Is Your Perspective Also My Perspective? Enriching Prediction with Subjectivity

Julia Romberg
Department of Social Sciences
Heinrich Heine University Düsseldorf, Germany
julia.romberg@hhu.de

Abstract

Although argumentation can be highly subjective, the common practice with supervised machine learning is to construct and learn from an aggregated ground truth formed from individual judgments by majority voting, averaging, or adjudication. This approach leads to a neglect of individual, but potentially important perspectives and in many cases cannot do justice to the subjective character of the tasks. One solution to this shortcoming are multi-perspective approaches, which have received very little attention in the field of argument mining so far.

In this work we present *PerspectifyMe*, a method to incorporate perspectivism by enriching a task with subjectivity information from the data annotation process. We exemplify our approach with the use case of classifying argument concreteness, and provide first promising results for the recently published CIMT PartEval Argument Concreteness Corpus.

1 Introduction

The analysis of arguments and especially their properties is challenging and often subjective, which renders the creation of suitable language resources for argument mining difficult (Stab and Gurevych, 2014; Lindahl et al., 2019). Uniform annotation often requires intensive training, and this costly approach has been shown to regularly result in at most moderate agreement among annotators (Aharoni et al., 2014; Rinott et al., 2015; Habernal and Gurevych, 2017; Shnarch et al., 2018). Alternative approaches such as crowd-sourcing share this problem, especially for demanding tasks like argument quality (Toledo et al., 2019).

Although the lack of consensus might clearly indicate that the annotation task is either ambiguous (Artstein and Poesio, 2008), too complex (Aroyo and Welty, 2015), or influenced by variables such as demographics and individual bias (Sap et al., 2022; Biester et al., 2022), the established procedure is to aggregate the individual judgments into a single ground truth at the end of the annotation process (by majority vote, averaging, or adjudication).

Learning from aggregated ground truth has several drawbacks. Minority voices are ignored, however valuable they may be, and only those in line with the mainstream are heeded (Noble, 2012). This rises also a fairness concern, as certain sociodemographic groups and their perspectives may be underrepresented (Prabhakaran et al., 2021). Finally, it is questionable whether the assumption of a single truth, i.e., that there is only one correct label for an example, holds at all for subjective tasks (Ovesdotter Alm, 2011; Aroyo and Welty, 2015).

Therefore, the question of multi-perspective approaches arises (Abercrombie et al., 2022). Basile et al. (2021) introduced the paradigm of *data perspectivism* in order to “integrate the opinions and perspectives of the human subjects involved in the knowledge representation step of ML processes”. One example for perspectivist data is argumentation (Hautli-Janisz et al., 2022; Romberg et al., 2022b).

However, many popular algorithms require a single ground truth to which the model can adapt. In this paper, (i) we thus introduce a method that combines collaborative and subjective viewpoints by complementing an aggregated label with a subjectivity score. More specifically, *PerspectifyMe* proposes to add the prediction of how perspectivist an input is as an additional sub-task. Providing this information can for example help a human decide when to rely on their own perspective. (ii) To exemplify our approach, we draw on a recently published perspectivist dataset for argument concreteness in public participation processes (Romberg et al., 2022b). We provide several baselines based on our proposed method for this subjective task. While these are certainly extendable, they already show promising results for automatic classification by concreteness. (iii) To the best of our knowledge, we are the first to automatically classify arguments

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 115–125
October 17, 2022.

in an explicitely perspectivist manner.

2 Related Work

Basile et al. (2021) provide a nice summarization of the previous work towards perspectivist machine learning, dividing the field in two groups.

The first aims at building unified ground labels that involve perspectivism by either only keeping instances on which a statistically significant majority agrees (Cabitza et al., 2020), by computing a weighting according to annotator reliability (Heinecke and Reyzin, 2019; Cabitza et al., 2020; Hovy et al., 2013), by replicating or weighting instances using provided labels or disagreement measures (Plank et al., 2014; Akhtar et al., 2019), or by participatory consensus building (Chang et al., 2017; Schaekermann et al., 2018).

The second group incorporates the perspectivism into the core machine learning workflow by either training an ensemble of models that rely on different ground truths (Akhtar et al., 2020; Campagner et al., 2021), by soft loss learning (Plank et al., 2014; Uma et al., 2020; Campagner et al., 2021), or by utilizing multi-task learning (Cohn and Specia, 2013; Guan et al., 2018; Sudre et al., 2019; Fornaciari et al., 2021; Davani et al., 2022).

Our approach ties into the latter idea by transforming the original problem into multiple subtasks. However, multi-task learning approaches for multi-perspectivist tasks have primarily aimed at improving model performance. To do so, the aggregated ground truth is learned along with the distribution of individual labels. Instead, we focus on outputting an indication of how perspectivist the model predictions are (namely, by adding a subjectivity score) to help interpret the results.

The only previous studies that specifically address argument mining are, to the best of our knowledge, two recently published non-aggregated datasets: QT30nonaggr (Hautli-Janisz et al., 2022) and the CIMT PartEval Argument Concreteness Corpus (Romberg et al., 2022b).

3 Use Case: Argument Concreteness in Public Participation

Public participation is a means regularly used by democratic authorities to involve citizens in policy-making processes (Dryzek et al., 2019). The manual evaluation workflow often includes reading the contributions, detecting duplicates, identifying arguments and opinions, and thematically clustering content before drawing conclusions from the input (Romberg and Escher, 2022).

One solution to reduce the workload of human evaluators is machine learning (OECD, 2003). Although there is a general consensus that such important democratic processes cannot be fully automated, automating sub-tasks such as topic classification or argument detection and analysis can support the evaluation.

