times to make it a charm. Technically, your vehicle could be subject to forfeiture proceedings after your first suspended license beef. Someone like you is exactly the reason the legislature designed that law, because your privilege to drive has been taken away from you and yet you obviously continue to drive. People like you are involved in an exponentially higher than average number of traffic accidents so the legislature figured maybe people like you should have your vehicles forfeited to the state if you just didn't go along with the game plan.Voila - I give you California Vehicle Code section 14607.6...and a link to it below. It would also be worth your time to review 14607.4, whether or not you live in California.You really need to stop driving. Really.
Missing Seriousness	is-porn-wrong	Porn is Wrong. mainly because they are Not Doing it Right. it should be Hi Def. in three years, it will be in 3-D.
Missing Openness	Pro-choice-vs-pro-life	There should be no argument in this really...whatever way yu see a fetus...its still a living form that has been created in a very intimate way... you shouldn't be changing what mothernature or God or fate or whatever has decided for you...and if you didn;t wannna get preggo in the first place...don't have sex or use protection. Yeh there are some women that get raped and it's very unfortunate but they should give the child up for adoption. It's not the child's fault that it was created. So why should the goring being have to pay the ultimate price of it's life?
Unclear Meaning	Evolution-vs-creation	Believing "Evolution" as in Darwinism and the like, is like believing the puzzle can be solved by pouring the pieces out because two pieces kind of stuck together.
Missing Relevance	Is it illegal to record a phone conversation?	The conversation can not be used as evidence in a court of law. I don't know what the lady hoped to gain from recording the conversation other than to create more drama. Some people are hooked on drama and they actually do what they can to create it. Run as far away and as fast as you can from these types. They will suck you dry.
Confusing Reasoning	If your spouse committed murder and he or she confided in you would you turn them in?	i would turn in my wife because its wrong to kill someone. it could have been an accident but it was still wrong and besides the police are going to find out who killed that person but i don't want her to leave me for a long period of time so i would tell but then again i wouldn't.
Deceptive Orthography	Is-the-school-uniform-a-good-or-bad-idea	it dose not show kids expressions and unforms dose not show is it
Reason Unclassified	Firefox-vs-internet-explorer	Firebug, WebDeveloper, TabMix, FaviconizeTab, GreaseMonkey, IETab (to use when you visit microsot.com). Just some reason why i prefer Firefox

Table 6: Examples of inappropriate arguments from our corpus for each of the nine sub-dimensions of our taxonomy.

D Annotation Interface



Issue

CMV: Jurassic Park is a good financial and scientific endeavour.

Argument

```
One of my favourite movies of all time! I believe John Hammond had a great vision, but was poorly advised by his HR department, thus severely underpaying his most vital e

Other than that:

- The Park would be a major boost to the economy of the Dominican Republic, as well as any other nation where it opens shop.
- it would motivate kids across the globe to venture into science disciplines.
-it would further our understanding of our planet.

The possibilities are endless. As long as we *spare no expense*

So, can you CMV?
```

Annotation

How appropriate is the argument given the issue?

- ○ Fully appropriate
- ◉ Partially inappropriate
- ○ Fully inappropriate

The argument is (partially/fully) inappropriate because **it appeals to emotions** ...

- ◉ Yes
- ○ No
  - ○ ... that are **too strong for the issue**
  - ◉ ... that are **unjustified for the issue**

The argument is (partially/fully) inappropriate because **it does not contribute to the resolution of the issue** ...

- ◉ Yes
- ○ No
  - ◉ ... since **the issue or discussion is not taken seriously**
  - ◉ ... since **it displays the refusal to consider arguments of the opposing point of view**

The argument is (partially/fully) inappropriate because **it is confusing/hard to follow** ...

- ○ Yes
- ◉ No
  - ○ ... since **the main point is unclear**
  - ○ ... since **it does not stick to the issue**
  - ○ ... since **the reasoning does not seem to follow a logical order**

The argument is (partially/fully) inappropriate because **of other reasons** ...

- ◉ Yes
- ○ No
  - ○ ... namely **notable grammatical issues that reduce readability**
  - ◉ ... namely **(please explain)**

The argument is inappropriate because...

Optional feedback

Provide any comments or additional feedback you may have. This will help us and is much appreciated.

