Citizen Involvement in Urban Planning - How Can Municipalities Be Supported in Evaluating Public Participation Processes for Mobility Transitions?

Julia Romberg
Institute of Social Sciences
Heinrich Heine University Düsseldorf
julia.romberg@hhu.de

Stefan Conrad
Institute of Computer Science
Heinrich Heine University Düsseldorf
stefan.conrad@hhu.de

Abstract

Public participation processes allow citizens to engage in municipal decision-making processes by expressing their opinions on specific issues. Municipalities often only have limited resources to analyze a possibly large amount of textual contributions that need to be evaluated in a timely and detailed manner. Automated support for the evaluation is therefore essential, e.g. to analyze arguments. In this paper, we address (A) the identification of *argumentative* discourse units and (B) their classification as *major position* or *premise* in German public participation processes. The objective of our work is to make argument mining viable for use in municipalities. We compare different argument mining approaches and develop a generic model that can successfully detect argument structures in different datasets of mobility-related urban planning. We introduce a new data corpus comprising five public participation processes. In our evaluation, we achieve high macro F_1 scores (0.76 - 0.80 for the identification of argumentative units; 0.86 - 0.93 for their classification) on all datasets. Additionally, we improve previous results for the classification of argumentative units on a similar German online participation dataset.

1 Introduction

In many democratic countries, political decisions are increasingly developed through the participation of citizens. *Public participation processes* allow citizens to voice their suggestions and concerns on specific issues, for example in urban planning, and thus influence decision-making processes. Participation can take place in formats that vary from on-site events such as citizen workshops, to written submissions via letter or e-mail, and to online platforms where citizens can discuss proposals digitally. Building on Scharpf (1999), we can distinguish two main goals of public participation processes. On the one hand, the additional input provided by citizens can influence the decision-making process and, potentially, lead to more effective policies. On the other hand, citizens are assumed to develop a higher acceptance of the output when given an opportunity to participate and, ultimately, the resulting decisions have a higher legitimacy.

In order to be able to include citizen comments in the further decision-making process, those comments first have to be evaluated. However, both offline and online participation formats have the potential to generate a high number of responses (Shulman, 2003; Schlosberg et al., 2008), e.g., thousands of contributions. Along with stringent schedules in decision-making processes, this often poses major challenges for municipalities. Still, participation contributions are commonly evaluated manually with considerable effort. Therefore, if municipalities do not have enough resources (human or monetary) to shoulder this effort, the detailed evaluation will have to be cut back. As a result, opinions might be completely omitted or not been taken into account equally. This in turn can have a negative influence on the goals of public participation processes. Filtering out individual or mass opinions risks loosing important clues for effective policies. It can also endanger citizens' confidence in the opportunity to participate in decision-making and weaken civic engagement (Mendelson, 2012). Besides, decision acceptance is influenced by perceived fairness (Esaiasson, 2010).

Automating the evaluation of public participation processes can help overcome these problems (OECD, 2004) and has been addressed by research initiatives such as the *Cornell eRulemaking Initiative* (CeRI)[1] and, more recently, the *Citizen participation and machine learning for a better democracy project*[2]. Over the years, several tasks that arise in the evaluation process have been high-

[1] https://scholarship.law.cornell.edu/ceri/

[2] https://www.turing.ac.uk/research/research-projects/citizen-participation-and-machine-learning-better-democracy

Proceedings of The 8th Workshop on Argument Mining, pages 89–99
Punta Cana, Dominican Republic, November 10–11, 2021. ©2021 Association for Computational Linguistics

lighted. These include thematic classification and clustering of citizen contributions (e.g. Kwon et al., 2007; Purpura et al., 2008; Arana-Catania et al., 2021; Teufl et al., 2009), summarization of similar content (e.g Arana-Catania et al., 2021), detection of duplicates (e.g. Yang et al., 2006), and analysis of arguments and opinions (e.g. Kwon et al., 2007; Park and Cardie, 2014; Lawrence et al., 2017).

In this paper, we focus on arguments in public participation processes that address sustainable mobility and land use in Germany. German cities have involved their citizens in hundreds of decision-making processes on these issues in recent years.[3] We look at five of them in detail, four of which are processes for concrete improvements to cycling infrastructure and one of which is a strategic process for creating a general mobility concept for a city. At the same time, we consider two very different participation formats, namely online platforms and questionnaires.

This paper's first objective is to analyze the strengths and weaknesses of previously published argument mining approaches for public participation processes when they are applied to different German datasets. Our attention is focused on the classification of text segments as *argumentative* or *non-argumentative*, as well as on the downstream classification of *argumentation components*. In addition to our datasets, we include the only other German public participation dataset (to the best of our knowledge) for argument mining (Liebeck et al., 2016) in the evaluation.

Our second objective is to improve the results obtained on the datasets under consideration by the previous approaches for both classification tasks. For this we apply BERT (Devlin et al., 2019) which is known to perform very well on many tasks including argument mining.

In practice, the use of argument mining to evaluate public participation processes only adds value when the benefits outweigh the effort. Manual coding of data and the training or fine-tuning of machine learning models are costly. In addition, machine learning requires expert knowledge and usually cannot be performed directly by the municipalities. An optimal solution would be a universally valid model that can be applied flexibly to new datasets. Our third objective is hence to investigate the extent to which trained models can recognize argument structures in other public participation processes that were not part of the training process.