Argument Mining for public participation has received considerable attention (Kwon et al., 2007; Liebeck et al., 2016; Lawrence et al., 2017; Park and Cardie, 2018; Romberg and Conrad, 2021). While works such as Park and Cardie (2014) and Niculae et al. (2017) have already addressed the evidence and verifiability of propositions, there has been no attempt to automatically classify their concreteness. Predicting the concreteness of propositions can assist a human analyst to speed up the evaluation by ranking them, since less concrete ideas tend to be more laborious to evaluate (Romberg et al., 2022b).

The CIMT PartEval Argument Concreteness Corpus (Romberg et al., 2022a) provides argumentative text units (ATU) in German extracted from mobility-related public participation processes. Each ATU consists of one to several sentences, consecutive in the original document, and a tag that describes the argumentative function (*major positions*: proposed courses of action and policy options or *premises*: attacking/supporting reasons). In total, the dataset contains $1,127$ ATUs, 614 of which are major positions and 513 are premises.

These ATUs have been categorised into three different degrees of concreteness:

- ATUs of **high concreteness** contain comprehensive details that describe the “what”, “how”, and “where”.
- ATUs of **intermediate concreteness** contain only partial specification of the “what”, “how” and “where”. There is room for interpretation in inferring specific actions (major positions) or in evaluating the actual reasons (premises).
- ATUs of **low concreteness** contain no detailed information of the “what”, “how” and “where”. A variety of measures could be derived and reasons remain vague.

Table 1 illustrates the three types to provide a better understanding of the dataset. Example A is a major position unit of high concreteness: it is clear what action is desired (protective cycle lanes next

Ex.	Unit text	Unit type	Concreteness
A	If the parking spaces along Friedrich-Breuer-Straße were removed, there would be enough space for protective cycle lanes next to the rails.	major position	high
B	The connection to the centre of Beuel through Obere Wilhelmsstraße is also not very pleasant to drive.	premise	intermediate
C	Rules for cycle paths	major position	low

Table 1: Examples of argumentative text units with argument types and concreteness ratings from the CIMT PartEval Argument Concreteness Corpus. To assist readers understand the content, the texts have been translated into English. (The examples presented here are cases in which the annotators were in complete agreement on the coding of concreteness.)

to the rails), where it is to be implemented (along Friedrich-Breuer-Straße) and how (free space by parking space removal). The premise unit in example B is of intermediate concreteness: it is clear, what the issue is and where (connection through Obere Wilhelmsstraße not very pleasant to drive). However, it remains unclear what makes driving through unpleasant. Example C shows a major position unit of low concreteness: the claim is very general and does not refer to specific locations, nor is it more specific about what rules are required.

The annotation of the data was performed by five coders. While finalizing the annotation guidelines, the coders annotated a selection of contributions, and inconsistencies were discussed in a group with the coders and two process supervisors. The guidelines were adjusted and the coders trained to the point where it became apparent that the divergent annotations were different perspectives rather than incorrect coding: In the discussion, the different coders were able to argue convincingly for their stance. Krippendorff's α_w (Krippendorff, 2013) of 0.46 confirms that the codings, although subjective, are not arbitrary.

4 PerspectifyMe

Previous work has incorporated perspectivism through distributions over individual labels. However, such distributions may be of limited use when provided to a human as a direct output, e.g. in human-machine interactions. In particular, providing such a diversity of perspectives that might apply (from the annotators' point of view - not necessarily from the point of view of the particular user) can be too complex and potentially confusing.

For items that trigger a subjective perception, it might make more sense (e.g., in a use case like ours) to inform the user about this and let them decide whether to make their own assessment or to go along with the collaborative opinion.

Therefore, we propose to enrich model predictions for subjective supervised machine learning tasks with the provision of a subjectivity score.

Task	Label	Support
Sub-Task $\mathcal{T}_H$: Concreteness	High	709 (62.9%)
	Intermediate	336 (29.8%)
	Low	82 (7.3%)
Sub-Task $\mathcal{T}_S$: Subjectivity	Objective	478 (42.4%)
	Rather objective	244 (21.7%)
	Rather subjective	275 (24.4%)
	Subjective	130 (11.5%)

Table 2: Overview of the label distributions for the tasks.

4.1 General Description

Given a task $\mathcal{T}$, we assume that there are both objective and subjective items in a corresponding dataset. This means that part of the dataset is annotated in a very consistent way, while the rest has elicited different views among coders. Our goal is then to predict a so-called hard label (aggregated by some method), and jointly inform on items for which there might be multiple correct outputs, depending on the perspective. We thus propose *PerspectifyMe*, a method to introduce perspectivism into the machine learning workflow by translating $\mathcal{T}$ into two sub-tasks $\mathcal{T}_\mathcal{H}$ and $\mathcal{T}_\mathcal{S}$. $\mathcal{T}_\mathcal{H}$ refers to the original prediction task using hard-labels as ground truth. $\mathcal{T}_\mathcal{S}$ refers to an artificial task of predicting the subjectivity of the input using a subjectivity score.

4.2 Application to Our Use Case

The perspectivity of judging argument concreteness is reflected in the CIMT PartEval Argument Concreteness Corpus through five single annotations. Following the previously introduced method, we conducted two transformation steps to yield the target variables for $\mathcal{T}_\mathcal{H}$ and $\mathcal{T}_\mathcal{S}$.