SUBMIT

E GAQCorpus Correlations

Quality Dimension	Description	Kendall's τ Correlation with Inappropriateness Dimensions													
		In	**TE**	**EI**	**ED**	**MC**	**MS**	**MO**	**MI**	**UM**	**MR**	**CR**	**OR**	**DO**	**RU**
Cogency	P acceptable / relevant / sufficient	**.32**	.23	.22	**.22**	.24	.14	.22	**.24**	**.18**	**.17**	**.11**	.09	.09	**.02**
Effectiveness	a persuades target audience	**.32**	.23	.22	.21	.25	**.15**	.22	.21	.16	.15	**.11**	**.10**	**.11**	.00
Reasonableness	A acceptable / relevant / sufficient	**.32**	**.25**	**.24**	**.22**	**.26**	**.15**	**.24**	.21	.16	.15	.10	.09	.10	.01
Overall quality	a/A is of high quality	**.32**	.24	.23	**.22**	.24	**.15**	.23	.22	.17	.16	**.11**	**.10**	.10	.01

Table 7: Kendall's τ correlation of the argument *quality dimensions* of Ng et al. (2020) with the mean ratings of the proposed *appropriateness dimensions* (see Table 1 for the meaning of the acronyms). P are all premises of an argument a that is used within argumentation A. The highest value in each column is marked in bold.

F Corpus Statistics

Dimension		(a) Count		(b) Agree.		(c) Kendall's τ Correlation												
		Yes	**No**	**Full**	α	**In**	**TE**	**EI**	**ED**	**MC**	**MS**	**MO**	**MI**	**UM**	**MR**	**CR**	**OR**	**DO**	**RU**
In	**Inappropriateness**	**609**	443	52%	**.55**		.48	.32	.38	.59	.40	.45	**.65**	.42	.43	.25	.22	.18	.10
TE	**Toxic Emotions**	263	789	80%	.41	.48		**.65**	**.79**	.36	.14	.36	.15	.01	.14	.06	.01	.00	.00
EI	Excessive Intensity	172	880	84%	.35	.32	.65		.22	.23	.07	.24	.12	.01	.10	.07	.02	.02	.01
ED	Emotional Deception	186	866	84%	.43	.38	**.79**	.22		.31	.15	.27	.11	.03	.10	.04	.01	.01	.00
MC	**Missing Commitment**	381	671	69%	.31	.59	.36	.23	.31		**.62**	**.78**	.22	.08	.21	.04	.01	.01	.00
MS	Missing Seriousness	135	917	90%	**.55**	.40	.14	.07	.15	.62		.12	.15	.10	.17	.01	.01	.02	.01
MO	Missing Openness	326	726	70%	.17	.45	.36	.24	.27	**.78**	.12		.17	.05	.16	.06	.00	.01	.01
MI	**Missing Intelligibility**	460	592	62%	.30	**.65**	.15	.12	.11	.22	.15	.17		**.64**	**.65**	**.41**	.12	.14	.01
UM	Unclear Meaning	300	752	73%	.18	.42	.01	.01	.03	.08	.10	.05	.64		.17	.21	.12	.16	.03
MR	Missing Relevance	297	755	74%	.25	.43	.14	.10	.10	.21	.17	.16	**.65**	.17		.07	.02	.01	.01
CR	Confusing Reasoning	135	917	87%	.17	.25	.06	.07	.04	.04	.01	.06	.41	.21	.07		.12	.13	.01
OR	**Other Reasons**	86	966	92%	.24	.22	.01	.02	.01	.01	.01	.00	.12	.12	.02	.12		**.87**	**.45**
DO	Detrimental Orthography	59	993	95%	.33	.18	.00	.02	.01	.01	.02	.01	.14	.16	.01	.13	**.87**		.00
RU	Reason Unclassified	28	**1024**	**97%**	.01	.10	.00	.01	.00	.00	.01	.01	.01	.03	.01	.01	.45	.00	

Table 8: Corpus statistics of the 1052 annotated **arguments** in the UKPConvArg2 (Habernal and Gurevych, 2016a) corpus: (a) Counts of annotations for each inappropriateness dimension, when being aggregated conservatively (i.e., at least one annotator chose *yes*). (b) *Full* agreement and Krippendorff's α agreement of all three annotators. (c) Kendall's τ correlation between the 14 inappropriateness dimensions, averaged over the correlations of all annotators. The highest value in each column is marked in bold.

