Argument Mining in Tweets: Comparing Crowd and Expert Annotations for Automated Claim and Evidence Detection

Neslihan Iskender[1(✉)], Robin Schaefer[2], Tim Polzehl[1,3], and Sebastian Möller[1,3]

[1] Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
{neslihan.iskender,sebastian.moeller}@tu-berlin.de
[2] Applied Computational Linguistics, University of Potsdam, Potsdam, Germany
robin.schaefer@uni-potsdam.de
[3] German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
tim.polzehl@dfki.de

Abstract. One of the main challenges in the development of argument mining tools is the availability of annotated data of adequate size and quality. However, generating data sets using experts is expensive from both organizational and financial perspectives, which is also the case for tools developed for identifying argumentative content in informal social media texts like tweets. As a solution, we propose using crowdsourcing as a fast, scalable, and cost-effective alternative to linguistic experts. To investigate the crowd workers' performance, we compare crowd and expert annotations of argumentative content, dividing it into claim and evidence, for 300 German tweet pairs from the domain of climate change. As being the first work comparing crowd and expert annotations for argument mining in tweets, we show that crowd workers can achieve similar results to experts when annotating claims; however, identifying evidence is a more challenging task both for naive crowds and experts. Further, we train supervised classification and sequence labeling models for claim and evidence detection, showing that crowdsourced data delivers promising results when comparing to experts.

Keywords: Argument mining · Crowdsourcing · Corpus annotation

1 Introduction

With the rapid development of social media sites, especially Twitter, have begun to serve as a primary media for argument and debate, leading to increasing interest in automatic argument mining tools [15]. However, they require considerable amounts of annotated data for the given topic to achieve acceptable performance, increasing the cost and organizational efforts of data set annotation by linguistic experts enormously [10]. As a result, crowdsourcing has become an attractive

E. Métais et al. (Eds.): NLDB 2021, LNCS 12801, pp. 275–288, 2021.
https://doi.org/10.1007/978-3-030-80599-9_25

alternative to expert annotation, helping researchers generate data sets quickly and in a cost-effective way [7]. Although some researchers have applied crowdsourcing to argument annotation [7,12,16], they did not focus on social media text which has character limitations and tends to be written informally without following specific rules for debate or opinion expression. So, focusing on social media increases the subjectivity and complexity of the argument annotation task [1,17]. Therefore, the appropriateness of crowdsourcing for it should be investigated.

This paper addresses this gap by conducting crowd and expert experiments on a German tweet data set[1], comparing annotations quantitatively, and investigating their performance for training argument mining tools. By placing a strong focus on the comparison of the crowd and expert annotations, we extend our previous study on tweet-based argument mining [13], which presents the first results for training performance of the expert annotations also used in this work. Like in our previous work, we apply a claim-evidence model, where *claim* is defined as a controversial opinion and *evidence* as a supportive statement related to a claim. Both components are further referred to as *Argumentative Discourse Units* (ADU) [11].

2 Related Work

Related work has investigated argument mining in tweets primarily from the viewpoint of corpus annotation and argument component detection. In an early work from 2016, the *Dataset of Arguments and their Relations on Twitter* (DART) was presented [2]. 4000 English tweets were annotated by three experts on the full tweet level for general argumentative content (stating high consistency as Krippendorff's α: 0.74 for inter-annotator agreement (IAA)), thereby refraining from further separating between claim and evidence. Also, topics were heterogeneous, including, for instance, tweets on product releases, which may contain different argumentation frequency, density and clarity. This may have facilitated individual annotation tasks. An applied logistic regression model yielded an F1 score of 0.78 on argument detection.

Another line of research approached argument mining on Twitter by focusing on evidence detection [1]. In contrast to our work, tweets were annotated for specific evidence types, e.g., news or expert opinion, and the annotators' level of expertise was not reported in the paper. Also, the full tweet was the unit of annotation, which reduced the task's complexity and might be reflected in their high Cohen's κ score of 0.79. An SVM classifier achieved an F1 score of 0.79 for the evidence detection task.

More recently, argument annotation work on Swedish social media was presented [9]. Annotators (one expert and seven "*trained annotators with linguistic backgrounds*") labeled argumentative spans in posts from discussion forums (Cohen's κ: 0.48). While this research did not focus on tweets, it still shows the difficulty of creating high-quality consistent argument annotations in social

[1] Corpus repository: https://github.com/RobinSchaefer/climate-tweet-corpus.

media data. Work on argument mining on data from various Greek social media sources, including tweets, was presented by [4]. The study included data annotation, however IAA was not presented, which hinders comparison. Moreover supervised classification and sequence labeling models were trained (F1: 0.77 and 0.42), which we adopt in our work.