Concreteness Score We first built an aggregated ground truth by calculating the average concreteness per unit. For this, we mapped the categorical labels to numerical values (high: 3, intermediate: 2, low: 1) and averaged them. To retain the origi-

		Concreteness		Subjectivity (4-class)		Subjectivity (2-class)	
		Macro-F$_1$	Accuracy	Macro-F$_1$	Accuracy	Macro-F$_1$	Accuracy
joint	Majority Baseline	0.26	0.63	0.15	0.42	0.39	0.64
	LR (length)	0.54 ± 0.06	0.74 ± 0.03	0.30 ± 0.02	$\mathbf{0.52} \pm 0.03$	0.68 ± 0.03	0.72 ± 0.02
	LR (bow)	0.53 ± 0.04	0.75 ± 0.02	0.33 ± 0.05	0.50 ± 0.03	0.69 ± 0.03	0.71 ± 0.03
	LR (length+bow)	0.54 ± 0.04	0.74 ± 0.03	0.34 ± 0.05	0.50 ± 0.04	0.69 ± 0.03	0.72 ± 0.03
	SVM (length)	0.59 ± 0.04	0.71 ± 0.02	0.34 ± 0.03	0.48 ± 0.03	0.70 ± 0.02	0.72 ± 0.02
	SVM (bow)	0.59 ± 0.04	0.74 ± 0.03	0.37 ± 0.05	0.49 ± 0.04	0.69 ± 0.02	0.71 ± 0.03
	SVM (length+bow)	0.62 ± 0.05	0.75 ± 0.03	0.37 ± 0.03	0.50 ± 0.03	0.70 ± 0.03	0.72 ± 0.02
	BERT	$\mathbf{0.67} \pm 0.05$	$\mathbf{0.79} \pm 0.02$	$\mathbf{0.42} \pm 0.04$	$\mathbf{0.52} \pm 0.03$	$\mathbf{0.72} \pm 0.02$	$\mathbf{0.74} \pm 0.02$
major position	Majority Baseline	0.25	0.60	0.14	0.40	0.39	0.64
	LR (length)	0.49 ± 0.06	0.70 ± 0.04	0.27 ± 0.04	0.46 ± 0.04	0.59 ± 0.11	0.68 ± 0.04
	LR (bow)	0.52 ± 0.06	0.69 ± 0.03	0.28 ± 0.06	0.42 ± 0.04	0.60 ± 0.10	0.67 ± 0.04
	LR (length+bow)	0.52 ± 0.06	0.69 ± 0.04	0.31 ± 0.06	0.44 ± 0.04	0.63 ± 0.10	0.68 ± 0.05
	SVM (length)	0.56 ± 0.04	0.69 ± 0.04	0.33 ± 0.04	0.44 ± 0.04	0.64 ± 0.05	0.67 ± 0.04
	SVM (bow)	0.53 ± 0.07	0.67 ± 0.04	0.28 ± 0.08	0.42 ± 0.04	0.63 ± 0.09	0.67 ± 0.06
	SVM (length+bow)	0.55 ± 0.06	0.70 ± 0.04	0.33 ± 0.06	0.44 ± 0.04	0.64 ± 0.06	0.68 ± 0.04
	BERT	$\mathbf{0.62} \pm 0.07$	$\mathbf{0.76} \pm 0.04$	$\mathbf{0.37} \pm 0.06$	$\mathbf{0.47} \pm 0.05$	$\mathbf{0.68} \pm 0.06$	$\mathbf{0.71} \pm 0.05$
premise	Majority Baseline	0.26	0.65	0.15	0.44	0.39	0.64
	LR (length)	0.57 ± 0.07	0.80 ± 0.02	0.32 ± 0.02	$\mathbf{0.56} \pm 0.04$	$\mathbf{0.73} \pm 0.05$	0.75 ± 0.04
	LR (bow)	0.52 ± 0.06	0.69 ± 0.03	0.34 ± 0.05	0.54 ± 0.05	0.71 ± 0.03	0.74 ± 0.03
	LR (length+bow)	0.61 ± 0.08	0.80 ± 0.03	0.35 ± 0.04	0.55 ± 0.04	0.72 ± 0.04	0.74 ± 0.04
	SVM (length)	0.60 ± 0.05	0.75 ± 0.03	0.33 ± 0.04	0.48 ± 0.05	0.72 ± 0.04	0.74 ± 0.04
	SVM (bow)	0.67 ± 0.05	0.79 ± 0.03	0.36 ± 0.05	0.53 ± 0.05	0.72 ± 0.04	0.74 ± 0.04
	SVM (length+bow)	$\mathbf{0.68} \pm 0.07$	0.81 ± 0.03	0.38 ± 0.07	0.53 ± 0.07	0.71 ± 0.04	0.74 ± 0.04
	BERT	$\mathbf{0.68} \pm 0.06$	$\mathbf{0.82} \pm 0.03$	$\mathbf{0.42} \pm 0.05$	$\mathbf{0.56} \pm 0.04$	$\mathbf{0.73} \pm 0.04$	$\mathbf{0.76} \pm 0.04$

Table 3: Excerpt from the results for the classification of ATUs according to concreteness and subjectivity.

nal concreteness scale, the rounded average scores were remapped to the original categories.

Subjectivity Score For each unit, we calculated the pairwise L1 distance of the numerical labels and summed them up to calculate an overall distance. We translated the resulting distances into a four-category and a two-category scheme of subjectivity (for more details see Appendix A.1).

Table 2 provides an overview of the resulting sub-tasks. While highly concrete ATUs predominate, low concreteness is rare. Over sixty percent of the units elicited a fairly objective perception, a large proportion of which were even coded in a completely consistent manner. At the same time, there is a notable proportion of perspectivist ATUs.