Our contributions are: (1) We present a new data corpus of five mobility-related public participation processes that vary in content and format. The German corpus comprises $17,306$ sentences coded with an argument scheme tailored to informal public participation processes. (2) We perform a broad comparison of previously published best approaches for argument mining in public participation processes, which so far have been evaluated mostly on distinct datasets. We compare the algorithms directly on our data corpus and compare the performances. (3) We show that BERT surpasses previously published argument mining approaches for public participation processes on German data for both tasks. Especially when classifying argument components, macro F_1 results improve by between 0.05 and 0.12 depending on the dataset. (4) In a cross-dataset evaluation, we show that BERT models trained on one dataset can recognize argument structures in other public participation datasets (which were not part of the training) with comparable goodness of fit. This finding is an important step towards practical application in municipalities.

2 Related Work

Mining arguments in the domain of citizen participation has been the subject of several studies. Much of this work centers on U.S. e-rulemaking initiatives, where citizens are given the opportunity for feedback on rule proposals. An early attempt to identify, classify, and relate arguments in e-rulemaking was made by Kwon et al. (2006); Kwon and Hovy (2007). Arguments were built as trees of claims and subclaims or main-support with support relations. Eidelman and Grom (2019) extended the detection of generic argument components (support and opposition) with corpus-specific argument types. Niculae et al. (2017), Galassi et al. (2018) and Cocarascu et al. (2020) differentiate between five proposition types (fact, testimony, value, policy, and reference) and evidence or reason relations. In addition, other research examined specific properties of argumentation and discourse in public participation processes. Park and Cardie (2014) identified the lack of appropriate justifications as a common problem in the analysis of citizen contributions and tried to predict whether and by what means a proposal is verifiable. Subsequent work

[3] The research project *Citizen Involvement in Mobility Transitions* (CIMT) has identified more than 350 processes directly related to mobility since 2015.

was presented by Park et al. (2015) and Guggilla et al. (2016). Furthermore, Lawrence et al. (2017) and Konat et al. (2016) investigated discourse analysis in more detail and measured controversy and divisiveness in argument graphs.

Besides e-rulemaking initiatives, informal public participation processes were considered. Our work shares most similarity to Liebeck et al. (2016) who focused on a German-language process about the restructuring of a former airport area. The authors developed an argumentation scheme specifically adapted to discursive online public participation processes. With regard to languages other than German, Fierro et al. (2017) and in a follow-up work Giannakopoulos et al. (2019) studied a corpus consisting of over $200,000$ political arguments in Chilean Spanish dialect, derived from a participatory process to form a new constitution for Chile. The arguments were classified thematically according to constitutional concepts and also as either *policies*, *facts* or *values*. Further work (Morio and Fujita, 2018a,b) paid attention to the complex structure of arguments in public online participation. Relying on a Japanese dataset, the authors presented an annotation scheme for discussion threads taking care of inner-post relations and inter-post interactions.

Although the work to date has produced encouraging results, most approaches are not yet mature for practical use (e.g. with German public participation processes). Only few previous research addressed the development of general models (see Cocarascu et al. (2020), who perform a cross-dataset comparison of baselines for relation prediction). Therefore, this paper investigates the cross-data transferability of trained models for the identification and classification of argument components in public participation processes, an investigation that is highly relevant for practical use.

3 Data Corpus

3.1 Datasets

Our five datasets originate from urban planning and are concerned with mobility. Four of them represent very specific processes for improving cycling as a mode of transportation, the fourth dataset stems from a more general strategic process for developing a mobility concept. These five datasets comprise different participation types, i.e., online platforms and questionnaires.

Cycling dialogues The *cycling dialogues* were a pilot project for improving the cycle traffic infrastucture in three German cities, namely Bonn, Cologne and Moers. During a five-week period in 2017, citizens were able to participate (make propositions, discuss and rate propositions or comments) in a map-based online consultation[4]. While in Bonn and Moers suggestions for improvement could be made city-wide, the focus in Cologne was on a specific city district. As a result, three datasets of similar online public participation processes from different local contexts emerged. In the following, these datasets will be referred to as *CD_B*, *CD_C* and *CD_M*. We focus on the initial text contributions in which citizens make new proposals. CD_B is the largest dataset comprising $12,103$ sentences from $2,364$ contributions, whereas CD_C and CD_M are considerably smaller, with 366 and 459 contributions consisting of $1,704$ and $2,193$ sentences, respectively. On average, the contributions consist of 4.83, 4.66 and 4.78 sentences ($\sigma = 2.63$, $\sigma = 3.00$ and $\sigma = 2.61$) with 15.94, 15.16 and 15.43 tokens ($\sigma = 10.92$, $\sigma = 10.45$ and $\sigma = 10.81$).

Mobility concept Since 2019, the German city of Krefeld has been planning how the city's mobility should look like in the future. In addition to various on-site events, multiple public participation processes were carried out online. The here presented dataset *MC_K* includes the $2,008$ sentences of the 337 initial contributions from two interrelated online processes. In the first process, citizens were informed about the drafts of seven citywide action plans. The fields of action were *urban development and regional cooperation*, *flowing motor vehicle traffic*, *commercial transport*, *stationary traffic*, *public transport*, *bicycle traffic*, and *foot traffic*. As part of the planning process, citizens were asked to comment on the planned actions. The second process gave citizens the opportunity to submit concrete propositions for actions in specified city districts. Citizens wrote an average of 5.96 sentences ($\sigma = 5.63$), slightly more than in the processes described above. The average 15.25 words per sentence ($\sigma = 10.80$) resemble the cycling dialogues.

