# Benchmark Evaluation for Tasks with Highly Subjective Crowdsourced Annotations: Case study in Argument Mining of Political Debates

**Rafael Mestre[†], Matt Ryan, Stuart E. Middleton, Richard Gomer, Masood Gheasi, Jiatong Zhu, Timothy J. Norman**

University of Southampton
[†]r.mestre@soton.ac.uk

**Abstract**

This paper assesses the feasibility of using crowdsourcing techniques for subjective tasks, like the identification of argumentative relations in political debates, and analyses their inter-annotator metrics, common sources of error and disagreements. We aim to address how best to evaluate subjective crowdsourced annotations, which often exhibit significant annotator disagreements and contribute to a "quality crisis" in crowdsourcing. To do this, we compare two datasets of crowd annotations for argumentation mining performed by an open crowd with quality control settings and a small group of master annotators without these settings but with several rounds of feedback. Our results show high levels of disagreement between annotators with a rather low Krippendorf's alpha, a commonly used inter-annotator metric. This metric also fluctuates greatly and is highly sensitive to the amount of overlap between annotators, whereas other common metrics like Cohen's and Fleiss' kappa are not suitable for this task due to their underlying assumptions. We evaluate the appropriateness of the Krippendorf's alpha metric for this type of annotation and find that it may not be suitable for cases with many annotators coding only small subsets of the data. This highlights the need for more robust evaluation metrics for subjective crowdsourcing tasks. Our datasets provide a benchmark for future research in this area and can be used to increase data quality, inform the design of further work, and mitigate common errors in subjective coding, particularly in argumentation mining.

## 1 Introduction

Crowdsourcing has been gaining popularity over the last years as a way of redistributing a large number of small tasks over the Internet to users ("the crowd") that can, either altruistically or by small payments, offer their computing resources, knowledge or time to help advance different fields of research. In social science, crowdsourcing is becoming an acceptable technique for recruiting participants in studies and information gathering. In computer science, machine learning-driven models, such as those based on natural language processing (NLP) and computer vision, often train using datasets that have crowdsourced annotations. In social science tasks, despite not completely solving the problem of obtaining a representative sample of the population of interest (Berinsky, Huber, and Lenz 2012), crowdsourcing has been demonstrated to provide high levels of agreement between in different tasks, for instance between crowd workers and political science experts at assessing the economic and social policy in political texts (Benoit et al. 2016). Other studies have explored whether using crowdsourcing can be a fix for the scientific "replication crisis" and have shown how studies in cognitive psychology and political science can be mostly reproduced with crowdworkers (Stewart, Chandler, and Paolacci 2017).

However, most of the studies focus on quantitative data. What happens when the task at hand is more subjective and qualitative? How do we deal with disagreement and evaluation metrics? Chen et al. (2018) analysed the unexplored connection between machine learning, "qualitative coding" and ambiguity when crowdsourcing social science data. The authors performed a short experiment with one of the most commonly used crowdsourcing platforms, Amazon's Mechanical Turk (MTurk), to study the relation between disagreement and ambiguity. They used master coders of MTurk, a distinction used for highly qualified contributors with high performance in previous tasks, and they found an expected correlation between perceived ambiguity in the task and the disagreement between annotations. Similarly, others have warned that the design of the task must be precisely well thought out when respondents' measures are subjective and qualitative (Kittur, Chi, and Suh 2008). Whereas some researchers might consider this subjectivity 'noise', there is a growing interest in modelling human label variation in a way that is embraced, rather than considered problematic (Plank 2022).

Disagreement in NLP tasks can come from individual differences, characteristics of the task and context (Basile et al. 2021). Traditionally, in NLP and related applications, a 'ground truth' is assumed, generally achieved by a majority vote, for models to be trained on. This assumption, however, can go against natural human subjectivity (Aroyo and Welty 2015), but can also reproduce biases and silence minority groups in tasks as sensitive as toxicity classification or detecting arguments in deliberations (Gordon et al. 2022). Certain paradigms have been proposed to expand the notion of annotation to allow for human label variation, such as label distribution learning (Geng 2016), in which each instance

of data is associated with a label distribution and evaluation metrics are based on the similarity or distance between predicted and real label distributions. Other approaches, like jury learning (Gordon et al. 2022), consider individual human characteristics that can help control for cultural or demographic differences without silencing minority groups.

The disagreement between annotators in these platforms is one of the reasons behind what has been called a "quality crisis" in crowdsourcing (Kennedy et al. 2020). Despite the popularity of MTurk in studies ranging from economics to sociology, with sometimes higher quality data than student or national samples, several researchers started to notice poor quality of response in their experiments. Researchers started to worry that respondents might not be engaging seriously with the task, answering nonsensically or randomly to open-ended or demographic questions. They also suspected that many annotators might be using bots to semi-automatically answer questions or virtual private servers to mask their location and answer questions in different languages with the help of automatic translation tools. This makes customary the use of quality control mechanisms (QCMs) to eliminate invalid annotations, but at the same time goes against the notion of embracing subjectivity in human labelled data and perpetuates the myths of human annotation (Aroyo and Welty 2015). Most of the research on the validity of crowdsourcing has been directed towards evaluating whether annotators provide 'valid' answers, but not much attention has been paid to investigating if these metrics are actually valid to assess crowd annotations.

In this work, our goal is to answer the research question: what are the best evaluation metrics and methods to deal with subjective crowdsourced annotations, that have the potential for significant annotator disagreements and contribute to a "quality crisis"? This work focuses on the field of argumentation, which has several applications, ranging from the automatic extraction of arguments in argumentation mining (AM) to the use of argument maps for decision making and deliberation. We use a crowdsourcing platform with two groups: 1) an open crowd with quality control settings, and 2) a small group of trustworthy master annotators (in-house workforce) without these settings but several rounds of feedback. This is partly a secondary data analysis work, as the open crowd dataset comes from a previous publication focused on multimodal machine learning (Mestre et al. 2021), and here we add an additional layer of analysis that enriches that work by comparing it with an in-house workforce, focusing on the agreement metrics and annotators' disagreements. We make available these datasets, with both quantitative and qualitative data, in our project's repository to be used as benchmarks[1]. Our objective is not to apply any of the approaches that embed subjectivity in evaluation or model training (Gordon et al. 2022; Geng 2016), but to understand if current metrics, such as Krippendorf's alpha, are appropriate for crowdsourcing subjective data. Finally, we qualitatively review the most common sources of error when annotating this type of data, which could be used to mitigate bias in future studies.

[1]https://github.com/rafamestre/subjective-crowdsourcing-argmin/.