Dimension		(a) Count		(b) Agree.		(c) Kendall's τ Correlation													
		Yes	No	Full	α	In	TE	EI	ED	MC	MS	MO	MI	UM	MR	CR	OR	DO	RU
In	**Inappropriateness**	**271**	267	53%	.22		.67	.49	.49	.59	.19	.56	.50	.34	.27	.22	.18	.17	.00
TE	**Toxic Emotions**	145	393	76%	**.28**	**.67**		**.69**	**.74**	.33	.04	.32	.04	.00	.06	.01	.01	.01	.00
EI	Excessive Intensity	109	429	80%	.15	.49	.69		.22	.32	.03	.31	.03	.00	.05	.02	.03	.03	.00
ED	Emotional Deception	98	440	83%	**.28**	.49	**.74**	.22		.19	.01	.20	.03	.01	.05	.03	.01	.01	.00
MC	**Missing Commitment**	174	364	68%	.07	.58	.33	.32	.19		**.35**	**.94**	.10	.02	.17	.01	.01	.01	.00
MS	Missing Seriousness	14	524	97%	.18	.19	.04	.03	.01	.35		.11	.04	.02	.06	.01	.01	.01	.00
MO	Missing Openness	166	372	70%	.07	.56	.32	.31	.20	**.94**	.11		.09	.01	.16	.02	.01	.01	.00
MI	**Missing Intelligibility**	113	425	79%	.10	.50	.04	.03	.03	.10	.04	.09		**.65**	**.57**	**.41**	.20	.22	.00
UM	Unclear Meaning	66	472	88%	.03	.34	.00	.00	.01	.02	.02	.01	**.65**		.06	.17	.32	.36	.00
MR	Missing Relevance	51	487	91%	.09	.27	.06	.05	.05	.17	.06	.16	.57	.06		.01	.01	.01	.00
CR	Confusing Reasoning	18	520	97%	.09	.22	.01	.02	.03	.01	.01	.02	.41	.17	.01		.17	.19	.00
OR	**Other Reasons**	11	527	98%	.23	.18	.01	.03	.01	.01	.01	.01	.20	.32	.01	.17		**.96**	.00
DO	Detrimental Orthography	10	528	98%	.24	.17	.01	.03	.01	.01	.01	.01	.22	.36	.01	.19	**.96**		.00
RU	Reason Unclassified	1	**537**	**100%**	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	

Table 9: Corpus statistics of the 538 annotated **arguments** in the GAQCorpus (Ng et al., 2020): (a) Counts of annotations for each inappropriateness dimension, when being aggregated conservatively (i.e., at least one annotator chose *yes*). (b) *Full* agreement and Krippendorff's α agreement of all three annotators. (c) Kendall's τ correlation between the 14 inappropriateness dimensions, averaged over the correlations of all annotators. The highest value in each column is marked in bold.

Dimension		(a) Count		(b) Agree.		(c) Kendall's τ Correlation													
		Yes	No	Full	α	In	TE	EI	ED	MC	MS	MO	MI	UM	MR	CR	OR	DO	RU
In	**Inappropriateness**	**279**	221	47%	.33		.68	.44	.54	.55	.30	.44	.59	.36	.42	.20	.15	.00	.00
TE	**Toxic Emotions**	171	329	71%	.34	**.68**		**.63**	**.78**	.37	.11	.35	.18	.04	.12	.11	.00	.00	.00
EI	Excessive Intensity	111	389	78%	.25	.44	.63		.18	.26	.07	.25	.16	.04	.09	.16	.03	.00	.00
ED	Emotional Deception	133	367	75%	.30	.54	**.78**	.18		.28	.09	.25	.10	.01	.08	.04	.01	.00	.00
MC	**Missing Commitment**	177	323	66%	.09	.55	.37	.26	.28		**.54**	**.79**	.25	.04	.20	.12	.06	.00	.00
MS	Missing Seriousness	32	468	94%	**.40**	.30	.11	.07	.09	.54		.07	.11	.01	.13	.05	.03	.00	.00
MO	Missing Openness	165	335	67%	.00	.44	.35	.25	.25	**.79**	.07		.25	.06	.19	.11	.03	.00	.00
MI	**Missing Intelligibility**	189	311	64%	.11	.59	.18	.16	.10	.25	.11	.25		**.60**	**.72**	**.33**	.11	.00	.00
UM	Unclear Meaning	90	410	82%	.07	.36	.04	.04	.01	.04	.01	.06	.60		.14	.16	**.17**	.00	.00
MR	Missing Relevance	150	350	71%	.04	.42	.12	.09	.08	.20	.13	.19	**.72**	.14		.04	.00	.00	.00
CR	Confusing Reasoning	20	480	96%	.04	.20	.11	.16	.04	.12	.05	.11	.33	.16	.04		.14	.00	.00
OR	**Other Reasons**	11	489	98%	.08	.15	.00	.03	.01	.06	.03	.03	.11	.17	.00	.14		.00	.00
DO	Detrimental Orthography	8	492	98%	.11	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00		.00
RU	Reason Unclassified	3	**497**	**99%**	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	