Citizen questionnaire on cycling Accompanying the cycling dialogues, a postal survey was con-

[4] In urban planning, propositions usually refer to specific places. Maps are often used to provide assistance.

	CD_B		CD_C		CD_M		MC_K		CQ_B	
non-arg	1,153	(11.3%)	197	(11.9%)	382	(17.8%)	431	(22.2%)	172	(12.4%)
mpos	2,589	(25.4%)	556	(33.6%)	359	(16.7%)	892	(46.0%)	960	(69.5%)
prem	6,438	(63.2%)	904	(54.6%)	1,407	(65.5%)	616	(31.8%)	250	(18.1%)
total	10,180		1,657		2,148		1,939		1,382	

Table 1: Distribution of sentences among the different coding categories per dataset (absolute and percentage).

ducted in a randomized sample of each city's population. The citizens were asked to submit suggestions for improvements to cycling in free-text fields. Respondents could fill out the questionnaire either by hand or online. In this paper, we focus on the 1,386 citizen contributions from the city of Bonn (*CQ_B*) which consist of 1,505 sentences. By comparing the length of the survey contributions (1.09 sentences on average ($\sigma = 0.37$), 7.75 tokens per sentence ($\sigma = 6.30$)) with the online platform contributions, we can clearly see that citizens write more succinct in surveys of this type.

3.2 Argumentation Model

A key aspect of public participation is that citizens can submit their own ideas on a given topic, such as the cycling infrastructure of a city or the development of a mobility concept. One contribution from CD_B, translated into English, e.g. states: "A new pavement is urgently needed here to be able to cycle along. The current pavement has grooves & cracks in the surface, so that cycling between Ringstraße & Kreuzherrenstraße is very risky, especially in wet conditions." The writer proposes to renew the pavement and substantiates this with the current poor and dangerous condition of the pavement. In urban planning processes, causes for suggested improvements are mostly descriptions of infrastructure problems or (perceived) planning deficits, while the propositions are measures to overcome these issues. Several interviews we conducted in 2020 with local authorities and urban planning practitioners emphasized the value in automatically recognizing the problems that citizens describe and the solutions they propose in text contributions (Romberg and Escher, 2020).

We follow the terminology of Liebeck et al. (2016), who developed an argumentation model for informal online public participation processes based on three argument components: *major positions* provide "options for actions or decisions that occur in the discussion". In simpler terms, these are the propositions that citizens make. *Premises* are "reasons that attack or support a major position, a claim or another premise". *Claims* are defined as "pro or contra stance towards a major position". In this work, we rely on the concepts of major positions and premises, as our focus is on the detection of propositions and underlying reasons. We leave for future work the detection of pro or contra stances expressed by fellow citizens in the feedback comments on initial proposals (in the case of dialogical processes).

3.3 Annotation Process

Coding guidelines were developed on 201 contributions from the cycling dialogues Bonn, which were excluded from the subsequent annotation process, reducing the sentences to be coded in CD_B to 10,442. Each sentence was labeled as *non-argumentative* (non-arg), *major position* (mpos) or *premise* (prem). In case a sentence contained multiple argumentation components, multi-labeling was allowed. Since contribution titles often contained parts of the argument, they were included as additional sentences.

We measured the inter-coder agreement on 10% of the contributions of each dataset, which were respectively annotated by three trained coders. In a subsequent curation step, disagreements were resolved by two supervisors to obtain unambiguous coding of the contributions used to measure the inter-coder agreement. High Fleiss' κ values prove the reliability of the codings: 0.76 (CD_B), 0.80 (CD_C), 0.77 (CD_M), 0.73 (MC_K), and 0.76 (CQ_B). During curation, certain edge cases became obvious. We believe that this subjectivity is also reflected in a human evaluation, which is why a small deviation in coding seems acceptable, also with regard to the training of the classification algorithms. The remaining 90% of the contributions were divided equally among the coders (each 30%) and annotated independently. These sentences were not curated; however, due to the high agreement on the over 1,700 sentences that were coded by all three annotators, we assume similar reliability on the sentences labeled by one person only.

Since the approaches we compare in this paper are tailored to single-label classifications, we omit sentences containing both major position and premise to be addressed in future work. This affects 548 sentences (262 in CD_B, 49 in CD_C, 45 in CD_M, 69 in MC_K, and 123 in CQ_B).

Table 1 shows the distribution of classes included in the evaluation across the five datasets. The majority of sentences in all datasets are argumentative, accounting for between 77.8% and 88.6%. Major positions and premises are distributed very differently throughout the datasets. While premises are made more frequently in the cycling dialogues, major positions are favored in MC_K and especially in CQ_B. The datasets are available under a Creative Commons License at https://github.com/juliaromberg/cimt-argument-mining-dataset/.

4 Methodology

Argument Mining can be divided into three subtasks: *segmentation*, *segment classification*, and *relation identification* (Peldszus and Stede, 2013). First, argumentative text is split into argument discourse units (ADUs). Second, ADUs are classified according to their function in the argument. Third, relations between ADUs are identified. Peldszus and Stede (2013) assume here that it is known which texts are argumentative or relevant for the argumentation. Lawrence and Reed (2019) widen the first task and include the *distinction between argumentative and non-argumentative units*.

In this work, we focus on (A) the classification of discourse units as argumentative (ADU) and non argumentative (non-ADU) and (B) the classification of ADUs according to contextual clausal properties for informal public participation processes. In the following, these two tasks will be referred to as *Task A* and *Task B*. We define each sentence as discourse unit, so that both tasks are sentence-level classification tasks.