## 2 Argumentation mining

In our study, we wanted to assess the feasibility of using crowdsourcing techniques to obtain high-quality annotated data for argumentation mining. Argumentation is an old field of research that can be traced back to Aristotles' Rhetoric (Cope and Sandys 2010) and more recent but influential works like Perelman and Olbrechts-Tyteca's New Rhetoric (1969). Since then, many different argumentation frameworks have been created with the aim of systematising how machines (and humans) should deal with arguments and information and how to draw conclusions from them. These frameworks are generally based on models of argumentation, like Toulmin's or Walton's models. Toulmin (2012), for instance, considered that the microstructure of arguments is divided into six categories: claim, grounds, warrant, backing, modality and rebuttal. Walton, Reed, and Macagno (2008), on the other hand, proposed the use of argument schemes, or templates of arguments that are used in ordinary conversation during an argument.

However, there is still not a clear consensus on which argumentation model is more appropriate for each specific task and, thus, efforts in the recent field of argumentation mining have been aimed at the automatic construction of argumentation frameworks (Cabrio and Villata 2012; Lippi and Torroni 2016). One of the most used frameworks (due to its simplified nature) to understand an argument is by dividing it into a final claim or conclusion and a set of premises (also evidence or reasons), together with a defined inference between them (Lippi and Torroni 2016; Walton 2009). This inference relation, using an extension of Dung's abstract framework (Dung 1995) can either attack or support relations (Lippi and Torroni 2016; Peldszus and Stede 2013). The field of argumentation mining has traditionally opted for identifying claims or premises in essays or dialogues and then identifying the relationship between them (Lawrence and Reed 2020). Carstens and Toni (2015) proposed a relation-based approach towards argumentation mining in which the relations between different arguments would be identified first, claiming that this relation is highly contextual, as a fact can be used as a supporting statement for an argument, but it can at the same time be attacking another one, or be completely unrelated to the argument in a different context. This approach has been successfully used in different works using neural networks to advance the field of argumentation mining (Bosc, Cabrio, and Villata 2016a,b; Cocarascu and Toni 2017).

It is clear that the detection of an argument is not an easy task, and the field of argumentation mining suffers from a lack of large training datasets that can be used in a variety of scenarios, ranging from student essays to political debates or internet discussions. Most of the available datasets come from annotations made by experts in argumentation or, at least, research assistants that have been trained in depth in it. This task thus takes a long time and is not cost efficient to be done at larger scales. Using crowdsourcing techniques comes with a series of obvious advantages, like higher throughput and less costs, but it also raises several

concerns. Given the subjectivity of the task, the level of disagreement between annotators can be much higher compared to other tasks. The complexity and context dependence of argumentation might be too high for crowd annotators who want to engage in simple and quick tasks that do not require deep thinking. A previous study a decade old (Peldszus and Stede 2013) showed that there was generally a low level of agreement between annotators identifying argumentative structures, especially if they were not trained in depth.

Although these disadvantages might be sufficiently compelling to abandon the idea of using crowdsourcing techniques in argumentation, there is one interesting aspect that persuades us to continue investigating it. In real-life applications of argumentation mining, which can range from assisted decision making in deliberations to understanding polarised discussions online, it is important to consider how real people understand arguments. Given the significant complexity of the theory of argumentation, it is not exceptional to think that people might have different opinions on what constitutes an acceptable argument in a discussion. There is a tension, however, on how to test this with crowdsourcing and whether we can even define a 'ground truth' in this case (Plank 2022; Aroyo and Welty 2015). In this study, we take a dual approach in which we first ask an open crowd to annotate arguments using expert 'ground truths' and then we use a smaller in-house workforce without ground truth to understand where differences emerge.

## 3 Quality control mechanisms

To partly solve the "quality crisis", crowdsourcing platforms provide an extensive library of settings to ensure the quality of the results. These QCMs might not be completely applicable for studies interested in surveying a representative sample of the population where there are "no wrong answers", but they come in handy for studies gathering training data for machine learning models applied to social science data. Here, we review some of the QCMs that can be found in platforms like MTurk and Appen (previously called Crowdflower), expanding upon the list of Alabduljabbar and Al-Dossari (2019):

- **Test questions:** Annotators can be assessed against a set of pre-annotated test questions (generally done by experts), which serve as a filter to eliminate poorly performing annotators. These test questions can be shown at two steps of the annotation task: 1) at the beginning, as an initial quiz; and 2) during the task itself, randomly distributed in each batch of annotations. A "trust theshold" can be defined, and those with scores below this threshold are removed from the task and their previous annotations discarded.
- **Demographic filtering:** Filtering based on demographic features like country and language is used to match annotators with tasks that require specific knowledge or skills. However, as pointed out by Kennedy et al. (2020), the use of virtual private servers can mask contributors' location to bypass this filtering. Some platforms also block plugins like Google Translate to ensure language proficiency.
- **Worker reputation or level:** Annotators gain reputation scores and levels as they participate in different tasks. High reputation scores and advanced levels grant access to more complex tasks that require experienced contributors and usually offer better pay.
- **Redundancy:** It is highly recommended that redundancy is added to the annotation task in the form of multiple annotators assessing the same data, resulting in agreement scores that help researchers evaluate the overall quality and ambiguity of the task. Platforms may allow dynamic judgments, where more annotations are requested if agreement falls below a specific threshold.
- **Response rules:** Researchers can set limitations or expectations based on prior knowledge of answer distributions to maintain quality and ensure annotators give thought to their responses. For example, setting limits on the number of positive and negative annotations for a specific contributor or requiring a minimum time spent on each annotation can help prevent random or biased annotations.

It is important to mention that these QCMs could go against the value of subjectivity and disagreement on data annotation and evaluation, as we assume that a 'ground truth' exists and that experts' annotations are 'better' (Aroyo and Welty 2015). However, without these test questions, crowdsourcing annotations becomes a difficult task, as one cannot assess whether annotators disagree because they have different opinions or because they did not understand the aims of the task or are giving random answers. Researchers have been working on implementing different methods to identify 'spammers', such as multi-annotator competence estimation (MACE) (Hovy et al. 2013), but annotation platforms are currently only implementing methods based on basic agreement. In theory, it could be possible to implement estimators such as MACE after the annotation process has finished, without discarding any crowdworkers during the process, but this can be costly, as a lot of annotators will be allowed to continue providing spam annotations without being stopped. For these reasons, in our study we decided to compare two different crowds, one with 'ground truths' in the form of test questions and another one without test questions but with several rounds of feedback to ensure the task was understood.

## 4 Methodology

We used the crowdsourcing platform Appen[2] (previously called CrowdFlower) to annotate for argumentative relation sentences from the US presidential debates of 2020 between Donald Trump and Joe Biden, as well as the vice-presidential debates between Mike Pence and Kamala Harris. We focused on the relational aspects of argumentation, that is, instead of asking annotators to identify the arguments in the candidates' speeches and then form argument maps connecting them, we followed the relational approach suggested by Carstens and Toni (2015). This approach is based

[2]https://appen.com/

on the assumption that the relation of support or attack between two argumentative units can be highly contextual. In one context, one argumentative unit might be an argument supporting one idea, but it might present a completely unrelated idea in another context. An example of the annotation task can be seen in Figure 1 of the Appendix and the instructions in the data repository.