As previous research on argument annotation of social media text reveals, the annotators were either experts [2,9], or their level of expertise was not reported or questioned [1,4]. Our research extends these studies by investigating the effect of annotator's expertise on the ADU annotation, focusing on claim detection in addition to general argument detection [2,4,9] and evidence detection [1] on the domain of highly controversial *climate change* tweets on Twitter.

3 Experiments

In our experiments, we used a data set with 300 German tweet pairs extracted from the Twitter API on the climate change debate. Each pair in the data set consists of a *context* tweet and a *reply* tweet as a response to the context tweet. The average word count of context tweets is 26.64, the shortest one with one word and the longest one with 49 words; the average word count of reply tweets is 27.44, the shortest one with one word, the longest one with 52 words.

3.1 Crowdsourcing Study

We collected crowd annotations using the Crowdee[2] Platform. We designed a task specific pre-qualification test for crowd worker selection. All crowd workers who passed Crowdee's German language test with a score of 0.9 or above were admitted for the pre-qualification test. In the pre-qualification test, we explained at first the general task characteristics and provided definitions and examples for the argumentative content dividing it into its two components *claim* and *evidence*. We defined *claim* as "the author's personal opinion, position or presumption" and *evidence* as "content intended to support a claim". In line with previous research, we decided on using relatively broad ADU definitions due to the rather informal nature of argumentation in tweets, which is hard to capture with more narrow definitions. Further, we provided text annotation guidelines such as only to annotate the smallest understandable part in a reply tweet as claim, only to annotate evidence if it relates to a claim from the tweets shown, and to ignore personal political beliefs, as well as the spelling or grammatical errors.

After reading the instructions, crowd workers were asked to annotate claim and evidence in tweet pairs. The first question "Is there any claim in the reply tweet?" was displayed with the two answer options "yes" and "no". The second question "Is there evidence in the reply tweet?" was displayed with the four answer options "yes, evidence in the reply tweet relates to a claim in the reply

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

tweet.", "yes, evidence in the reply tweet relates to a claim in the context tweet.", "yes, evidence in the reply tweet relates to a claim in both tweets.", and "no, there is no evidence.". We refer to these questions as voting questions in Sect. 4. If crowd workers selected an answer option with "yes" in any of the voting questions, they were asked to label the text part containing claim or evidence, which we refer to as text annotation in Sect. 4.

Each question was displayed on a separate page, and the pre-qualification task included the annotation of three different tweet pairs. Crowd workers could achieve a maximum of 12 points for answering each of the voting questions correctly, and we kept crowd workers exceeding 8 points. Additionally, the author's team evaluated manually crowd workers' answers for three text annotation questions and eliminated crowd workers who labeled the non-argumentative content in tweets as claim or evidence. Overall, 101 crowd workers participated in the pre-qualification test completing the task in 15 h with an average work duration of 546 s. Based on our selection criteria, 54 crowd workers were accepted for the main task.

Out of 54 admitted crowd workers, 42 crowd workers participated in the main task. Further, five unique crowd workers per tweet pair annotated claim and evidence using the same task design as in the pre-qualification test, resulting in 1500 crowd answers. We published a total of 1500 tasks in batches, and each batch was completed within a maximum of five days, with an average work duration of 394 s. Here, we observed that the main task's average task completion duration was lower than for the pre-qualification task, although the main task included the annotation of two more tweet pairs. The reason for this is probably the following: after doing the task a couple of times, crowd workers did not need to read the definitions and instructions at the beginning of the task, which led to a lower task completion duration.

3.2 Expert Evaluation

Two experts, one of them a Ph.D. student at a linguistics department and co-author of this paper, and the other one a student in linguistics, annotated the same 300 tweet pairs using the same task design as the crowdsourcing study. At first, they annotated the tweet pairs separately using the Crowdee platform. After the first separate evaluation round, the IAA scores, Cohen's κ, showed that the experts often diverted in their assessment. To reach consensus among experts, we arranged physical follow-up meetings with the two experts, which we refer to as mediation meetings. In these meetings, experts discussed the reasons and backgrounds of their annotations for tweet pairs in case of substantial disagreement and eventually aligned them if consensus was obtained. Eventually, acceptable IAA scores were reached for the voting questions of claim and evidence. This procedure also led to several suggestions regarding the refinement of annotation guidelines which will be discussed in Sect. 6.