4.1 Previously Applied Argument Mining Approaches for Public Participation

Our first objective is to compare the previously used approaches for solving Task A and Task B in public participation processes on our datasets. In the following, we provide an overview of these algorithms and describe in detail the setups we chose for our experiments (e.g. input features, hyperparameter selection). The results of our experiments are described and discussed in Section 5. For every dataset in consideration, we used a 5-fold cross-validation, dividing the datasets into 80% training and 20% test data each time. We tuned algorithm hyperparameters using a grid search with cross-validation (5 folds) for each split of the (outer) cross-validation.

4.1.1 Task A

All of the works considering the distinction between ADUs and non-ADUS have predefined sentences as elementary discourse units, as we do.

SVM Kwon et al. (2006), Liebeck et al. (2016) and Morio and Fujita (2018a) used support vector machines (Cortes and Vapnik, 1995) to detect ADUs with F_1 scores between 0.52 and 0.70.

For our experiments, we adopted the best setup of Liebeck et al. (2016) since their dataset is most similar to ours. Sentences were represented as a combination of unigrams and grammatical features, more precisely a L_2-normalized POS-Tag distribution[5] and a L_2-normalized distribution of dependencies[6]. We used the radial basis function kernel, and considered $C \in \{1, 10, 100\}$ and $\gamma \in \{0.001, 0.01, 0.1\}$ in the grid search. We further weighted the training samples inversely proportional to the class frequencies to take care of the strong class imbalance of our datasets.

fastText Eidelman and Grom (2019) suggested the use of fastText (Joulin et al., 2017) and proposed balancing the training data for highly imbalanced datasets. By downsampling the majority class in the corresponding dataset, they improved the macro F_1 outcome from 0.80 to 0.90.

In our experiments, we trained two fastText models per dataset: One on the original, imbalanced dataset and one on a balanced version of the dataset where the majority class was undersampled by randomly picking samples. We used pretrained fastText embeddings for German with 50 dimensions, and included learning rates of $1e-1$, $5e-1$ and $9e-1$, and 5 or 10 epochs of training in the grid search.

4.1.2 Task B

More attention has been paid to the classification of ADUs in previous work.

SVM Kwon et al. (2006), Park and Cardie (2014), Liebeck et al. (2016) and Morio and Fujita

[5] STTS tagset (Thielen and Schiller, 2011)

[6] TIGER scheme (Albert et al., 2003)

(2018a) classified argument components in public participation processes with SVMs. Depending on the dataset and argumentation scheme, they yielded macro F_1 values in the range of 0.56 to 0.77.

For our experiments, we again relied on the closely related work of Liebeck et al. (2016) and used the same setup as described in Section 4.1.1.

fastText In Fierro et al. (2017) and Eidelman and Grom (2019), fastText provided the best results (0.65 and 0.78). Of particular interest is that, on the Spanish dataset (Fierro et al., 2017), fastText surpassed the SVM. We were curious to see if this behavior applies to our datasets as well.

In our experiments, we replicated the implementation of Fierro et al. (2017) using pretrained fastText embeddings (we chose 50 dimensions) and word bigrams in the classification. Grid search considered learning rates of $1e-1$, $5e-1$ and $9e-1$, and 5 or 10 epochs of training. Similar to Task A, classes were imbalanced in our datasets, and we thus trained models with and without undersampling.

ECGA Further deep learning architectures have been considered by Guggilla et al. (2016) and Giannakopoulos et al. (2019). While Guggilla et al. (2016) showed that the use of convolutional neuronal networks (CNN) (LeCun et al., 1998) can marginally improve the results of an SVM, the advantages of deep learning become more obvious in the work of Giannakopoulos et al. (2019). Using an ensemble method called ECGA, a combination of multiple learners, they improved the results of Fierro et al. (2017) by 0.07. Each learner is composed of a CNN followed by bidirectional gated recurrent units (BiGRU) (Cho et al., 2014), connected to an attention layer (Bahdanau et al., 2015). The class predictions of the multiple learners are averaged to obtain final predictions. FastText embeddings build the input matrix. For argument classification, Giannakopoulos et al. (2019) proposed the use of two learners with kernel sizes of 2 and 3 as well as 512 filters in the convolution and 256 GRU units.

Since the proposed architecture failed to produce reasonable results on our datasets, we reduced the number of GRU units in our experiments to 64 and the number of convolution filters in to 128. We took our cue from the authors' best model for solving a different task, textual churn detection, with a smaller corresponding dataset. Despite the reduced model architecture, ECGA still tended to neglect the minority class in our datasets. To counteract this, we additionally evaluated ECGA with undersampling. We tried batch sizes of 2, 4, and 8, as well as 1 and 2 kernels or 2 and 3 kernels for the two learners. The training ran for 200 epochs with the option of early stopping if the loss did not improve within 10 epochs.

4.2 Bidirectional Encoder Representations from Transformers for Argument Mining in Public Participation Processes

Our second objective is to improve the results obtained by the previous approaches on our datasets for both classification tasks. To this end, we use BERT (Devlin et al., 2019) which has already provided promising results for Task A and Task B in other text domains, such as on persuasive online forums (Chakrabarty et al., 2019) and on heterogeneous sources of argumentative content (Reimers et al., 2019). With public participation processes, BERT has so far only been used to identify relations between ADUs (Cocarascu et al., 2020).