Table 10: Corpus statistics of the 500 annotated **forum posts** in the GAQCorpus (Ng et al., 2020): (a) Counts of annotations for each inappropriateness dimension, when being aggregated conservatively (i.e., at least one annotator chose *yes*). (b) *Full* agreement and Krippendorff's α agreement of all three annotators. (c) Kendall's τ correlation between the 14 inappropriateness dimensions, averaged over the correlations of all annotators. The highest value in each column is marked in bold.

Dimension		(a) Count		(b) Agree.		(c) Kendall's τ Correlation														
		Yes	**No**	**Full**	α	**In**	**TE**	**EI**	**ED**	**MC**	**MS**	**MO**	**MI**	**UM**	**MR**	**CR**	**OR**	**DO**	**RU**
In	**Inappropriateness**	**23**	78	79%	.44		.78	.56	.63	.00	.00	.00	.52	.00	.48	.00	-	-	-
TE	**Toxic Emotions**	15	86	87%	.41	**.78**		**.73**	**.83**	.00	.00	.00	.16	.00	.17	.00	-	-	-
EI	Excessive Intensity	10	91	90%	.31	.56	.73		.43	.00	.00	.00	.14	.00	.15	.00	-	-	-
ED	Emotional Deception	10	91	92%	**.50**	.63	**.83**	.43		.00	.00	.00	.05	.00	.06	.00	-	-	-
MC	**Missing Commitment**	3	98	97%	.01	.00	.00	.00	.00		.00	.00	.00	.00	.00	.00	-	-	-
MS	Missing Seriousness	2	99	98%	.00	.00	.00	.00	.00	.00		.00	.00	.00	.00	.00	-	-	-
MO	Missing Openness	1	**100**	**99%**	.00	.00	.00	.00	.00	.00	.00		.00	.00	.00	.00	-	-	-
MI	**Missing Intelligibility**	12	89	88%	.16	.52	.16	.14	.05	.00	.00	.00		.00	**.92**	.00	-	-	-
UM	Unclear Meaning	3	98	97%	.01	.00	.00	.00	.00	.00	.00	.00	.00		.00	.00	-	-	-
MR	Missing Relevance	10	91	90%	.20	.48	.17	.15	.06	.00	.00	.00	**.92**	.00		.00	-	-	-
CR	Confusing Reasoning	1	**100**	**99%**	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00		-	-	-
OR	**Other Reasons**	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DO	Detrimental Orthography	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
RU	Reason Unclassified	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 11: Corpus statistics of the 101 annotated **reviews** in the GAQCorpus (Ng et al., 2020): (a) Counts of annotations for each inappropriateness dimension, when being aggregated conservatively (i.e., at least one annotator chose *yes*). (b) *Full* agreement and Krippendorff's α agreement of all three annotators. (c) Kendall's τ correlation between the 14 inappropriateness dimensions, averaged over the correlations of all annotators. The highest value in each column is marked in bold.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☑ A2. Did you discuss any potential risks of your work?
9

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☑ A4. Have you used AI writing assistants when working on this paper?
Assistance purely with the language of the paper.

B ☑ Did you use or create scientific artifacts?

4

☑ B1. Did you cite the creators of artifacts you used?
4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Upon acceptance, the collected dataset will be published under the Creative Commons Attribution 4.0 International

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
All data is completely anonymized, and annotators are only identified by a unique ID that is not connected to personal information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
5

C ☑ Did you run computational experiments?

6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix E

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix D

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Supplementary material

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix C

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
4

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
All annotators were informed during recruitment about the task and what the annotated data will be used for.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Since we collect no personal information, the university internal ethics board deemed the annotation study to be ethical.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
4