Using crowdsourcing techniques allowed us to investigate the "knowledge of the crowd" regarding argumentation. This means that since annotators cannot be trained in depth about argumentative frameworks or philosophy of argumentation, their judgements of what constitutes a support/attack relation might be subjective and different to those provided by the theories of argumentation. We believe this experiment can provide useful information about how people approach arguments in real life, since an argumentation mining tool will need to differentiate arguments made by people, who aren't always guided by rational logic alone.

We performed our study using two sets of contributors from the Appen platform: an open crowd and an in-house workforce. For the first case, we use a secondary dataset from a previous study in which we annotated argumentative relations for multimodal argumentation mining (Mestre et al. 2021). We refer to that study detailed information about the performance of the machine learning models, whereas here we discuss in depth the QCMs and inter-annotator agreement metrics. This open crowd consisted of contributors from the United States of Level 3 (the higher level of the platform). After this work, we performed a follow-up study with an in-house workforce: a set of 9 contributors from an internal team of the company based in the Philippines (although we ended up with 8, see Section 6), who have significant experience in data annotation and are assigned in batches to specific jobs when higher accuracy and control are needed. According to the company, this crowd presented diverse demographics in terms of gender (3 female, 4 male, 2 non-binary) and age (5 between 21-30 and 4 between 31-40). Although they were all Filipino, their self-assessed level of English ranged from C2-C1 (4 people) to B2-B1 (5 people) and they all had a bachelors' degree. Besides nationality, this group could be comparable to a team of graduate or under-graduate students generally hired for coding tasks in social science. This in-house workforce would be analogous to the "master level" contributors of MTurk, albeit perhaps with a slightly higher degree of control.

Whereas the open crowd required a significant amount of testing and QCMs in place, as will be discussed later, the in-house workforce did not need quality settings such as test questions, as direct contact with the manager of the team during the testing phase for feedback ensured that the contributors would not be trying to 'trick the system' as the open crowd might. This higher trust allowed us to include a free-text box for the in-house workforce, where they gave reasons on why they chose the annotation, allowing us to understand potential sources of errors and reasons for disagreement. We acknowledge that there might be cultural differences affecting our results, as the open crowd was based in the US and the in-house workforce in the Philippines. This platform, however, would not let us choose an in-house workforce from a different country. It would have been interesting to control for gender, level of education or political leaning in the open crowd, for instance to potentially apply a jury learning-type of model that considers this (Gordon et al. 2022). The features of the platform made this a difficult task, as it is not possible to show annotators demographic questions only once during the annotation task.

In Table 1 we review three of the most common inter-annotator agreements metrics to evaluate crowdsourcing tasks. Kappa statistics like Cohen's or Fleiss' are widely used in the social sciences, and they measure the reliability of the coding considering the agreement expected by chance (Cohen 1960; Fleiss 1975). Cohen's kappa is a robust metric that can be used with only two annotators, whereas Fleiss' kappa is a generalisation that allows for multiple annotators. Krippendorff's alpha, although it also measures the agreement compared to that expected by chance, is the most flexible of the three, as it allows for many data types and can handle missing data (Krippendorff 2018). In our crowdsourcing study the kappa metrics cannot be used, as they assume that all the annotators code each single data point. Due to the nature of crowdsourcing, this is not entirely possible, as crowdworkers annotate small subsets of the data that have little overlap with each other. For that reason, Krippendorff's alpha is the only metric that can be used in this case, as it corrects for crowdworkers not annotating all the data available, but a subset of it. However, one important criticism of these metrics is that they reproduce the ideal of a 'ground truth' and that higher agreement values mean higher data quality (Aroyo and Welty 2015).

A previous study (Peldszus and Stede 2013) showed that untrained annotators tended to perform poorly on argumentation tasks. Thus, as open crowd annotators (even if they are from the upper quality level) have less experience in annotation task than the in-house workforce (or master annotators), our first hypothesis was:

**H1:** The agreement between contributors in the open crowd is lower than the agreement between contributors in the in-house workforce.

To assess the relationship between agreement and ambiguity or difficulty in the task, we asked the contributors to self-assess their confidence when providing annotations in a Likert scale from 1 (not at all confident) to 5 (very confident). We decided to use this self-confidence score instead of ambiguity, as Chen et al. (2018) used, because, when identifying the relationship between arguments, the context and the meaning of the sentences might be completely unambiguous, but the annotators might still struggle to identify the relation. This self-confidence score might help filter out low-quality annotations for which the contributors were not sure at all. Therefore, our second hypothesis was:

**H2:** Contributors in both the open crowd and the in-house workforce agree more often in annotations provided with higher confidence.

Finally, although we use a fairly simple argumentation scheme (support/attack/neither) to account for argumentative context and avoid increasing the complexity of the task, we acknowledge that this framework asks, indirectly, for two evaluations. First, they need to decided whether two sen-

| Metric | Nb annotators | Type of data | Handles missing data |
|---|---|---|---|
| Cohen's Kappa | 2 | Nominal | No |
| Fleiss' Kappa | More than 2 | Nominal | No |
| Krippendorff's Alpha | More than 2 | Nominal, ordinal, interval, and ratio | Yes |

Table 1: Most common inter-annotator metrics and some of their characteristics. Missing data can be understood as non-overlapping annotations, like it occurs in crowdworking, where each annotator only sees a very small subset of the data.

tences form an argument or not and, if so, evaluate if they are supporting or attacking each other. In this context, our last hypothesis is:

**H3:** The majoritarian source of disagreement is between the support/attack class and the neither class. Few confusions occur between support and attack only.

## 5 Open crowd

### Quality control mechanisms

As previously mentioned, the annotation task with the open crowd required a series of QCMs to ensure that the obtained annotations were of high quality, mainly fuelled by the "quality crisis" of crowdsourcing platforms, even if they go against human subjectivity (Plank 2022). Before the final launch of the annotation task, several iterations of test jobs were performed to understand whether the instructions were clear, the quality settings appropriate and the pay per annotation fair.

The following settings were finally established, based on those reviewed in 3. Contributors were selected at level 3 from the platform (the higher available) and from the United States to complete the annotation task. Contributors were tested with a quiz before annotating data and with test questions randomly distributed in the main task. They had to always score above a 'trust score' of 80% or they were dismissed. To ensure quality, a minimum time of 90 seconds per page was set (each page containing 4-6 questions), and contributors who completed pages too quickly were released from the task. Answer distribution rules were enabled after 20 judgments, and contributors who judged over 60% of relations as "support" or over 35% as "attack" were dismissed. A minimum of 3 annotations per pair of sentences were requested, but dynamic collection of judgments was enabled if the annotation agreement fell below the selected threshold of 70%, up to 7 judgements.