We expected BERT to also perform well for Task A and Task B on public participation datasets and to outperform the other algorithms in the evaluation. We used case-sensitive German BERT[7] with an additional linear layer for sequence classification. For fine-tuning, we relied on the suggestions of Devlin et al. (2019) and included batch sizes of 16 and 32, learning rates of $5e-5$, $3e-5$ and $2e-5$, and 1 to 4 epochs of training in the grid search.

4.3 Model Generalizability

This work's third objective is to investigate model generalizability in a cross-dataset evaluation. The previous two evaluation objectives were to determine which approach generates the best results for each dataset. To this end, both the training and the test data stem from the same dataset. In a practical application, this would mean that a sufficiently large amount of citizen contributions would have to be coded manually by local authorities. However, a more feasible and cost-effective solution would be to provide a pretrained classification model that can reliably recognize argument structures in new participation processes without the need for further training. Our goal is to provide such a model for public participation processes of mobility-related urban planning. The diversity in subjects and for-

[7] https://www.deepset.ai/german-bert

mats in our data corpus is well suited for testing the transferability to a range of processes.

For the cross-dataset evaluation, we used the evaluation setup described in Section 4.1 (5-fold cross validation, hyperparameter tuning) and trained on CD_B in our experiments. We intentionally chose the largest dataset for training to provide reliable models. For every approach, we then applied the five resulting models to the remaining datasets and averaged the results for each dataset to obtain an average macro F_1 score. Algorithms were implemented as described in Sections 4.1 and 4.2.

For Task A, we evaluated SVM, fastText without undersampling (as will be shown in Section 5.1.1, undersampling of CD_B provided no advantage), and BERT. For Task B, we chose to evaluate models trained on undersampled data and models trained on the original data alongside. Our decision was due to the very different distribution of ADU-types in our datasets: while premises prevail in the cycling dialogues (62%-80% prem), major positions are more present in MC_K (59% mpos) and in CQ_B (80% mpos). We thus wanted to investigate whether models trained on balanced data could provide more stable results across the different datasets. To sum up, we compared the behavior of eight approaches in the cross-dataset evaluation for Task B: SVM, fastText, ECGA, and BERT trained on the original CD_B dataset, and trained on an undersampled CD_B dataset.

5 Results and Discussion

5.1 Comparison of the Approaches

In the following, we evaluate for both classification tasks the approaches from previous work (see Section 4.1) and BERT (see Section 4.2) on our corpus from Section 3. For completeness, we also have a look at the only other German public participation dataset for argument mining, *THF Airport ArgMining Corpus* (Liebeck et al., 2016). *THF* provides 2,078 argumentative and 355 non-argumentative sentences for Task A, and 509 major positions, 1,170 premises, and 311 claims for Task B.[8]

5.1.1 Task A

Results for the classification of ADUs and non-ADUs are given in Table 2. For each dataset, only the results of the superior fastText model are listed. Undersampling models are marked with an asterisk. Overall, BERT performed best with macro F_1 values up to 0.80, improving most SVM scores by at least 0.03.[9] However, on THF the SVM yielded slightly better results. FastText struggled with the minority class. The problem was particularly evident in the three datasets with the fewest non-argumentative samples, where undersampling could improve the results at least to some degree.

		SVM	*fastText*	*BERT*
CD_B	arg	0.93 (0.00)	0.94 (0.00)	**0.95** (0.00)
	non-arg	0.52 (0.04)	0.41 (0.04)	**0.57** (0.04)
	macro	0.73 (0.02)	0.68 (0.02)	**0.76** (0.02)
CD_C	arg	0.93 (0.01)	0.87 (0.02)*	**0.95** (0.01)
	non-arg	0.53 (0.10)	0.42 (0.06)*	**0.58** (0.12)
	macro	0.73 (0.06)	0.64 (0.04)*	**0.77** (0.07)
CD_M	arg	0.90 (0.01)	0.92 (0.01)	**0.94** (0.01)
	non-arg	0.59 (0.05)	0.51 (0.05)	**0.67** (0.04)
	macro	0.75 (0.03)	0.71 (0.03)	**0.80** (0.03)
MC_K	arg	0.86 (0.01)	0.89 (0.01)	**0.91** (0.01)
	non-arg	0.53 (0.03)	0.45 (0.02)	**0.62** (0.06)
	macro	0.69 (0.02)	0.67 (0.01)	**0.77** (0.04)
CQ_B	arg	0.94 (0.02)	0.85 (0.02)*	**0.96** (0.01)
	non-arg	0.53 (0.07)	0.42 (0.04)*	**0.56** (0.16)
	macro	0.73 (0.05)	0.63 (0.03)*	**0.76** (0.09)
THF	arg	0.91 (0.01)	0.79 (0.03)*	**0.92** (0.01)
	non-arg	**0.48** (0.03)	0.37 (0.03)*	0.46 (0.05)
	macro	**0.70** (0.02)	0.58 (0.03)*	0.69 (0.03)

Table 2: Results for Task A on the individual datasets. Scores are mean F_1 values of the five test sets, standard deviation is given in parentheses.

5.1.2 Task B

Table 3 shows the findings for argument component classification. For fastText and ECGA, two model variants were evaluated (with and without undersampling), of which the better one is listed. Undersampling models are marked with an asterisk. While undersampling slightly increased the macro performance of ECGA on all datasets, there was no enhancement with fastText. Contrary to our expectations, ECGA performed worse than fastText and could only keep up with the other approaches for datasets that have sufficient samples in the minority class. BERT showed outstanding results and could significantly advance the classification, especially for the minority classes: Compared to the also good SVM, the prediction of major positions

[8] We evaluate the dataset according to our methodology instead of the suggested train-test split by Liebeck et al. (2016).