4 Comparing Crowd with Expert

Results are presented for the two voting questions (claim and evidence) and the text annotations from the crowdsourcing and expert evaluation. We analyzed 1500 crowd answers using majority vote as the aggregation method for the voting questions, leading to 300 majority voted crowd answers and 600 expert answers for 300 tweet pairs. Further, we investigate the general annotation of argumentative content by combining claim and evidence annotation under the label *argument*.

4.1 Comparing Voting

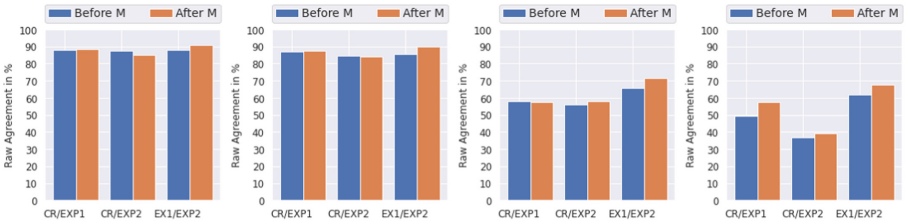
Before comparing expert votings for argument, claim and evidence with the crowd, we calculated Cohen's κ and Krippendorff's α scores to measure the IAA between two experts and the raw agreement scores in %. We analyzed both the voting with four answer options and binary evidence voting deducted from four answer options.

Table 1. Raw agreement in %, Cohen's κ and Krippendorff's α scores between two experts for argument, claim and evidence votings before mediation and after mediation

	Before mediation			After mediation		
	Agr. in %	κ	α	Agr. in %	κ	α
Argument	87.7	0.47	0.47	90.7	0.62	0.62
Claim	85.7	0.45	0.45	90	0.62	0.62
Evidence (binary)	65.7	0.34	0.31	71.7	0.44	0.43
Evidence (4 options)	61.7	0.32	0.31	67.7	0.41	0.41

Looking at Table 1, we see that the mediation meetings increase all of the agreement scores, and the Cohen's κ score for argument and claim reaches a substantial level (0.6–0.8] [6]. However, the mediation meetings increase the Cohen's κ scores for evidence only from fair (0.20–0.40] to moderate (0.40–0.60]. Also, we calculated Krippendorff's α, which is technically a measure of evaluator disagreement rather than agreement. Although the mediation meetings increase the Krippendorff's α scores, still they leave room for improvement ($\alpha < 0.667$) [5]. This result shows that identifying argumentative content, especially evidence, is even for experts a subjective and ambiguous task, which is also reflected by the raw agreement scores in % for evidence.

Next, we calculated raw agreement in % between crowd and experts, and between the two experts before and after mediation as shown in Fig. 1. Here, we observe that both before and after mediation, crowd workers reach comparable results as experts in terms of the raw agreement in %, achieving an agreement above 85 % for argument and claim. However, crowd-expert agreements for evidence is lower than the expert-agreement, especially when using the scale with four answer options. It shows that evidence identification by determining to



(a) Argument **(b)** Claim **(c)** Evidence (binary) **(d)** Evidence (4 options)

Fig. 1. Barplots of raw agreement in percentage for argument, claim, evidence (binary), and evidence (4 options) between crowd and experts, and between experts before and after mediation (M = Mediation, CR = Crowd, EXP = Expert)

which tweet evidence relates is a complex and subjective task, notably for crowd workers. Therefore, we use the results from the binary evidence votings in our further analysis.

To investigate the differences between crowd and expert for voting questions, we calculated the non-parametric T-Test, Mann-Whitney U Test. The test results revealed significant differences for argument and claim between crowd and experts both before and after mediation. The median values of crowd and experts clearly showed that the crowd workers identified arguments and claims in more tweets than the experts (argument: $N_{cr} = 282$, $N_{exp1} = 273$, $N_{exp2} = 255$; claim: $N_{cr} = 261$, $N_{exp1} = 255$, $N_{exp2} = 251$). Moreover, the Mann-Whitney U test results for evidence also revealed significant differences between crowd and expert 2 and between two experts both before and after mediation. Looking at the median values, we observed that expert 2 identified more evidence in tweets than expert 1 and crowd workers ($N_{cr} = 162$, $N_{exp1} = 166$, $N_{exp2} = 175$). The significant difference between the two experts for evidence is in line with our previous expert IAA analysis.