In this annotation task, 104 trusted contributors out of 287 who attempted it participated (26.6%) and a total of 103 test questions were used. Overall, 21,646 trusted annotations were collected, with 5,746 belonging to test questions and 15,900 to random pairs. Additionally, 1,663 annotations were deemed untrustworthy due to the quality settings. In total, 4,104 pairs of sentences were labelled as support, attack or neither. Full information about the construction of the dataset can be found in its original source (Mestre et al. 2021) and our the GitHub of this work[3].

[3]https://github.com/rafamestre/subjective-crowdsourcing-argmin/.

| Measure | With test questions | Without test questions |
|---|---|---|
| Single disagreements | 1401 | 1339 |
| Tri-disagreements | 310 | 278 |
| No disagreement | 2340 | 2331 |
| Percentage agreement | 57.76% | 59.04% |
| Total sentences | 4051 | 3948 |
| Krippendorff's $\alpha$ | 0.43 | 0.24 |

Table 2: Descriptive statistics of agreement between annotators from the open crowd force.

### Statistics of agreement

Table 2 shows the statistics of agreement for the open crowd annotation task. As already discussed, each pair of sentences was annotated by at least 3 trusted contributors, increasing up to 7 when the disagreement was high. Test questions, however, were annotated by many contributors, since they were shown during the initial quiz and as test questions. This table shows that there was a 59.04% of agreement (annotations where every single contributor agreed on the label) without including test questions, and 57.76% if those are included in the calculation. Only 7% of the annotations (278 out of 3948) had tri-disagreements, i.e., the contributors disagreed on the three labels (having one annotations of each support/attack/neither), indicating that generally the disagreement occurred only between two of the labels.

We also report Krippendorff's alpha as a way to measure the inter-annotator agreement. Out of all the agreement metrics, Krippendorff's alpha is probably the most flexible one and the only one that adapts to our type of crowdsorced data (Krippendorff 2018). Despite its flexibility, we noted a striking difference between the value including test questions ($\alpha = 0.43$) and not including them ($\alpha = 0.24$), even though the overall percentage agreements were roughly the same. The fact that including such a small number of sentences (103 test questions), albeit with many annotations each, had such a strong influence on the value of Krippendorff's alpha makes us think that this metric is highly sensitive to the data points annotated by a large number of contributors, that is, the regions of overlap between different annotators. As test questions are annotated by many contributors, they have a strong influence on skewing the metric towards a higher agreement. In any case, the range 0.24-0.43 indicates a rather low agreement Krippendorff (2018). This effect is especially apparent when we filter annotations by the trust of the annotator (the percentage of correctly answered test questions). Although one could think that higher trust should lead to larger agreement between annotators, Krippendorff's

| | All annotations | | | | | Self-confidence = 5 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Trust** | $\alpha$ | **Workers** | **# Annots.** | **Supports** | **Attacks** | $\alpha$ | **Workers** | **# Annots.** | **Supports** | **Attacks** |
| $\geq 0.80$ | **0.24** | 95 | 15698 | 2568 | 1225 | 0.31 | 79 | 4318 | 600 | 302 |
| $\geq 0.85$ | 0.24 | 90 | 14767 | 2373 | 1160 | 0.32 | 75 | 3954 | 565 | 293 |
| $\geq 0.90$ | 0.23 | 70 | 12187 | 1880 | 910 | 0.30 | 58 | 3238 | 394 | 233 |
| $\geq 0.95$ | 0.23 | 47 | 6580 | 820 | 449 | 0.29 | 35 | 1685 | 154 | 91 |
| $= 1.0$ | -0.04 | 21 | 462 | 144 | 77 | -0.20 | 15 | 117 | 31 | 16 |

Table 3: Descriptive statistics of agreement between annotators in the open crowd according to their reported self-confidence and the trust score of the annotators, without including test questions. Negative Krippendorff's $\alpha$ can be potentially explained by the low overlap of annotators when the results are filtered (see Section 5). 'Trust' is defined as the amount of test questions correctly answered by the annotators (see 'Quality control mechanisms' in Section 5).

alpha can be negative when we only include annotators with trust = 1 (those that did not fail any test question). In these cases, the overlap between annotators is so low that a couple of disagreements can heavily tip the balance.

To assess H2, i.e., whether annotators agree more often in annotations they provide with higher confidence, we filtered the results by self-confidence score and trust in the annotator. Table 3 shows the results without including the test questions. We can see that the 'trust score' of the annotator does not seem to significantly affect the inter-annotator agreement, but a high self-confidence is correlated with a higher inter-annotator agreement. It needs to be noted that the results become less reliable when we heavily filter the amount of data, e.g., by increasing the trust threshold to 1. As the number of contributors decreases, it is very unlikely that they annotated the same units of data and therefore the calculation of the inter-annotator agreement contains very few data points to give an accurate result.

## 6 In-house workforce

Given the complexity of annotating argumentative relations and the aforementioned "quality crisis" of crowdsourcing platforms (where contributors use all sorts of means to earn rewards without properly engaging in the task), we wanted to further investigate whether using an in-house workforce of highly trained and professional annotators (see Section 4) from the same platform would improve the results from the previous section. The in-house workforce came with several advantages compared to the open crowd. It provided with a higher degree of control which workers contributed to the task, as it was managed internally. The instructions for the task remained identical to those used with the open crowd, so that we could compare our results. We also included a free-text box for the annotator to provide reasons on why they chose their annotations. In our initial testing with the open crowd, we discovered that crowd annotators would not engage with a free-text box, even if we made it compulsory. Direct access to the in-house workforce through their manager assured us that they would provide valuable feedback through this text box, albeit at a higher cost.

The in-house workforce was a team of annotators based in the Philippines with a proficient level of English so that they were able to understand the task. We had direct contact with the manager of the team, who helped us define the task and transmit to the contributors our expectations. We launched 4 test jobs before the final job. After the 4 tests, the team of annotators was reduced to 9, who showed a good understanding of the task at hand after our feedback (but eventually reduced to 8, as one annotator still showed a lack of understanding, see Table 4). The first test helped the team manager assess what was needed from workers and estimate a price per task, which was higher than the price set in the open crowd, as it was a dedicated team of workers. The other three tests helped us identify misconceptions in the task and common sources of errors in a way that was not possible with the open crowd. We review some of those common sources of error in more detail in Section A of the Appendix, but we summarise here that there was a tendency to assume the intentions of the speakers, instead of focusing on the arguments themselves. This lead to considering sentences like "thank you" as supportive or even the presence of transcription labels like "[crosstalk]" as an attack relation, as that meant they were interrupting each other. In another case, the presence of a sentence that simply contained "..." due to a transcription error prompted an annotator to consider it 'support', since for them it was "a silent statement to re-affirm confidence in the first sentence".

| **Measure** | **All 9 annotators** | **Removing ID 60000** |
|---|---|---|
| Single disagreements | 743 | 621 |
| Tri-disagreements | 64 | 24 |
| No disagreement | 1145 | 1307 |
| Percentage agreement | 58.66% | 66.96% |
| Total sentences | 1952 | 1952 |
| Krippendorff's $\alpha$ | 0.16 | 0.21 |

Table 4: Descriptive statistics of agreement between annotators from the in-house workforce.