[9] BERT models show a high standard deviation in the minority classes of CD_C and CQ_B. Variance is due to the small number of non-arg sentences in the cross-validation for hyperparameter tuning. Fixed hyperparameters yield comparable F_1 values and much lower standard deviation.

		SVM	*fastText*	*ECGA*	*BERT*
CD_B	mpos	0.82 (0.01)	0.79 (0.01)	0.78 (0.03)	**0.90** (0.01)
	prem	0.93 (0.01)	0.93 (0.00)	0.92 (0.01)	**0.96** (0.00)
	macro	0.88 (0.01)	0.86 (0.01)	0.85 (0.02)	**0.93** (0.01)
CD_C	mpos	0.77 (0.02)	0.74 (0.02)	0.76 (0.03)*	**0.89** (0.02)
	prem	0.85 (0.01)	0.86 (0.01)	0.84 (0.01)*	**0.93** (0.01)
	macro	0.81 (0.02)	0.80 (0.02)	0.80 (0.02)*	**0.91** (0.02)
CD_M	mpos	0.67 (0.03)	0.58 (0.03)	0.52 (0.08)*	**0.84** (0.06)
	prem	0.92 (0.01)	0.92 (0.00)	0.86 (0.05)*	**0.91** (0.04)
	macro	0.80 (0.03)	0.75 (0.02)	0.69 (0.06)*	**0.90** (0.03)
MC_K	mpos	0.83 (0.03)	0.83 (0.03)	0.84 (0.03)*	**0.88** (0.02)
	prem	0.75 (0.03)	0.74 (0.04)	0.74 (0.05)*	**0.84** (0.03)
	macro	0.79 (0.03)	0.78 (0.04)	0.79 (0.05)*	**0.86** (0.03)
CQ_B	mpos	0.93 (0.02)	0.92 (0.01)	0.89 (0.03)*	**0.97** (0.01)
	prem	0.70 (0.08)	0.58 (0.06)	0.55 (0.10)*	**0.88** (0.03)
	macro	0.81 (0.05)	0.75 (0.04)	0.72 (0.06)*	**0.93** (0.02)
THF	mpos	0.53 (0.05)	0.46 (0.03)	0.46 (0.04)*	**0.68** (0.03)
	prem	0.78 (0.01)	0.79 (0.01)	0.60 (0.06)*	**0.84** (0.03)
	claim	0.60 (0.03)	0.59 (0.06)	0.51 (0.06)*	**0.63** (0.06)
	macro	0.64 (0.02)	0.61 (0.03)	0.52 (0.04)*	**0.72** (0.04)

Table 3: Results for Task B on the individual datasets. Scores are mean F_1 values of the five test sets, standard deviation is given in parentheses.

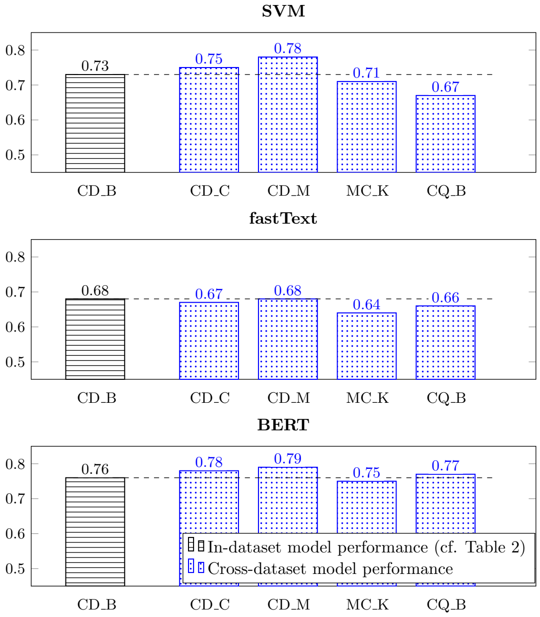

Figure 1: Cross-dataset evaluation for Task A. Results are averaged macro F_1 values of the five models trained on CD_B.

(CD_B, CD_C, CD_M, THF) improved by at least 0.08 up to 0.17. Premises were predicted with an improvement of 0.09 and 0.18 (MC_K, CQ_B).

5.2 Cross-Dataset Evaluation

Next, we look at the generalization performance of the learned models for both classification tasks.

5.2.1 Task A

Figure 1 shows the cross-dataset results of the CD_B models on the other datasets. BERT could consistently achieve good macro F_1 values (between 0.75 and 0.79) for all datasets, close to the score of 0.76 that BERT achieved on the refence dataset CD_B ($\sigma = 0.02$). The obtained values are also comparable to the results of dataset-internal results from Section 5.1. Equally stable was fastText ($\sigma = 0.02$), but results were on average 0.10 points lower. SVM predictions varied more ($\sigma = 0.04$), especially when transferring to CQ_B and MC_K.

5.2.2 Task B

Results for the cross-dataset classification of argument components are presented in Figure 2. Both BERT model variants generalized very well and achieved an average macro F_1 score of 0.90 across the different datasets. With $\sigma = 0.01$, the undersampling model predicted remarkably stable on our datasets ($\sigma = 0.02$ for the non-undersampling model). SVM, ECGA and fastText strongly benefited from balanced training data. With undersampling, the latter two approaches could surpass the in-dataset results from Section 5.1 and thus achieved best values for all datasets. SVM struggled with generalization on MC_K and CQ_B ($\sigma = 0.03$). Likewise fastText showed some weaknesses in generalization ($\sigma = 0.03$), which were particularly noticeable in the performance drop on CQ_B (0.76) compared to the reference value (0.84). ECGA achieved more uniform results with an average macro F_1 value of 0.83 ($\sigma = 0.02$), which, however, do not come close to the high values of BERT.