5 Experiments

5.1 Classification Baselines

We evaluate several classification baselines: The traditional approaches logistic regression (LR), support vector machines (SVM), and random forests (RF) were combined with text length (in tokens) and a bag-of-words as features. The language model BERT was initialized with a case-sensitive base model for German (110M parameters) [1]. We fitted separate classifiers for the two sub-tasks.

5.2 Experimental Setup

We evaluated model performance on the dataset with and without respect to the types of arguments (major position/premise vs. joint) to see whether there are differences in predicting concreteness and subjectivity. To obtain reliable results, we used a repeated 5-fold cross-validation setup (Krstajic et al., 2014) (10 repetitions) and kept 10% for validation (i.e. splitting the dataset each time in $70/10/20$ for train/val/test). The hyperparameters were tuned with a grid search in each fold (an overview of the search space is given in Appendix A.2). F$_1$ and accuracy are the evaluation scores.[2]

5.3 Results

Table 3 shows a selection of the results for the classification of ATUs. A complete overview, including class scores, can be found in Appendix A.3.

When predicting degrees of concreteness, BERT achieved the best results (F$_1$ as well as accuracy). Looking at the other models, it turned out that simple length was already a good indicator for concreteness. When analyzing correlation effects with Spearman's rank correlation coefficient this finding was supported by a strong correlation of the target variables with the text length (concreteness: $\rho = 0.657$, subjectivity: $\rho = -0.525$). Adding

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

[2] Code available at github.com/juliaromberg/ArgMining2022

		rather objective	rather subjective
Macro-F$_1$	LR (length)	0.50 ± 0.08	0.45 ± 0.06
	LR (bow)	0.49 ± 0.05	0.44 ± 0.05
	LR (length+bow)	0.51 ± 0.07	0.45 ± 0.05
	SVM (length)	0.64 ± 0.06	0.46 ± 0.05
	SVM (bow)	0.61 ± 0.06	0.47 ± 0.05
	SVM (length+bow)	0.64 ± 0.07	0.49 ± 0.07
	BERT	0.70 ± 0.06	0.51 ± 0.07
Accuracy	LR (length)	0.80 ± 0.03	0.62 ± 0.05
	LR (bow)	0.82 ± 0.03	0.62 ± 0.05
	LR (length+bow)	0.81 ± 0.03	0.62 ± 0.05
	SVM (length)	0.84 ± 0.04	0.49 ± 0.05
	SVM (bow)	0.83 ± 0.03	0.57 ± 0.05
	SVM (length+bow)	0.84 ± 0.03	0.57 ± 0.07
	BERT	0.88 ± 0.02	0.63 ± 0.06

Table 4: Differences in predictions (joint classification) between rather objective and rather subjective ATUs.

semantic information by bag-of-words could nevertheless mostly improve prediction, especially for SVM and with respect to premises.

We further looked at predicting the subjectivity of ATUs and considered two granularities. While in the 2-class case all classifiers scored rather similar in the joint evaluation, in the 4-class case the differences became more obvious: In terms of F$_1$ score, BERT can outperform the other classifiers. Overall, it appears that our baseline models can already make some meaningful predictions for the complex task of whether an ATU triggers a subjective perception regarding its concreteness.

As for the different types of arguments, it shows that predicting concreteness and subjectivity is more difficult for major positions than for premises.

To gain further insight into the relationship between the task at hand and subjectivity, we examined the differences in the models' predictions of concreteness between "rather objective" and "rather subjective" ATUs (see Table 4). We found that all models did significantly better with the objective ATUs than with the subjective ones. We therefore hypothesize that the difficulty of assigning a standardized value to subjective ATUs is also shared by machine learning models due to the perspectivist scope.

6 Discussion

The evaluation of public participation can be supported by machine learning in a human-machine interaction. Not only machine prediction, but also pointing out cases where the user might potentially disagree can help with good evaluation practice. Perspectives can differ for a variety of reasons. First, it is due to the task itself, which is subjective. In addition, personal biases of the analyst may also contribute, such as their professional background (e.g., in our application case, whether they studied urban planning or administrative sciences). Furthermore, process-related demands on the evaluation may require the analyst to adjust their view. All these factors argue for a perspectivist approach.

As exemplified, our method can be integrated into workflows by adding a model for the sub-task of predicting subjectivity. While $\mathcal{T}_H$ reflects the prevailing opinion of the crowd, $\mathcal{T}_S$ can indicate how different coders' perceptions were when rating the unit - a valuable piece of information that is lost in non-perspectivist approaches. However, a potential barrier to applying our method to further use cases is the need for a non-aggregated dataset. The publication of annotations on an individual level is not yet common (Basile et al., 2021).

We found that objective ATUs (regarding their concreteness) can already be filtered out with an F$_1$ score between 0.73 and 0.80, depending on the granularity level (cf. Table 7 in Appendix A.3). However, the distinction between different degrees of subjectivity yielded weak results. Further research is needed to determine whether the problem lies in the task of predicting subjectivity, insufficient classification models, the dataset itself, or the transfer of the non-aggregated annotations to the labels for $\mathcal{H}_S$.

Concerning the original task of classifying the concreteness of arguments, the degree of concreteness (hard label) could be predicted with an accuracy of 0.80 and an F$_1$ of 0.67, which can already be helpful for supporting the manual evaluation of public participation processes.

7 Conclusion & Future Work

We introduced PerspectifyMe, a simple method to include perspectivism in machine learning workflows. Using argument concreteness as an example, we have shown that our baseline approaches can assess the subjective perception of ATUs.

In future work, we plan to apply advanced multi-task learning models as previous work has shown that they can lead to an increase in performance (Davani et al., 2022). Furthermore, we have tailored the transformation of the spectrum of annotations into a subjectivity score specific to the use case at hand. It would be of great interest to develop a more general (task-independent) algorithm.