Analyzing the Spearman correlation coefficients between the crowd and expert, we saw that crowd-expert correlation for argument ($r_{cr/exp1} = 0.35, r_{cr/exp2} = 0.31, p < 0$) was at a weak level, where experts reached a moderate correlation before mediation ($r = 0.47, p < 0$). On the contrary, crowd correlation with expert 1 for claim ($r_{cr/exp1} = 0.42, r_{cr/exp2} = 0.31, p < 0$) achieved a similar level of correlation as the correlation between two experts ($r = 0.45, p < 0$). After expert mediation, the correlation between two experts increased to 0.62 both for argument and claim, while correlation between crowd and expert remained at the same level for argument ($r_{cr/exp1} = 0.36$, $r_{cr/exp2} = 0.25$, $p < 0$) and claim ($r_{cr/exp1} = 0.42$, $r_{cr/exp2} = 0.30$, $p < 0$). Note, the crowd and expert correlations for evidence were of an overall weak level regardless of the mediation (before mediation: $r_{cr/exp1} = 0.15$, $r_{cr/exp2} = 0.15$, $r_{exp1/exp2} = 0.36$, after mediation: $r_{cr/exp1} = 0.14$, $r_{cr/exp2} = 0.17$, $r_{exp1/exp2} = 0.46$, $p < 0$). These weak/moderate correlations before mediation demonstrate again the subjectivity of the task, especially for evidence.

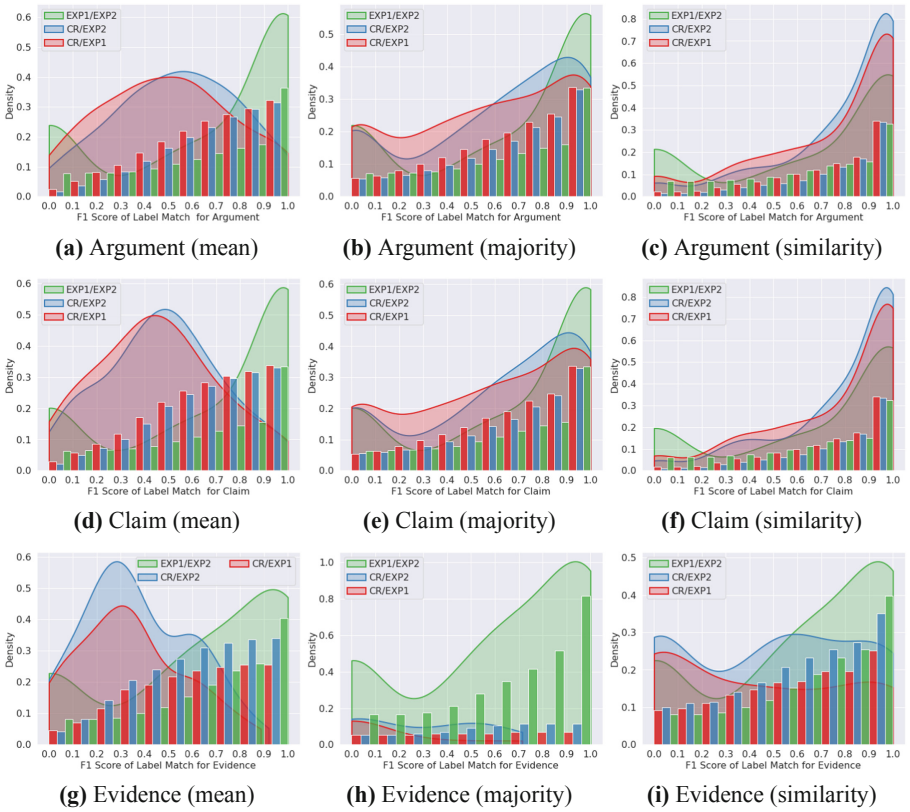
As our last analysis on the voting consistency, we calculated Fleiss' κ scores between crowd and two experts. Before mediation, they reached a Fleiss' κ score of 0.36 for argument and 0.37 for claim, which is also at a similar level of expert-agreement before mediation. After mediation, the Fleiss' κ score increased to 0.40 for argument and 0.43 for claim. This shows that mediation meetings contribute to the robustness of expert votings, indicating that a similar approach between crowd workers could increase the crowd votings' robustness as well. Similarly, the Fleiss' κ score for evidence increased from 0.20 to 0.24 after mediation, however, still remaining at a weak level.