### Statistics of agreement

As the price per judgement was higher for the in-house workforce (twice the cost), especially due to the free-text box, we only asked them to annotate a random sub-set of the original annotations so that we could compare, totalling 1,952[4]. Table 4 displays the statistics of the in-house

[4]Our initial dataset size was of 2,000. However, 48 of those pairs of sentences randomly turned out to be previous test questions for the open crowd. Although there were no test questions for

| | Attack | Neither | Support |
|---|---|---|---|
| Attack | 0% | 45.09% | 4.83% |
| Neither | | 0% | 50.08% |
| Support | | | 0% |

Table 5: Disagreement matrix for annotations with the in-house workforce.

| **Self-conf.** | **Kripp. alpha** | **# Annot.** | **# Supp.** | **# Att.** |
|---|---|---|---|---|
| 1 | 0.21 | 4738 | 512 | 470 |
| 2 | 0.21 | 4737 | 512 | 470 |
| 3 | 0.21 | 4730 | 511 | 470 |
| 4 | 0.22 | 4501 | 463 | 413 |
| 5 | 0.21 | 3685 | 320 | 315 |

Table 6: Inter-annotator agreement for in-house workforce annotators (after removing outlier annotator) according to reported self-confidence. Krippendorf's $\alpha$ column refers to its value for clusters of annotators with the same reported self-confidence.

workforce annotation. Interestingly, we can see that the percentual level of agreement is almost identical to that of the open crowd, 58.66%. The number of cases in which tri-disagreements occurred (one annotator said support, another attack and another neither) was reduced to 3%, an indication that most of the disagreements occur between two specific labels. The Krippendorff's alpha metric is rather low in this case, $\alpha = 0.16$, which indicates a very poor agreement even compared with the open crowd. We again advise taking this metric with care, as we found it above to be extremely sensitive to the annotation overlap of different annotators. Since in this case we do not have a "trust score" to assess whether the annotators can be trusted, we performed a leave-one-out analysis in which we filter out the annotations from one single annotator and we calculate again the inter-annotator agreement (Table 9 in the Appendix). We found the agreement scores to be fairly similar except when contributor ID 60000[5] was left out. This contributor seemed to be a large source of disagreement and when we removed their contributions, the overall agreement increased to 66.96% and $\alpha = 0.21$, with a reduction of tri-disagreements to 1.2%.

Table 5 shows the disagreement matrix for this annotation task after removing this contributor. We can see that disagreeing between support and attack relations happened very few times, about 4.83%. The largest source of disagreement is between support and neither, and between attack and neither. This goes in line with our third hypothesis (H3). It is not a matter of completely misunderstanding the direction of the argument, e.g., believing an argumentative attack is a support, but having problems understanding if two statements form part of the same argument or not.

Our hypothesis H2 stated that crowd workers that provided annotations with higher reported self-confidence would agree more on their annotations. Our first study with an open crowd seems to support this hypothesis, as the inter-annotator agreement score provided by Krippendorff's alpha increased when we filtered for annotations with higher self-confidence. Table 6 shows the corresponding results for the in-house workforce, after eliminating the outlier annotator. In this case, we see no changes in Krippendorff's alpha after filtering for self-confidence, which could be related to the higher amount of overlap between in-house annotations. In any case, it is unclear whether providing annotations with higher self-confidence actually leads to higher agreement between annotators and we cannot confirm H2.

Finally, our first hypothesis (H1) stated that annotators in the open crowd would perform more poorly than those from the in-house workforce. However, from Tables 2 and 4 we find very similar levels of agreement (approximately 57-59%) between both crowds. Filtering the outlier annotator in the in-house workforce, though, increased the level of agreement to 67%. Interestingly, if we use the inter-annotator agreement of Krippendorff's alpha as a metric, we find the exact opposite trend: annotators from the open crowd tend to agree more often than in-house annotators. We raise a word of advice again about this metric, which seems to be highly dependent on the overlap between workers annotating the same row of data multiple times. Although this metric was designed to deal with exceptions such as multiple coders not annotating the same rows of data, it was never designed to be used in crowdsourcing, where this is the norm, rather than the exception. It is also interesting to analyse where both sets of annotators disagree with one another, after comparing their final annotations (the majority label) as annotations provided by two different coders. Table 7 shows the statistics of agreement, where we can see that they agree on about 84% of the annotations. The inter-annotator agreement scores show, however, low levels of agreement, due to the fact that the dataset is imbalanced towards the neither case and, although they have a high percentage of agreement, it is most likely coming from the neither class. Both Fleiss' and Cohen's kappa[6] can be consider a fair agreement between crowds, although the same value for Krippendorff's alpha is considered to be rather low. The acceptable agreement ranges are in constant debate, as researchers in different disciplines disagree on what can be considered good, fair or poor (McHugh 2012). In Table 8 we can see that the source of disagreement comes, as hypothesised in H3, from confusion between support and neither relations and attack and neither relations. Of all disagreements, only 3.22% came from support and attack, indicating that annotators can differentiate both cases rather well.

The reasons behind these disagreements are varied and the written feedback by the contributors helped disentangle

the in-house workforce, we decided to eliminate those 48 pairs of sentences from our analysis so we could compare only final, and not test questions.

[5]To avoid confusion between the open crowd worker IDs, which were labelled starting from 0, we gave the in-house workforce annotators IDs starting from 10000, up to 90000.

[6]In this case, since we only have two sets of annotators (open crowd and in-house workforce), each annotating every single data row, we can use Fleiss' and Cohen's data, which could not be used before when we considered individual annotations.

| Measure | Value |
|---|---|
| Number of disagreements | 311 |
| Number of no disagreements | 1642 |
| Percentage agreement | 84.08% |
| Total sentences | 1953 |
| Fleiss kappa | 0.24 |
| Cohen's kappa | 0.25 |
| Krippendorff's alpha | 0.24 |

Table 7: Statistics of agreement between crowd-force and in-house force, out of 1953 in common.

| | Attack | Neither | Support |
|---|---|---|---|
| Attack | 0% | 35.69% | 3.22% |
| Neither | | 0% | 61.09% |
| Support | | | 0% |

Table 8: Disagreement matrix for final annotation labels of the open crowd and in-house workforce.

them. In fact, annotators fully agreed on support or attack only 20 and 19 times, respectively. In general, those were cases with very clear context and speech, and those with more nuanced or more chaotic resulted in high level of disagreement. The sources of disagreement were cases where the speaker said something affirmative like "yes, it was" or "agreed", which tended to be assumed as *argumentative* support, even though there is no argument in it. For instance, in the following pair of sentences: "Was it a mistake to support it?", "Yes, it was". The annotators claimed that this was a support relation "because [he] is showing that he agrees with the first sentence".