Acknowledgements

Thanks to the anonymous reviewers for their helpful comments. This publication is based on research in the project CIMT/Partizipationsnutzen, which is funded by the Federal Ministry of Education and Research of Germany as part of its Social-Ecological Research funding priority, funding no. 01UU1904. Responsibility for the content of this publication lies with the author.

References

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, Marseille, France.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI*IA 2019 – Advances in Artificial Intelligence*, pages 588–603, Cham. Springer International Publishing.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):151–154.

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

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across nlp tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France. European Language Resources Association.

Federico Cabitza, Andrea Campagner, and Luca Maria Sconfienza. 2020. As if sand were stone. new concepts and metrics to probe the ground on which to build trustable ai. *BMC Medical Informatics and Decision Making*, 20(1):1–21.

Andrea Campagner, Davide Ciucci, Carl-Magnus Svensson, Marc Thilo Figge, and Federico Cabitza. 2021. Ground truthing from multi-rater labeling with three-way decision and possibility theory. *Information Sciences*, 545:771–790.

Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 2334–2346, New York, NY, USA. Association for Computing Machinery.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42, Sofia, Bulgaria. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

John S. Dryzek, André Bächtiger, Simone Chambers, Joshua Cohen, James N. Druckman, Andrea Felicetti, James S. Fishkin, David M. Farrell, Archon Fung, Amy Gutmann, Hélène Landemore, Jane Mansbridge, Sofie Marien, Michael A. Neblo, Simon Niemeyer, Maija Setälä, Rune Slothuus, Jane Suiter, Dennis Thompson, and Mark E. Warren. 2019. The crisis of democracy and the science of deliberation. *Science*, 363(6432):1144–1146.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Annette Hautli-Janisz, Ella Schad, and Chris Reed. 2022. Disagreement space in argument analysis. In

Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pages 1–9, Marseille, France. European Language Resources Association.

Shelby Heinecke and Lev Reyzin. 2019. Crowdsourced pac learning under classification noise. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):41–49.

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

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage publications.

Damjan Krstajic, Ljubomir J Buturovic, David E Leahy, and Simon Thomas. 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1):1–15.

Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W. Shulman. 2007. Identifying and classifying subjective claims. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, dg.o '07, page 76–81. Digital Government Society of North America.

John Lawrence, Joonsuk Park, Katarzyna Budzynska, Claire Cardie, Barbara Konat, and Chris Reed. 2017. Using argumentative structure to interpret debates in online deliberative democracy and erulemaking. *ACM Trans. Internet Technol.*, 17(3).

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? mining arguments in the German online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153, Berlin, Germany. Association for Computational Linguistics.

Anna Lindahl, Lars Borin, and Jacobo Rouces. 2019. Towards assessing argumentation annotation - a first step. In *Proceedings of the 6th Workshop on Argument Mining*, pages 177–186, Florence, Italy. Association for Computational Linguistics.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Jennifer A. Noble. 2012. Minority voices of crowdsourcing: Why we should pay attention to every member of the crowd. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*, CSCW '12, page 179–182, New York, NY, USA. Association for Computing Machinery.

OECD. 2003. *Promise and Problems of E-Democracy*. OECD.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Julia Romberg and Stefan Conrad. 2021. Citizen involvement in urban planning - how can municipalities be supported in evaluating public participation processes for mobility transitions? In *Proceedings of the 8th Workshop on Argument Mining*, pages 89–99, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julia Romberg and Tobias Escher. 2022. Automated topic categorisation of citizens' contributions: Reducing manual labelling efforts through active learning. In *Electronic Government*, pages 369–385, Cham. Springer International Publishing.

Julia Romberg, Laura Mark, and Tobias Escher. 2022a. *CIMT PartEval Corpus - Argument Concreteness (Subcorpus)*. ISLRN 776-577-161-062-9. https://github.com/juliaromberg/cimt-argument-concreteness-dataset.

Julia Romberg, Laura Mark, and Tobias Escher. 2022b. A corpus of german citizen contributions in mobility planning: Supporting evaluation through multidimensional classification. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2874–2883, Marseille, France. European Language Resources Association.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Carole H. Sudre, Beatriz Gomez Anson, Silvia Ingala, Chris D. Lane, Daniel Jimenez, Lukas Haider, Thomas Varsavsky, Ryutaro Tanno, Lorna Smith, Sébastien Ourselin, Rolf H. Jäger, and M. Jorge Cardoso. 2019. Let's agree to disagree: Learning highly debatable multirater labelling. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 665–673, Cham. Springer International Publishing.

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

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):173–177.

A Appendix

A.1 Details on the Dataset Transformation

Table 5 gives further insights into the generation of the subjectivity scores for the dataset.

High	Interm.	Low	#	L1	Subjectivity	
					4-class	2-class
5	0	0	439	0	O	RO
4	1	0	162	8	RO	RO
3	2	0	90	12	RS	RS
2	3	0	57	12	RS	RS
2	2	1	43	20	S	RS
1	3	1	38	16	RS	RS
0	3	2	38	12	RS	RS
3	1	1	37	20	S	RS
0	2	3	31	12	RS	RS
0	1	4	29	8	RO	RO
0	4	1	28	8	RO	RO
1	2	2	26	20	S	RS
1	4	0	25	8	RO	RO
0	5	0	20	0	O	RO
0	0	5	19	0	O	RO
4	0	1	18	16	RS	RS
1	1	3	11	20	S	RS
2	1	2	9	24	S	RS
1	0	4	3	16	RS	RS
2	0	3	2	24	S	RS
3	0	2	2	24	S	RS