4.2 Comparing Text Annotations

In this section, we compare the text annotations for claim and evidence given by crowd and experts. As explained in Sect. 4.1, the mediation meetings did not affect the relationship between crowd and expert votings remarkably, therefore we only focus on the annotations after mediation in this section. To compare the text annotations with each other, we follow a similar logic to ROUGE-1, which describes the overlap of unigrams (each word) between the system and reference summaries [8]. In our case, we compare the location of labeled text characters by crowd and expert, computing the precision ($Precision = \frac{\text{location of crowd labeled characters} \cap \text{location of expert labeled characters}}{\text{location of crowd labeled characters}}$) and recall ($Recall = \frac{\text{location of crowd labeled characters} \cap \text{location of expert labeled characters}}{\text{location of expert labeled characters}}$) to calculate the F1 score ($F1\ score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$).

We applied three different methods for comparing text annotations: *mean*, *majority vote* and *similarity*. In the first approach, we considered all five different crowd annotations for each tweet pair and computed the F1 score between each of five crowd workers and experts, calculating the mean of the five F1 scores as a final result. In our second approach, we followed a similar strategy to voting and calculated the majority vote for each annotated character location by comparing annotations from five different crowd workers. The resulting majority-voted character locations were used to calculate F1 scores between crowd and experts. In the last approach, we calculated F1 scores between each of five crowd workers for each tweet pair and selected the individual crowd worker whose text annotation has the highest average F1 score with other crowd workers. Then, we used this crowd worker's answer for calculating the F1 score between crowd and expert. It should be noted that we calculated the F1 scores only in case of positive claim or evidence voting from both naive and expert annotators.

Figure 2 shows the cumulative histograms and density estimation plots of argument, claim and evidence F1 scores for all three approaches. As the density plot for all text annotations between the two experts shows, the experts either do not agree on the text annotations or they agree 100 %. However, crowd's and experts' text annotations F1 score is distributed equally centered around the score 0.5 for argument and claim (see Fig. 2a and Fig. 2d) and around the score 0.3 for evidence (see Fig. 2g) using the mean approach. For the majority vote approach, we observe that argument and claim annotations get close to the score



(a) Argument (mean) **(b)** Argument (majority) **(c)** Argument (similarity)

(d) Claim (mean) **(e)** Claim (majority) **(f)** Claim (similarity)

(g) Evidence (mean) **(h)** Evidence (majority) **(i)** Evidence (similarity)

Fig. 2. Cumulative histograms and density estimation plots for the annotation match for argument (first row), claim (second row) and evidence (third row) between crowd and experts, and between two experts (CR = Crowd, EXP = Expert)

1, but still, its density is not at the level of the experts' F1 score (see Fig. 2b and Fig. 2e); and the crowd workers cannot agree on the text annotations for evidence (see Fig. 2h). As the Figs. 2c, 2f and 2i demonstrate, the similarity approach produce results most similar to experts' F1 score, especially for argument and claim. Therefore, we recommend using this approach when collecting data from multiple crowd workers.

5 Training Argument Mining Models on Annotated Tweets

In this section, we present experimental results from training supervised classification and sequence labeling models on full tweet and ADU annotations of

crowds and experts, respectively. As features BERT [3] embeddings were created by using deepset.ai's pretrained *bert-base-german-cased* model[3].

We compare different annotation sets (crowd vs expert) and layers (argument vs claim vs evidence). Models are trained both on individual expert and crowd annotations and on combinations of these. Models are tested either with test sets obtained from a train-test split (Tables 2 and 4) or by using expert annotations as gold standard (Tables 3 and 5). As shown in Sect. 4.1, all argument classes form the respective majority class, which is why we report weighted F1 scores. For comparison we also show unweighted macro scores in Tables 2 and 4. All scores are 10-fold cross-validated.