It turned out that the models generalize surprisingly well across the different processes. In both tasks, BERT showed superior results, but other methods were also able to provide stable predictions across the different test datasets. This suggests that universally valid patterns of argument structures could be learned, generalizing to a very different data type (from deliberative online platforms to questionnaire data), as well as to a process with a more general topic (from specific cycling to a comprehensive mobility concept).

6 Conclusion and Future Work

We investigated (A) the distinction of ADUs and non-ADUs and (B) the classification of major posi-

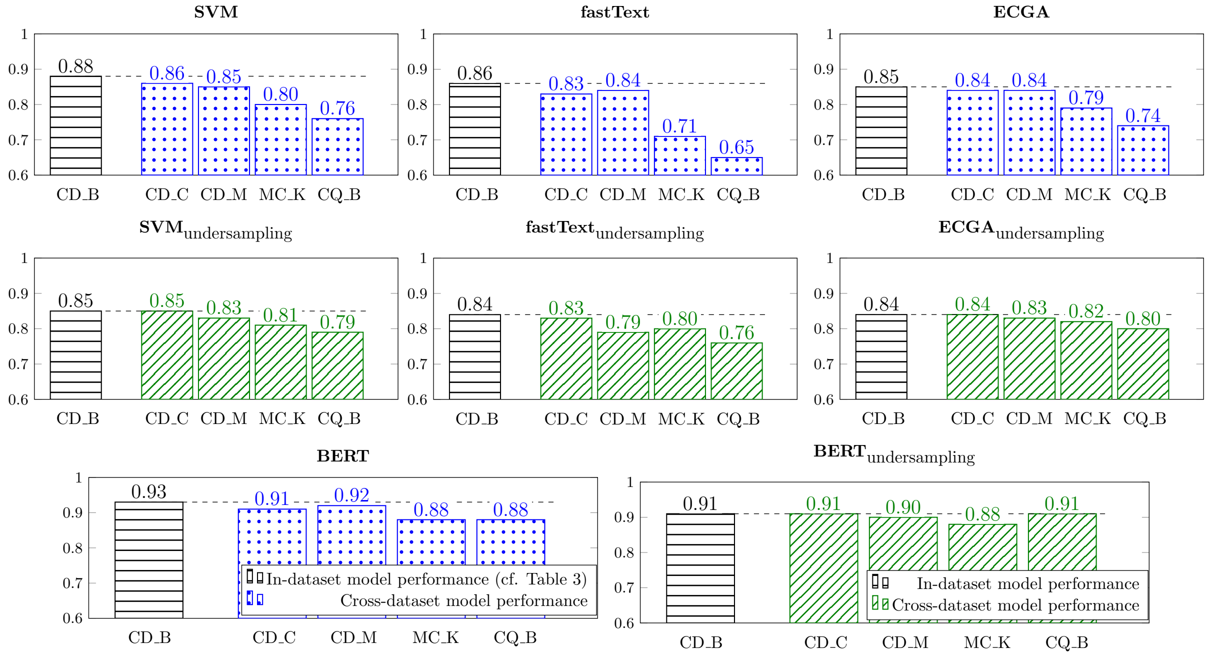

Figure 2: Cross-dataset evaluation for Task B. Results are averaged macro F_1 values of the five models trained on CD_B. (Note that in-dataset performance of CD_B with undersampling has not been reported in Table 3),

tions and premises for German public participation processes from urban planning. For this purpose, we introduced a new data corpus comprising five diverse mobility-related processes. Our first objective was to identify previously published approaches to solving the two classification tasks on public participation processes and test their performance on our datasets. Among these works, SVM achieved the best results in both tasks. Our second objective was to improve the previous results. We proposed the use of BERT and successfully demonstrated that the results of both tasks improved. On our datasets, BERT yielded highly promising macro F_1 scores, between 0.76 and 0.80 for Task A and between 0.86 and 0.93 for Task B. We additionally showed, that our approach outperforms previous results for Task B on a similar German online participation dataset. We further argued, that the use of pretrained models is one way to make argument mining applicable in municipalities. Our third objective was to prove the feasibility for processes from urban planning that differ in topic or format. We showed that BERT models outperform the other approaches, achieving average macro F_1 values of 0.77 ($\sigma = 0.02$) for Task A and 0.90 ($\sigma = 0.01$) for Task B in the cross-dataset evaluation. Our results are very positive and show that practical support for municipalities in evaluating mobility-related public participation processes is within reach by providing pretrained models.

In future work, we plan to investigate whether our best model can generalize to non-mobility public participation processes in urban planning to cover a broader range of topics. To further improve our models, we will concentrate on improving the detection of argumentative discourse units. Although we were able to achieve promising results, it has become apparent that distinguishing ADUs from non-ADUs is a particular challenge. Additionally, we will extend the classification for sentences that include multiple argument components (major position and premise) and address stance detection.

Acknowledgements

We are especially grateful to Tobias Escher, who provided significant support for this work through funding acquisition, the provision of data resources and comments to the manuscript. We thank him, Katharina Huseljić, Matthias Liebeck and Laura Mark for valuable discussions and feedback on earlier forms of this work. We thank 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 as part of its Social-Ecological Research funding priority, funding no. 01UU1904. Responsibility for the content of this publication lies with the author.