Likewise, it was very common to label pairs are argumentatively attacking if one of them was simply stating the opposite of the other sentence. Common examples were pairs of sentence where one of them stated "that is not true". However, annotators were able to find attack relations in pairs of sentences that were not obvious, such as: "And I'm going to eliminate those tax cuts", "But why didn't you do it over the last 25 years?". These sentences were labelled as attack, as the question could be rephrased to mean "you're not going to eliminate the tax cuts because you had 25 years to do it and you didn't do it." More detailed examples can be found in Section B of the Appendix.

## 7 Conclusion

This paper investigates the feasibility of using crowdsourcing techniques for subjective data annotation and examines inter-annotator metrics, common sources of error, and disagreements. As part of our analysis, we assess the potential of crowdsourcing for identifying argumentative relations in political debates, using annotations from an open crowd of annotators and a smaller in-house workforce from the same company. Our study reveals the complexity and subjectivity of this task, which may not be well-suited for the fast-paced levity of crowd annotation. Nonetheless, the data provides valuable insights into how people interpret arguments without extensive training in argumentation theory. For example, we found that people tend to associate interrupting someone with an argumentative attack relation, and that it is important to distinguish between a personal attack and an argumentative attack. Similarly, people tend to associate support towards someone with simple politeness, such as thanking the other person or simply agreeing on a point. Nonetheless, our study also highlights the subjective nature of argumentation, as most disagreements tend to occur between the labels support and neither or between attack and neither. Due to that, one of the main conclusions of our study is that quality control mechanisms (for an open crowd) or extensive feedback rounds (for a dedicated workforce) are necessary to increase agreement, but they still might not be enough for subjective coding, especially since defining 'ground truths' for subjective tasks is problematic (Basile et al. 2021).

In particular, our work sheds light into the issue of crowdsourcing and coding in subjective tasks, as well as finding adequate benchmarks and evaluation metrics for them. Our results with both sets of annotators (the open crowd with quality control mechanisms and the in-house workforce) had low Krippendorf's alpha (going from 0.21 to 0.43). This is unfortunately the only one of the most common inter-annotation metrics (like Cohen's or Fleiss' kappa) that can be used for this task, as it is robust to missing annotations from annotators (as they code different regions of the dataset). We also noticed that this value fluctuated a lot when filtering the data according to the annotator self-confidence and test questions. This leads us to believe that this metric is highly sensitive to the amount of overlap between annotators. As we had in the first case 104 contributors providing 3-7 judgements per pair of sentences, the overlap region between annotators that coded at least one pair in common should be rather low. This makes the evaluation metric in this case highly sensitive to disagreements in this small area of overlap. When test questions are included in the calculation, as all annotators were continuously tested on the same 103 questions, the region of overlap increases, as well as value of the alpha metric.

Although Krippendorff's alpha is an adequate metric to test agreement in cases with multiple annotators that do not necessarily annotate all the same data, we raise some concern around its use on crowdsourcing. This metric is robust with a few annotators coding large amounts of data and is robust to them missing a few data points. However, as it relies on overlaps between annotators, it might not be completely suitable for cases with many annotators only coding small amounts of data. Our work emphasises the need for evaluation metrics that are robust in crowdsourcing tasks where low overlap is the norm rather than the exception. Moreover, we believe our results and datasets could be useful as benchmarks for future research in this area, as well as for increasing data quality, informing the design of further work and mitigating common errors in subjective coding, in particular, in argumentation mining. We believe that follow-up studies should focus on developing crowdsourcing-robust evaluation metrics and studying the extent of the impact of annotation overlap in Krippendorff's alpha, for instance through simulated data.

## Ethical statement

Ethics approval for this research was received from the University of Southampton's Faculty of Social Science Ethics and Research Governance committee, Ref: 66226, Date 22/07/2021. This work could have potential positive societal implications, such as improving the quality of argumentation analysis, which can aid in decision-making and deliberation in various domains, such as politics, law, and education. The research also provides insights into how people interpret arguments without extensive training in argumentation theory, which can be valuable in improving communication and understanding between individuals and groups with diverse backgrounds and perspectives. However, we acknowledge concerns regarding the use of automated models based on natural language processing in decision making, with serious impacts related to accountability, transparency, and biases. Our findings suggest that subjectivity plays a significant role in annotating this type of data, which could potentially impact the performance of automated models trained on such data. Therefore, we caution against the use of such models in automated decision making, and emphasise the importance of thorough human review and oversight in all decision-making processes that involve subjective data. Furthermore, we recognise the potential negative societal implications of crowdsourcing techniques, such as the exploitation of low-wage workers or the perpetuation of biases and stereotypes. Therefore, we took measures to ensure that our study adheres to ethical standards, such as providing fair compensation above their country's minimum wage, and using quality control mechanisms to ensure the accuracy and fairness of annotations.

## Acknowledgements

This work has been funded by UK Research and Innovation (UKRI) funding (grant ref MR/S032711/1) and by the Web Science Institute of the University of Southampton (project PP-2020-Mestre). This work was also supported by the Natural Environment Research Council (NE/S015604/1) and Economic and Social Research Council (ES/V011278/1).

## References

Alabduljabbar, R.; and Al-Dossari, H. 2019. A dynamic selection approach for quality control mechanisms in crowdsourcing. *IEEE Access*, 7: 38644–38656.

Aroyo, L.; and Welty, C. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1): 15–24.

Basile, V.; Fell, M.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; Poesio, M.; Uma, A.; et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, 15–21. Association for Computational Linguistics.

Benoit, K.; Conway, D.; Lauderdale, B. E.; Laver, M.; and Mikhaylov, S. 2016. Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, 110(2): 278–295.

Berinsky, A. J.; Huber, G. A.; and Lenz, G. S. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political analysis*, 20(3): 351–368.

Bosc, T.; Cabrio, E.; and Villata, S. 2016a. DART: A dataset of arguments and their relations on Twitter. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, 1258–1263.

Bosc, T.; Cabrio, E.; and Villata, S. 2016b. Tweeties squabbling: Positive and negative results in applying argument mining on social media. *COMMA*, 2016: 21–32.

Cabrio, E.; and Villata, S. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 208–212.

Carstens, L.; and Toni, F. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, 29–34.

Chen, N.-C.; Drouhard, M.; Kocielnik, R.; Suh, J.; and Aragon, C. R. 2018. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2): 1–20.