5.1 Supervised Classification

We trained supervised classification models on full tweet annotations derived from the ADU annotations (voting questions in experiments). Thus, a classifier's task is to separate tweets containing an ADU from non-argumentative tweets. Results (Tables 2 and 3) are obtained using eXtreme Gradient Boosting. Models trained on non-mediated expert annotations mostly yield promising weighted F1 scores (0.71–0.91). Unweighted F1 scores are comparatively low. This indicates the models' problems with identifying minority classes, which is intensified by the small corpus size. Notably, the reduction appears especially to be caused by low recalls. Models trained on mediated expert data show less variance between annotators. Also, training on combined expert annotation sets yields substantially better results than training on individual expert annotation sets.

Results obtained by crowd annotations show an interesting pattern. While models trained on all crowd annotations can generally compete with expert models, weighted F1 scores derived from crowd majority annotations are reduced (F1: 0.57/0.58) with the exception of evidence targets. For argument and claim targets the difference between weighted and unweighted F1 scores is less severe than for expert annotations. Also, utilizing combined crowd and expert annotations yields acceptable results. Testing models trained on mediated expert data with gold annotations (see Table 3) yields mainly similar results to the scores shown in Table 2. However, testing all crowd annotation sets with expert annotations does not perform well. Adding expert annotations to the training set notably improves results with the exception of evidence annotations.

5.2 Sequence Labeling

Sequence labeling models were trained on the ADU annotations in order to build a system that can extract argumentative spans from tweets (text annotations from crowd and experts). We applied Conditional Random Fields for this task. Here, we use the similarity method instead of majority for deriving a single set from the crowd annotations, as this showed best results during text annotation

[3] https://huggingface.co/bert-base-german-cased.

Table 2. Supervised classification results (M = mediation; CS = corpus size; p = partial (i.e. only experts are mediated); w = weighted).

Annotator	M	CS	Argument				Claim				Evidence			
			F1 (w)	F1	P	R	F1 (w)	F1	P	R	F1 (w)	F1	P	R
Expert 1	–	300	0.84	0.60	0.68	0.59	0.81	0.57	0.66	0.57	0.57	0.54	0.56	0.55
Expert 2	–	300	**0.91**	0.77	0.93	0.72	0.86	0.71	0.85	0.67	**0.71**	0.70	0.73	0.70
Expert (both)	–	600	0.90	0.77	0.79	0.78	**0.88**	0.78	0.78	0.78	0.69	0.69	0.71	0.69
Expert 1	+	300	0.87	0.66	0.79	0.64	0.84	0.63	0.72	0.61	0.62	0.60	0.63	0.61
Expert 2	+	300	0.90	0.76	0.91	0.74	0.87	0.72	0.87	0.69	0.69	0.68	0.70	0.68
Expert (both)	+	600	**0.95**	0.89	0.93	0.87	**0.93**	0.86	0.90	0.84	**0.75**	**0.75**	0.77	0.75
Crowd (majority)	–	300	0.57	0.53	0.55	0.53	0.58	0.53	0.54	0.54	**0.81**	0.46	0.43	0.50
Crowd (all)	–	1,500	**0.87**	0.86	0.87	0.86	**0.81**	0.79	0.80	0.79	0.78	0.61	0.65	0.60
Crowd + Expert	p	2,100	0.80	0.80	0.81	0.80	0.78	0.78	0.79	0.78	0.76	0.69	0.73	0.67

Table 3. Supervised classification results, tested with gold annotations (Expert 1 or Expert 2). Expert annotations are mediated. Only weighted F1 scores are reported.

Annotator	Argument		Claim		Evidence	
	Expert 1	Expert 2	Expert 1	Expert 2	Expert 1	Expert 2
Expert 1	–	0.86	–	0.83	–	0.61
Expert 2	0.90	–	0.88	–	0.59	–
Expert (both)	0.89	0.88	0.86	0.87	0.61	0.66
Crowd (majority)	0.47	0.49	0.48	0.47	0.55	0.38
Crowd (all)	0.21	0.21	0.11	0.14	0.43	0.28
Crowd + Expert	0.81	0.84	0.74	0.72	0.49	0.33

analysis (see Sect. 4.2). Looking at Table 4, models trained on non-mediated data yields promising results for argument (0.83) and evidence detection (0.70). Weighted F1 scores for claim detection are comparatively low. However, training on both expert sets results in a notable improvement on this task. Compared to classification, unweighted precision and recall show less divergence. Training sequence labeling models on mediated expert data hardly changes results. However, improvements are achieved by utilizing both expert annotation sets.