Table 5: Overview of the different combinations of individual annotations, their occurence, the overall L1 distance and the mappings to subjectivity categories for both the 4-class and the 2-class schema. (O: Objective, RO: Rather Objective, RS: Rather Subjective, S: Subjective)

A.2 Hyperparameter-Tuning

For LR we tested the L1 and L2 norms for the penalty and set the regularization parameter C to take a value from $[0.001, 0.1, 1, 10, 100]$. Furthermore the classes were either weighted to simulate a balanced distribution or not weighted at all. We used an SVM with RBF kernel and a balanced class weighting. The regularization parameter C was set to be from $[0.001, 0.1, 1, 10, 100]$ and the kernel coefficient to be from $[1, 0.1, 0.01, 0.001]$. In RF the split quality was either measured with the Gini index or the Shannon information gain. Regarding the imbalance of the classes, we tested balancing weights and none.

For fine-tuning BERT we used the AdamW optimizer with beta coefficients of 0.9 and 0.999, and an epsilon of $1e{-}8$, and set the maximum sequence length to 128. We further trained for 5 epochs with a batch size from $[16, 32]$ and a learning rate from $[5e-5, 4e-5, 3e-5]$. For reproducibility of the experiments, we fixed the random seeds.

A.3 Full Overview of the Results

Table 6 and Table 7 list the full overview of results from the experiments.

		low	intermediate	high	macro-F$_1$	accuracy
major position	Baseline Majority	0.00	0.00	0.75	0.25	0.60
	RF (length)	0.19 ± 0.17	0.50 ± 0.07	0.81 ± 0.03	0.50 ± 0.07	0.69 ± 0.04
	RF (bow)	0.22 ± 0.13	0.58 ± 0.06	0.81 ± 0.03	0.54 ± 0.06	0.71 ± 0.04
	RF (length+bow)	0.17 ± 0.14	0.57 ± 0.06	0.82 ± 0.03	0.52 ± 0.06	0.71 ± 0.04
	LR (length)	0.13 ± 0.19	0.52 ± 0.08	0.81 ± 0.04	0.49 ± 0.06	0.70 ± 0.04
	LR (bow)	0.20 ± 0.13	0.55 ± 0.06	0.80 ± 0.03	0.52 ± 0.06	0.69 ± 0.03
	LR (length+bow)	0.22 ± 0.17	0.54 ± 0.06	0.80 ± 0.04	0.52 ± 0.06	0.69 ± 0.04
	SVM (length)	**0.45** ± 0.08	0.39 ± 0.09	0.83 ± 0.04	0.56 ± 0.04	0.69 ± 0.04
	SVM (bow)	0.28 ± 0.16	0.52 ± 0.11	0.79 ± 0.04	0.53 ± 0.07	0.67 ± 0.04
	SVM (length+bow)	0.33 ± 0.13	0.50 ± 0.09	0.82 ± 0.03	0.55 ± 0.06	0.70 ± 0.04
	BERT	0.38 ± 0.18	**0.63** ± 0.07	**0.86** ± 0.02	**0.62** ± 0.07	**0.76** ± 0.04
premise	Baseline Majority	0.00	0.00	0.79	0.26	0.65
	RF (length)	0.21 ± 0.18	0.63 ± 0.07	0.88 ± 0.02	0.57 ± 0.07	0.78 ± 0.03
	RF (bow)	0.32 ± 0.17	0.63 ± 0.06	0.89 ± 0.02	0.61 ± 0.06	0.79 ± 0.03
	RF (length+bow)	0.26 ± 0.17	**0.68** ± 0.05	0.90 ± 0.02	0.61 ± 0.06	0.81 ± 0.03
	LR (length)	0.16 ± 0.21	0.67 ± 0.04	0.90 ± 0.02	0.57 ± 0.07	0.80 ± 0.02
	LR (bow)	0.20 ± 0.13	0.55 ± 0.06	0.80 ± 0.03	0.52 ± 0.06	0.69 ± 0.03
	LR (length+bow)	0.25 ± 0.23	0.67 ± 0.05	0.90 ± 0.02	0.61 ± 0.08	0.80 ± 0.03
	SVM (length)	0.43 ± 0.09	0.47 ± 0.08	0.89 ± 0.02	0.60 ± 0.05	0.75 ± 0.03
	SVM (bow)	0.50 ± 0.12	0.63 ± 0.06	0.89 ± 0.02	0.67 ± 0.05	0.79 ± 0.03
	SVM (length+bow)	**0.51** ± 0.15	0.64 ± 0.08	0.90 ± 0.02	**0.68** ± 0.07	0.81 ± 0.03
	BERT	0.45 ± 0.16	**0.68** ± 0.06	**0.91** ± 0.02	**0.68** ± 0.06	**0.82** ± 0.03
joint	Baseline Majority	0.00	0.00	0.77	0.26	0.63
	RF (length)	0.15 ± 0.11	0.59 ± 0.05	0.86 ± 0.02	0.53 ± 0.04	0.75 ± 0.02
	RF (bow)	0.22 ± 0.13	0.61 ± 0.04	0.85 ± 0.02	0.56 ± 0.05	0.75 ± 0.02
	RF (length+bow)	0.28 ± 0.11	0.62 ± 0.04	0.86 ± 0.02	0.59 ± 0.05	0.76 ± 0.02
	LR (length)	0.16 ± 0.18	0.61 ± 0.04	0.84 ± 0.02	0.54 ± 0.06	0.74 ± 0.03
	LR (bow)	0.11 ± 0.11	0.62 ± 0.04	0.85 ± 0.02	0.53 ± 0.04	0.75 ± 0.02
	LR (length+bow)	0.16 ± 0.13	0.61 ± 0.05	0.85 ± 0.02	0.54 ± 0.04	0.74 ± 0.03
	SVM (length)	0.45 ± 0.07	0.46 ± 0.06	0.85 ± 0.02	0.59 ± 0.04	0.71 ± 0.02
	SVM (bow)	0.35 ± 0.10	0.58 ± 0.06	0.85 ± 0.02	0.59 ± 0.04	0.74 ± 0.03
	SVM (length+bow)	0.42 ± 0.11	0.58 ± 0.08	0.86 ± 0.02	0.62 ± 0.05	0.75 ± 0.03
	BERT	**0.47** ± 0.12	**0.66** ± 0.04	**0.88** ± 0.02	**0.67** ± 0.05	**0.79** ± 0.02