Cocarascu, O.; and Toni, F. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1374–1379.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Cope, E. M.; and Sandys, J. E. 2010. *Aristotle: Rhetoric*, volume 2. Cambridge University Press.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2): 321–357.

Fleiss, J. L. 1975. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 651–659.

Geng, X. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7): 1734–1748.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19.

Hovy, D.; Berg-Kirkpatrick, T.; Vaswani, A.; and Hovy, E. 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*, 1120–1130.

Kennedy, R.; Clifford, S.; Burleigh, T.; Waggoner, P. D.; Jewell, R.; and Winter, N. J. 2020. The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, 8(4): 614–629.

Kittur, A.; Chi, E. H.; and Suh, B. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 453–456.

Krippendorff, K. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Lawrence, J.; and Reed, C. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4): 765–818.

Lippi, M.; and Torroni, P. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2): 1–25.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282.

Mestre, R.; Milicin, R.; Middleton, S. E.; Ryan, M.; Zhu, J.; and Norman, T. J. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, 78–88.

Peldszus, A.; and Stede, M. 2013. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, 196–204.

Perelman, C.; and Olbrechts-Tyteca, L. 1969. The new rhetoric: a treatise on argumentation, trans. *John Wilkinson and Purcell Weaver (Notre Dame, IN: University of Notre Dame Press, 1969)*, 19.

Plank, B. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10671–10682. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Stewart, N.; Chandler, J.; and Paolacci, G. 2017. Crowdsourcing samples in cognitive science. *Trends in cognitive sciences*, 21(10): 736–748.

Toulmin, S. E. 2012. The uses of argument.

Walton, D. 2009. Argumentation theory: A very short introduction. In *Argumentation in artificial intelligence*, 1–22. Springer.

Walton, D.; Reed, C.; and Macagno, F. 2008. *Argumentation schemes*. Cambridge University Press.

| Worker ID eliminated | Krippendorff's alpha |
|---|---|
| 10000 | 0.15285635191593694 |
| 20000 | 0.15075059419550296 |
| 30000 | 0.15988598626132033 |
| 40000 | 0.14917142926577298 |
| 50000 | 0.16616592648959272 |
| 60000 | 0.20789420905709088 |
| 70000 | 0.15865013196179212 |
| 80000 | 0.14464617120113876 |
| 90000 | 0.14813681176757842 |

Table 9: Ablation study eliminating one by one each worker and assessing the effect in Krippendorff's alpha.

# A Common misconceptions during test in-house annotations

In the test launches, we noticed that there was a tendency to assume that statements from different speakers are an argumentative attack by default. For instance, consider this extract from one of the debates between Donald Trump and Joe Biden:

- Joe Biden: "Number two-"
- Donald Trump: "Chris."
- Joe Biden: **"Number three."**
- Donald Trump: **"They said it would take... No, you're on number two."**
- Chris Wallace: "No."

The contributors were asked to annotate the argumentative relation between the highlighted sentences, which are not argumentative, as the extract is from series of interruptions between both candidates. In the first iteration, however, two of the three annotators thought that was an attack relation, simply because it was Donald Trump trying to interrupt Biden. A similarly interesting case occurred with certain 'support' relations. We noticed that contributors tended to consider a relation as support when the message was positive. See for instance this interaction between Mike Pence and Kalama Harris:

- Mike Pence: "The Green New Deal is on their campaign website."
- Mike Pence: **"And as USA TODAY said, it's essentially the same plan as you co-sponsored with AOC when she submitted it in the Senate."** [...]
- Mike Pence: "We don't need a massive $2 trillion Green New Deal that would impose all new mandates on American businesses and American families."
- Kamala Harris: **"Thank you."**

This is part of a long speech by Mike Pence about the Green New Deal and the second sentence by Kamala Harris is simply "Thank you", most likely thanking Pence or the moderator for giving her a turn to speak. The three annotators that judged this considered the relation to be of 'support'. Despite the instructions stating clearly and with examples that the sentences needed to be somewhat argumentative, we encountered several similar examples. In this particular case, it was transmitted to us by the manager that the workers assumed that 'thank you' was an acknowledgement and agreement of what Pence said, and thus support. We reminded them that simple agreement was not sufficient to be considered argumentative support.

There was a tendency to over-read much of the intentions of the speakers and over-complicate the task. For instance, in this exchange:

- Savannah Guthrie: "You retweeted it."
- Donald Trump: "That was a retweet."
- Donald Trump: **"That was an opinion of somebody-"**
- Savannah Guthrie: "But-"
- **Donald Trump: "...."**

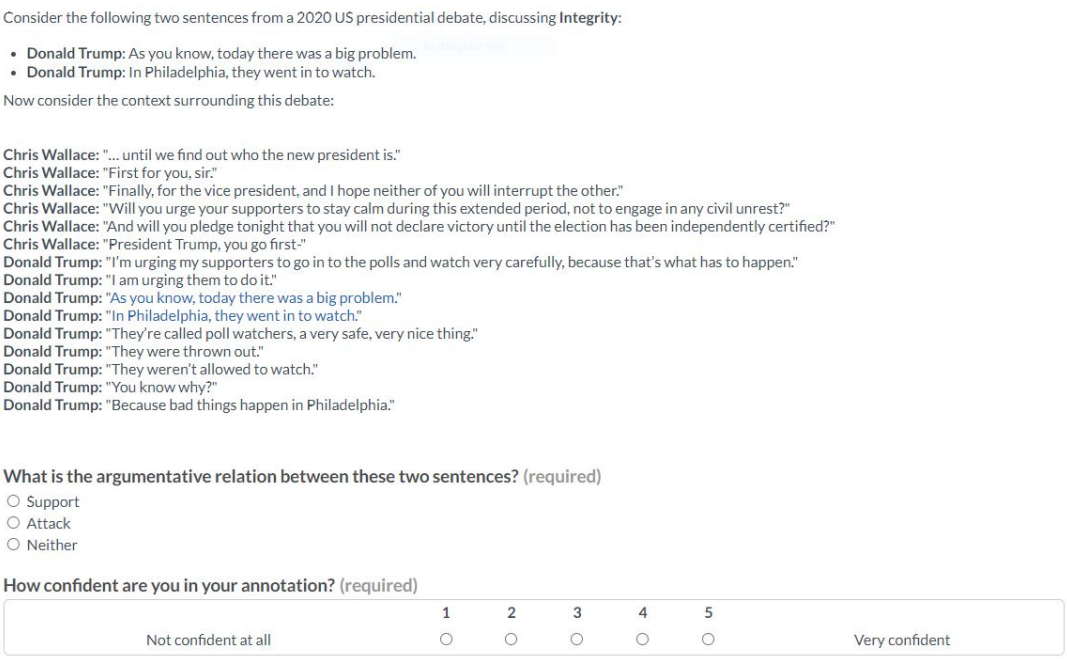


Figure 1: Example of the annotation task as seen by annotators.