Using all crowd annotations results in reduced scores for argument labels, and comparable results for claim and evidence labels in comparison to experts. Combining crowd and expert annotations improves the results. Testing models with gold annotations (see Table 5) shows patterns similar to previously discussed results. Importantly, crowd similarity annotations yield results comparable to expert annotations or better when tested with gold annotations.

Table 4. Sequence labeling results. (M = mediation; CS = corpus size, p = partial (i.e. only experts are mediated); w = weighted).

Annotator	M	CS	Argument				Claim				Evidence			
			F1(w)	F1	P	R	F1(w)	F1	P	R	F1(w)	F1	P	R
Expert 1	–	300	0.72	0.62	0.62	0.62	0.57	0.58	0.59	0.58	**0.70**	0.60	0.60	0.61
Expert 2	–	300	**0.83**	0.68	0.69	0.68	0.57	0.60	0.61	0.60	0.62	0.62	0.63	0.62
Expert (both)	–	600	0.80	0.70	0.72	0.71	**0.65**	0.67	0.69	0.66	**0.70**	0.67	0.71	0.68
Expert 1	+	300	0.72	0.61	0.61	0.62	0.57	0.59	0.60	0.59	0.71	0.61	0.61	0.61
Expert 2	+	300	0.81	0.67	0.67	0.68	0.57	0.60	0.61	0.60	0.62	0.61	0.62	0.61
Expert (both)	+	300	**0.86**	0.78	0.80	0.78	**0.69**	0.71	0.73	0.70	**0.76**	0.72	0.75	0.72
Crowd (similarity)	–	300	0.53	0.54	0.55	0.53	**0.55**	0.56	0.58	0.55	**0.81**	0.64	0.64	0.64
Crowd (all)	–	1500	**0.64**	0.60	0.64	0.58	0.54	0.59	0.63	0.57	0.64	0.61	0.65	0.59
Crowd + Expert	p	2100	0.72	0.65	0.71	0.62	0.59	0.63	0.67	0.61	0.68	0.63	0.69	0.60

Table 5. Sequence labeling results, tested with gold annotations (Expert 1 or Expert 2). Expert annotations are mediated. Only weighted F1 scores are reported.

Annotator	Argument		Claim		Evidence	
	Expert 1	Expert 2	Expert 1	Expert 2	Expert 1	Expert 2
Expert 1	–	0.77	–	0.57	–	0.62
Expert 2	0.74	–	0.56	–	0.66	–
Expert (both)	0.74	0.79	0.58	0.59	0.67	0.63
Crowd (similarity)	0.73	0.73	0.66	0.65	0.75	0.62
Crowd (all)	0.75	0.83	0.57	0.61	0.74	0.64
Crowd + Expert	0.75	0.80	0.57	0.61	0.74	0.64

6 Discussion and Outlook

Our extensive empirical comparison of crowd and expert ADU annotations in Sect. 4 showed that this task has a high level of subjectivity and ambiguity, even for experts. Even after mediation, experts only reached moderate IAA scores for evidence, indicating that distinguishing between claim and evidence is even harder than claim identification. We observed similar results when comparing crowd and expert annotations, where crowd workers could reach a comparable level of raw agreement in % as experts for argument and claim, while crowd-expert agreement for evidence remained at moderate level for both expert and crowd assessment. Also, the results from Sect. 4.2 confirmed this finding. Here, we also demonstrated a method for determining the "reliable" crowd worker for text annotation who can achieve similar results as experts.

Despite the annotation differences, the results from Sect. 5.1 showed that training with all crowd annotations delivers similar results as experts. However, when using gold annotations for testing classification models, the crowd could not achieve comparable results to experts. For sequence labeling (see Sect. 5.2), training with crowd annotations produced close results to single experts for claim and evidence, but combining both experts led to better results than for

the crowd. Also, when using an expert data set as the test set for sequence labeling, crowd text annotations achieved expert-level F1 scores. As results of models trained on crowd worker annotations derived by the similarity method and on expert annotations are comparable when tested with gold annotations, we argue that the similarity annotations are reliable.

The reasons for different annotations between crowd and experts, especially for evidence, may be due to the text structure of tweets, which are characterized by a certain degree of implicitness, thereby entailing substantial subjectivity for the annotation task. Further, subjectivity also complicates the decision on the exact boundary between claim and evidence units. As evidence is defined as occurring only in relation to a claim, determining claim-evidence boundaries is of particular importance. So, one may consider separating evidence annotation from claim annotation. Annotating claims in a first step, followed by subsequent evidence labeling, would reduce annotators' degrees of freedom and thereby possibly increase the IAA. Limiting the allowed number of ADU annotations per tweet could positively affect IAA scores as fewer boundaries between claim and evidence have to be drawn.