Table 6: Complete overview of all experiment results for sub-task $\mathcal{T}_H$: Concreteness.

		4-class					
		objective	rather objective	rather subjective	subjective	macro-F$_1$	accuracy
major position	Baseline Majority	0.57	0.00	0.00	0.00	0.14	0.40
	RF (length)	0.61 ± 0.04	0.21 ± 0.06	0.30 ± 0.08	0.30 ± 0.11	0.36 ± 0.05	0.42 ± 0.04
	RF (bow)	0.65 ± 0.04	0.16 ± 0.08	0.37 ± 0.08	0.18 ± 0.11	0.34 ± 0.04	0.45 ± 0.04
	RF (length+bow)	0.65 ± 0.04	0.12 ± 0.07	0.35 ± 0.08	0.20 ± 0.11	0.33 ± 0.04	0.46 ± 0.04
	LR (length)	0.65 ± 0.04	0.00 ± 0.00	**0.39** ± 0.11	0.02 ± 0.07	0.27 ± 0.04	0.46 ± 0.04
	LR (bow)	0.61 ± 0.05	0.10 ± 0.11	0.31 ± 0.13	0.11 ± 0.12	0.28 ± 0.06	0.42 ± 0.04
	LR (length+bow)	0.64 ± 0.05	0.11 ± 0.11	0.34 ± 0.10	0.15 ± 0.14	0.31 ± 0.06	0.44 ± 0.04
	SVM (length)	0.64 ± 0.05	0.09 ± 0.10	0.23 ± 0.11	**0.34** ± 0.10	0.33 ± 0.04	0.44 ± 0.04
	SVM (bow)	0.62 ± 0.05	0.10 ± 0.10	0.18 ± 0.15	0.23 ± 0.15	0.28 ± 0.08	0.42 ± 0.04
	SVM (length+bow)	0.64 ± 0.05	0.11 ± 0.09	0.26 ± 0.11	0.29 ± 0.11	0.33 ± 0.06	0.44 ± 0.04
	BERT	**0.69** ± 0.05	**0.24** ± 0.10	0.34 ± 0.08	0.22 ± 0.15	**0.37** ± 0.06	**0.47** ± 0.05
premise	Baseline Majority	0.62	0.00	0.00	0.00	0.15	0.44
	RF (length)	0.68 ± 0.05	0.19 ± 0.08	0.46 ± 0.08	0.05 ± 0.10	0.35 ± 0.04	0.49 ± 0.04
	RF (bow)	0.74 ± 0.04	0.10 ± 0.07	0.50 ± 0.06	0.19 ± 0.12	0.38 ± 0.05	0.56 ± 0.04
	RF (length+bow)	0.74 ± 0.04	0.10 ± 0.08	0.51 ± 0.06	0.18 ± 0.14	0.38 ± 0.05	**0.57** ± 0.04
	LR (length)	0.74 ± 0.04	0.01 ± 0.02	**0.53** ± 0.06	0.00 ± 0.03	0.32 ± 0.02	0.56 ± 0.04
	LR (bow)	0.72 ± 0.05	0.09 ± 0.10	0.51 ± 0.07	0.05 ± 0.08	0.34 ± 0.05	0.54 ± 0.05
	LR (length+bow)	0.73 ± 0.05	0.10 ± 0.09	0.52 ± 0.06	0.06 ± 0.09	0.35 ± 0.04	0.55 ± 0.04
	SVM (length)	0.71 ± 0.07	0.20 ± 0.10	0.19 ± 0.14	0.24 ± 0.10	0.33 ± 0.04	0.48 ± 0.05
	SVM (bow)	0.73 ± 0.05	0.11 ± 0.07	0.38 ± 0.20	0.21 ± 0.14	0.36 ± 0.05	0.53 ± 0.05
	SVM (length+bow)	0.72 ± 0.11	0.13 ± 0.10	0.40 ± 0.16	**0.27** ± 0.12	0.38 ± 0.07	0.53 ± 0.07
	BERT	**0.77** ± 0.05	**0.25** ± 0.09	0.51 ± 0.06	0.15 ± 0.13	**0.42** ± 0.05	0.56 ± 0.04
joint	Baseline Majority	0.60	0.00	0.00	0.00	0.15	0.42
	RF (length)	0.67 ± 0.03	0.15 ± 0.05	0.41 ± 0.05	0.14 ± 0.12	0.34 ± 0.04	0.47 ± 0.03
	RF (bow)	0.70 ± 0.03	0.12 ± 0.04	0.47 ± 0.06	0.18 ± 0.08	0.37 ± 0.04	0.51 ± 0.03
	RF (length+bow)	0.71 ± 0.03	0.09 ± 0.05	0.48 ± 0.06	0.18 ± 0.09	0.36 ± 0.03	**0.52** ± 0.03
	LR (length)	0.71 ± 0.03	0.00 ± 0.00	**0.49** ± 0.05	0.01 ± 0.05	0.30 ± 0.02	**0.52** ± 0.03
	LR (bow)	0.68 ± 0.04	0.09 ± 0.11	0.46 ± 0.05	0.07 ± 0.11	0.33 ± 0.05	0.50 ± 0.03
	LR (length+bow)	0.69 ± 0.04	0.11 ± 0.10	0.47 ± 0.06	0.10 ± 0.12	0.34 ± 0.05	0.50 ± 0.04
	SVM (length)	0.70 ± 0.04	0.13 ± 0.08	0.24 ± 0.09	**0.30** ± 0.06	0.34 ± 0.03	0.48 ± 0.03
	SVM (bow)	0.69 ± 0.03	0.15 ± 0.07	0.35 ± 0.14	0.27 ± 0.07	0.37 ± 0.05	0.49 ± 0.04
	SVM (length+bow)	0.70 ± 0.03	0.14 ± 0.07	0.37 ± 0.09	0.28 ± 0.08	0.37 ± 0.03	0.50 ± 0.03
	BERT	**0.73** ± 0.03	**0.27** ± 0.08	0.44 ± 0.05	0.25 ± 0.09	**0.42** ± 0.04	**0.52** ± 0.03