References

Stefanie Albert, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Sabine Brants, Thorsten Brants, Vera Demberg, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Hagen Hirschmann, Juliane Janitzek, Carolin Kirstein, Robert Langner, Lukas Michelbacher, Oliver Plaehn, Cordula Preis, Marcus Pussel, Marco Rower, Bettina Schrader, Anne Schwartz, Smith George, and Hans Uszkoreit. 2003. TIGER Annotationsschema. Technical report, Universität des Saarlandes, Universität Stuttgart, Universität Potsdam.

Miguel Arana-Catania, Felix-Anselm Van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. 2021. Citizen participation and machine learning for a better democracy. *Digit. Gov.: Res. Pract.*, 2(3).

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations*.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions. 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 2933–2943.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. Dataset independent baselines for relation prediction in argument mining. In *Computational Models of Argument - Proceedings of COMMA 2020*, pages 45–52.

Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning*, 20:273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Vlad Eidelman and Brian Grom. 2019. Argument Identification in Public Comments from eRulemaking. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 199–203.

Peter Esaiasson. 2010. Will citizens take no for an answer? What government officials can do to enhance decision acceptance. *European Political Science Review*, 2(3):351–371.

Constanza Fierro, Claudio Fuentes, Jorge Pérez, and Mauricio Quezada. 2017. 200K+ Crowdsourced Political Arguments for a New Chilean Constitution. In *Proceedings of the 4th Workshop on Argument Mining*, pages 1–10.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2018. Argumentative Link Prediction using Residual Networks and Multi-Objective Learning. In *Proceedings of the 5th Workshop on Argument Mining*, pages 1–10.

Athanasios Giannakopoulos, Maxime Coriou, Andreea Hossmann, Michael Baeriswyl, and Claudiu Musat. 2019. Resilient Combination of Complementary CNN and RNN Features for Text Classification through Attention and Ensembling. In *6th Swiss Conference on Data Science (SDS)*, pages 57–62.

Chinnappa Guggilla, Tristan Miller, and Iryna Gurevych. 2016. CNN-and LSTM-based Claim Classification in Online User Comments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2740–2751.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Barbara Konat, John Lawrence, Joonsuk Park, Katarzyna Budzynska, and Chris Reed. 2016. A Corpus of Argument Networks: Using Graph Properties to Analyse Divisive Issues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3899–3906.

Namhee Kwon and Eduard Hovy. 2007. Information Acquisition using Multiple Classifications. In *Proceedings of the 4th International Conference on Knowledge Capture*, pages 111–118.

Namhee Kwon, Stuart W. Shulman, and Eduard Hovy. 2006. Multidimensional Text Analysis for eRulemaking. In *Proceedings of the 2006 International Conference on Digital Government Research*, pages 157–166.

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*, pages 76–81.

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 Transactions on Internet Technology*, 17(3):1–22.

John Lawrence and Chris Reed. 2019. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

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 3rd Workshop on Argument Mining*, pages 144–153.

Nina A. Mendelson. 2012. Should Mass Comments Count? *Mich. J. Envtl. & Admin. L. 2*, 2:173–183.

Gaku Morio and Katsuhide Fujita. 2018a. Annotating Online Civic Discussion Threads for Argument Mining. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 546–553.

Gaku Morio and Katsuhide Fujita. 2018b. End-to-End Argument Mining for Discussion Threads Based on Parallel Constrained Pointer Architecture. In *Proceedings of the 5th Workshop on Argument Mining*, pages 11–21.

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.

OECD. 2004. *Promise and Problems of E-Democracy*. OECD Publishing.

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.

Joonsuk Park, Arzoo Katiyar, and Bishan Yang. 2015. Conditional Random Fields for Identifying Appropriate Types of Support for Propositions in Online User Comments. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 39–44.

Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Stephen Purpura, Claire Cardie, and Jesse Simons. 2008. Active Learning for e-Rulemaking: Public Comment Categorization. In *Proceedings of the 2008 International Conference on Digital Government Research*, pages 234–243.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578.

Julia Romberg and Tobias Escher. 2020. Analyse der Anforderungen an eine Software zur (teil-)automatisierten Unterstützung bei der Auswertung von Beteiligungsverfahren. Working Paper 1, CIMT Research Group, Institute for Social Sciences, Heinrich Heine University Düsseldorf.

Fritz W. Scharpf. 1999. *Governing in Europe: Effective and Democratic?* Oxford: Oxford University Press.

David Schlosberg, Stephen Zavestoski, and Stuart W. Shulman. 2008. Democracy and E-Rulemaking: Web-Based Technologies, Participation, and the Potential for Deliberation. *Journal of Information Technology & Politics*, 4(1):37–55.

Stuart W. Shulman. 2003. An experiment in digital government at the United States National Organic Program. *Agriculture and Human Values*, 20:253–265.

Peter Teufl, Udo Payer, and Peter Parycek. 2009. Automated Analysis of e-Participation Data by Utilizing Associative Networks, Spreading Activation and Unsupervised Learning. In *International Conference on Electronic Participation*, pages 139–150.

Christine Thielen and Anne Schiller. 2011. Ein kleines und erweitertes Tagset fürs Deutsche. In *Lexikon und Text: Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, pages 193–204. Max Niemeyer Verlag.

Hui Yang, Jamie Callan, and Stuart Shulman. 2006. Next Steps in Near-Duplicate Detection for eRulemaking. In *Proceedings of the 2006 International Conference on Digital Government Research*, pages 239–248.