In future work, we suggest adjustments to the definitions of argument components based on the results from expert mediation sessions. Given the peculiarities of tweets, we consider it appropriate to utilize a relatively broad interpretation, especially of the concept *claim*. Still, it may be fruitful to define more narrow claim and evidence definitions resulted from expert mediation sessions. For example, one could focus on *major claims* [14], which could be defined as a tweet's single main position or opinion, i.e., the argumentative reason why it was created. This may decrease the task subjectivity. Additionally, evidence might relate to a tweet outside the presented tweet pairs, so showing more than one context tweet may help the evidence annotation process. Another helpful approach may be arranging mediation sessions between crowd workers since the mediation between experts increased their agreement.

Despite the limitations, this paper makes an important contribution to human annotation research of argument mining in tweets. The organizational efforts and the cost of expert annotation at scale can be enormous, which is a great challenge in a fast-moving field like argument mining. Therefore, finding reliable ways of using crowdsourcing can be a promising solution, and we hope to see more research in this field.

References

1. Addawood, A., Bashir, M.: What is your evidence? A study of controversial topics on social media. In: Proceedings of the Third Workshop on Argument Mining (ArgMining2016), August 2016, pp. 1–11. Association for Computational Linguistics, Berlin, Germany (2016). https://doi.org/10.18653/v1/W16-2801
2. Bosc, T., Cabrio, E., Villata, S.: DART: a dataset of arguments and their relations on Twitter. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 2016, pp. 1258–1263. European Language Resources Association (ELRA), Portorož, Slovenia (2016)

3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), June 2019, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423
4. Goudas, T., Louizos, C., Petasis, G., Karkaletsis, V.: Argument extraction from news, blogs, and social media. In: Likas, A., Blekas, K., Kalles, D. (eds.) Artificial Intelligence: Methods and Applications, pp. 287–299. Springer International Publishing, Cham (2014)
5. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology, Sage publications, Thousand Oaks (1980)
6. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
7. Lavee, T., et al.: Crowd-sourcing annotation of complex NLU tasks: a case study of argumentative content annotation. In: Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP, November 2019, pp. 29–38. Association for Computational Linguistics, Hong Kong (2019). https://doi.org/10.18653/v1/D19-5905
8. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries, pp. 74–81 (July 2004)
9. Lindahl, A.: Annotating argumentation in Swedish social media. In: Proceedings of the 7th Workshop on Argument Mining, December 2020, pp. 100–105. Association for Computational Linguistics, Online (2020)
10. Miller, T., Sukhareva, M., Gurevych, I.: A streamlined method for sourcing discourse-level argumentation annotations from the crowd. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), June 2019, pp. 1790–1796. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1177
11. Peldszus, A., Stede, M.: From argument diagrams to argumentation mining in texts: a survey. Int. J. Cogn. Inform. Nat. Intell. **7**(1), 1–31 (2013). https://doi.org/10.4018/jcini.2013010101
12. Reisert, P., Vallejo, G., Inoue, N., Gurevych, I., Inui, K.: An annotation protocol for collecting user-generated counter-arguments using crowdsourcing. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) Artificial Intelligence in Education, pp. 232–236. Springer International Publishing, Cham (2019)
13. Schaefer, R., Stede, M.: Annotation and detection of arguments in tweets. In: Proceedings of the 7th Workshop on Argument Mining, December 2020, pp. 53–58. Association for Computational Linguistics, Online (2020)
14. Stab, C., Gurevych, I.: Identifying argumentative discourse structures in persuasive essays. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), October 2014, pp. 46–56. Association for Computational Linguistics, Doha, Qatar (2014). https://doi.org/10.3115/v1/D14-1006
15. Stede, M., Schneider, J.: Argumentation Mining, Synthesis Lectures in Human Language Technology, vol. 40. Morgan & Claypool (2018)

16. Toledo-Ronen, O., Orbach, M., Bilu, Y., Spector, A., Slonim, N.: Multilingual argument mining: Datasets and analysis. In: Findings of the Association for Computational Linguistics: EMNLP 2020, November 2020, pp. 303–317. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.29
17. Šnajder, J.: Social media argumentation mining: The quest for deliberateness in raucousness (2016)