		2-class			
		rather objective	rather subjective	macro-F$_1$	accuracy
major position	Baseline Majority	0.78	0.00	0.39	0.64
	RF (length)	0.70 ± 0.05	0.49 ± 0.09	0.59 ± 0.05	0.62 ± 0.04
	RF (bow)	0.76 ± 0.03	**0.58** ± 0.07	0.67 ± 0.05	0.70 ± 0.04
	RF (length+bow)	0.77 ± 0.03	**0.58** ± 0.06	**0.68** ± 0.04	0.70 ± 0.04
	LR (length)	0.77 ± 0.04	0.42 ± 0.22	0.59 ± 0.11	0.68 ± 0.04
	LR (bow)	0.75 ± 0.04	0.45 ± 0.23	0.60 ± 0.10	0.67 ± 0.04
	LR (length+bow)	0.75 ± 0.04	0.52 ± 0.20	0.63 ± 0.10	0.68 ± 0.05
	SVM (length)	0.74 ± 0.04	0.54 ± 0.10	0.64 ± 0.05	0.67 ± 0.04
	SVM (bow)	0.73 ± 0.11	0.54 ± 0.16	0.63 ± 0.09	0.67 ± 0.06
	SVM (length+bow)	0.75 ± 0.04	0.53 ± 0.12	0.64 ± 0.06	0.68 ± 0.04
	BERT	**0.78** ± 0.04	**0.58** ± 0.09	**0.68** ± 0.06	**0.71** ± 0.05
premise	Baseline Majority	0.78	0.00	0.39	0.64
	RF (length)	0.78 ± 0.04	0.65 ± 0.04	0.71 ± 0.03	0.73 ± 0.03
	RF (bow)	0.81 ± 0.03	0.64 ± 0.06	**0.73** ± 0.04	0.75 ± 0.04
	RF (length+bow)	**0.82** ± 0.03	0.65 ± 0.06	**0.73** ± 0.04	**0.76** ± 0.04
	LR (length)	0.81 ± 0.03	0.64 ± 0.07	**0.73** ± 0.05	0.75 ± 0.04
	LR (bow)	0.79 ± 0.04	0.63 ± 0.05	0.71 ± 0.03	0.74 ± 0.03
	LR (length+bow)	0.79 ± 0.03	0.65 ± 0.05	0.72 ± 0.04	0.74 ± 0.04
	SVM (length)	0.80 ± 0.04	0.64 ± 0.05	0.72 ± 0.04	0.74 ± 0.04
	SVM (bow)	0.79 ± 0.04	0.64 ± 0.05	0.72 ± 0.04	0.74 ± 0.04
	SVM (length+bow)	0.80 ± 0.03	0.63 ± 0.06	0.71 ± 0.04	0.74 ± 0.04
	BERT	0.81 ± 0.03	**0.66** ± 0.06	**0.73** ± 0.04	**0.76** ± 0.04
joint	Baseline Majority	0.78	0.00	0.39	0.64
	RF (length)	0.76 ± 0.03	0.58 ± 0.03	0.67 ± 0.02	0.70 ± 0.02
	RF (bow)	0.79 ± 0.02	0.63 ± 0.03	0.71 ± 0.02	0.73 ± 0.02
	RF (length+bow)	**0.80** ± 0.02	0.62 ± 0.03	0.71 ± 0.02	**0.74** ± 0.02
	LR (length)	0.78 ± 0.02	0.58 ± 0.06	0.68 ± 0.03	0.72 ± 0.02
	LR (bow)	0.77 ± 0.03	0.60 ± 0.05	0.69 ± 0.03	0.71 ± 0.03
	LR (length+bow)	0.77 ± 0.03	0.61 ± 0.04	0.69 ± 0.03	0.72 ± 0.03
	SVM (length)	0.78 ± 0.02	0.63 ± 0.03	0.70 ± 0.02	0.72 ± 0.02
	SVM (bow)	0.77 ± 0.03	0.62 ± 0.04	0.69 ± 0.02	0.71 ± 0.03
	SVM (length+bow)	0.78 ± 0.02	0.61 ± 0.04	0.70 ± 0.03	0.72 ± 0.02
	BERT	**0.80** ± 0.02	**0.64** ± 0.04	**0.72** ± 0.02	**0.74** ± 0.02

Table 7: Complete overview of all experiment results for sub-task $\mathcal{T}_S$: Subjectivity.