- Donald Trump: "and that was a retweet."

The highlighted sentence "..." was a transcription error and the instructions stated to label these cases simply as 'neither'. One of the contributors, however, labelled it as 'support'. They explained in the free-text box that "[t]he ... assuming a pause is a silent statement to re affirm confidence in the first sentence".

In other cases, transcription labels confused some of the annotators:

- Donald Trump: "Many car companies came in from Germany, from Japan, went to Michigan, went to Ohio and they didn't come in with you."
- Donald Trump: **"[crosstalk 00:21:47]."**
- Joe Biden: [...] **"He talks about these great trade deals."**
- Joe Biden: "He talks about the art of the deal."

The contributors should have labelled this case as 'neither', since the first sentence is a transcription label. However, two of the contributors labelled it as attack, most likely because interrupting someone can be considered an attack towards the other speaker. In other instances, sentences that said something like 'I need to respond to that' or 'Please, let me respond' tended to be associated with attacks, whereas positive sentences like 'Good' or 'Yes', tended to be annotated as support.

## B The reasons behind disagreements

We then assessed the annotations and the reasons provided by the contributors in two cases that we thought to be of particular interest: i) when the contributors completely agree on "support" or "attack"; ii) when the contributors disagree at least once. Focusing on cases when the contributors completely agree on "neither" is not particularly useful, as they represent most of the dataset - and are generally because the sentences are clearly unrelated or not arguments at all ("thank you", "no", etc.). Focusing on the cases in which the contributors completely agree on support or attack and comparing them with the cases in which they are not so sure provided insight into which type of argumentative relations are clear enough to not induce confusions, and when confusion starts to set in.

Out of the 1145 cases in which the annotators did not disagree, only 20 of them were support relations and 19 of them attacks, which highlights the difficulty on agreeing on this type of relation. In general, we observed that the annotators agreed on support relations when the claims were very clearly stated. For instance:

- Chris Wallace: "Are you questioning the efficacy of masks?"
- Donald Trump: **"No, I think masks are okay."**
- Donald Trump: "You have to understand, if you look... I mean, I have a mask right here."
- Donald Trump: "I put a mask on when I think I need it."
- Donald Trump: **"Tonight, as an example, everybody's had a test and you've had social distancing and all of the things that you have to, but I wear masks-"**

Here, Trump claims very clearly "I think masks are okay" and he later explains using an argument by example, supporting his own claim. All annotators agreed that this was a support relation. One common source of confusion with

support statements was the cases in which one speaker was agreeing with the other. For instance:

- George Stephanopoulos: **"Was it a mistake to support it?"**
- Joe Biden: **"Yes, it was."**

Here, Biden is answering a question and the three contributors agreed that this relationship was of support because "its very clear that is supporting argument" and "Biden agrees with George". For the other contributor, "this is "support," because Joe is showing that he agrees with the first sentence." The fact that George Stephanopoulos precisely uses the word "support" might have fuelled this idea that this was a support relation, but there is no argument behind it, just an opinion. This was not isolated and happened in other examples too:

- Joe Biden: "You'll not hear me dividing."
- Joe Biden: "You'll hear me trying to unify, and bring people together."
- Joe Biden: **"When I said I was running because I wanted to unify the country, people said, "Well, there are the old days." We better be able to do it again"**
- Audience Member 11: **"Agreed."**

Using similar arguments, the contributors thought this was a support relation "[. . . ] because Audience Member 11 is showing that he/she agrees with the first sentence."

Natural dialogue is full of nuance and does not follow in general the argumentative structure found in argumentation frameworks. Whereas some cases are straightforward and clear for the annotators (see the first examples), those that require more thought, background knowledge, contextual information or mental rephrasing can be a real challenge and highly subjective. Nevertheless, sometimes the results are rather positive. Take, for instance, the following segment:

- Kamala Harris: "So, Susan, I'm glad you asked about transparency because it has to be across the board."
- Kamala Harris: "Joe has been incredibly transparent over many, many years."
- Kamala Harris: **"The one thing we all know about Joe, he puts it all out there."**
- Kamala Harris: **"He is honest, he is forthright, but Donald Trump on the other hand has been about covering up everything"**

Here, Harris is confident in her claim: Joe Biden puts it all out there. She uses the second statement to further prove her point by emphasising some of his attributes, while at the same time attacking Donald Trump. The contributors also pick up on this: "the argument was supporting the sentence but some how they attacking [sic] the other person." This type of example emphasises an added difficulty in this task, which is its highly contextual nature. This is one of the reasons why we decided to follow this relational approach instead of identifying first claims and premises and then linking them. Whereas in one context Harris' second claim might have been an attack if she was speaking about Trump's fitness for presidency, it was a support relation in this context to the claim that Biden puts it all out there and will be a good president. The following segment highlights this contextual nature:

- Joe Biden: "That's why I'm going to eliminate the Trump tax cuts."
- Joe Biden: **"And I'm going to eliminate those tax cuts."**
- Donald Trump: "That's okay."
- Joe Biden: "And make sure that we invest in the people who in fact need the help."
- Joe Biden: "People out there need help."
- Donald Trump: **"But why didn't you do it over the last 25 years?"**

When asked to identify the relation between these two sentences, all the annotators agreed this was an attack. This is not obvious as the second sentence is expressed as a question and needs some rephrasing in this context to understand that Trump implies "you're not going to eliminate the tax cuts because you had 25 years to do it and you didn't do it; you cannot be trusted". Indeed, one of the contributors recognises that this is not a simple question, but a "contentious query" and another that is "showing attack argumentative on this context".

Finally, there was one type of dialogue that tended to attract attack relations, which is when one person simply states the opposite to what the other person says. For instance:

- Donald Trump: **"He was thrown out dishonorably discharged"**
- Joe Biden: **"That's not true he was not dishonorably discharged."**

Or:

- Donald Trump: **"He doesn't have any law support."**
- Donald Trump: "He has no law enforcement."
- Joe Biden: **"That's not true."**

Or:

- Donald Trump: **"Once you became vice president he made a fortune in Ukraine, in China, in Moscow and various other places."**
- Joe Biden: **"That is not true."**

These are some cases that are particularly relevant when investigating how people understand arguments. All these pairs of sentences were labelled without disagreement as an argumentative "attack". Can simply negating what the other person just said be considered an argumentative attack? Following the logic of argumentation, if one statement p is valid, its negation ¬p cannot be valid following deductively the same set of premises. Therefore, we cannot use ¬p as an argument attacking p. However, the contributors consistently annotated these cases as attacks, which raises questions like: do people consider that saying "that is not true", without providing a valid set of premises, a valid argument? The reasons provided by the contributors did not clarify this issue, as they simply stated that they were disagreeing with one another, and that seemed to be sufficient to label it as "attack".