Machine-assisted Text Classification of Public Participation Contributions

Inaugural-Dissertation

zur

Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Julia Romberg

aus Kaiserslautern

Düsseldorf, August 2023

aus dem Institut für Informatik der
Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Stefan Conrad
2. Jun.-Prof. Dr. Tobias Escher

Tag der mündlichen Prüfung: 22.11.2023

Erklärung

Ich versichere an Eides Statt, dass die vorliegende Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Desweiteren erkläre ich, dass ich eine Dissertation in der vorliegenden oder in ähnlicher Form noch bei keiner anderen Institution eingereicht habe. Ich habe keinerlei andere Promotionsversuche unternommen.

Düsseldorf, 17. August 2023 Julia Romberg

Acknowledgements

Throughout the process of writing this thesis, I have been fortunate to receive support from numerous people, to whom I would like to express my sincere gratitude. First of all, I would like to thank my two supervisors, Stefan Conrad and Tobias Escher. Stefan, you strongly encouraged me to enter research and contemplate a doctorate. Thank you for always sharing your vast knowledge with me. I have been beyond grateful for your support as I considered the move to Tobias' research group. Tobias, your guidance has allowed me to become a better researcher and author. Thank you for always taking the time to critically question my research and provide valuable input. Our engaging discussions were consistently motivating, as you have a wonderful talent for combining the mutual perspectives of the social sciences and computer science. I would also like to thank Gunnar Klau, who accompanied me as an official mentor.

Thanks to my colleagues in the social sciences, especially my amazing research group: Katharina Holec, Laura Mark and Nicole Rohan, I have cherished every moment of being a part of the team and having the opportunity to work alongside all of you. I learned a lot in our interdisciplinary collaboration, which helped me gain a broader perspective and deepen the focus of my work.

Thanks to my colleagues in computer science, especially from the databases and information systems chair: Alexander Askinadze, Fabian Billert, Kirill Bogomasov, Daniel Braun, Manh Khoi Duong, Sabine Freese, Thomas Germer, Philipp Grawe, Ludmila Himmelspach, Gerhard Klassen, Sergej Korlakov, Martha Krakowski, Matthias Liebeck, Lisa Lorenz, Pashutan Modaresi, Leonie Selbach, Michael Singhof, and Boris Thome. I am grateful for the many fun times, some of which we shared back when we were still young students, to our hilarious first joint trip together as a team to Stuttgart, to the evenings we spent together at the pub, and the collaborative hat-making for those who moved on over time. It was also a pleasure to ponder over research with you. Dear Guido Königstein, master of various annotation software, thank you for your technical support, which I could always count on even after I moved to another chair.

To my family and friends, thank you for your support and patience throughout my journey. I would especially like to thank my parents, who inspired me with their interest in science from an early age. Mama, your in-depth and patient manner of explaining "the world" to me will always remain of irreplaceable significance to me. Papa, while I may not possess the same enthusiasm for APL as you do, we do share a common fascination for numbers and statistics. Iris, Franzi and Johannes, I am very grateful for the valuable input you provided during the final stages of my dissertation.

Dear André, I've heard that being resilient is a must nowadays. The last few years have certainly helped you improve this skill. Thank you for everything!

Abstract

Engaging citizens in decision-making processes is a widely implemented instrument in democracies. Such public participation processes serve the goal of achieving a more informed procedure to potentially improve the process outcome and increase the public acceptance of decisions made. As public officials try to evaluate the often large quantities of citizen input by hand, they regularly face challenges due to restricted resources. For textual contributions, the most common form of citizen input, natural language processing offers the prospect of automatic support for evaluation. Still, many methods are inadequate due to insufficient accuracy, lack of robustness across datasets, or a neglect of important aspects of practical application. This thesis explores how existing research gaps can be overcome with text classification methods, focusing on the tasks of thematic structuring and argument analysis in the manual evaluation cycle.

We start with a systematic literature review of previous approaches to the machine-assisted evaluation of textual contributions. Given the identified shortage of language resources, we subsequently create a multidimensionally annotated corpus to facilitate the development of text classification models for German-language public participation.

Once the groundwork is laid, our initial focus is on the thematic structuring of public input, particularly considering the uniqueness of many public participation processes in terms of content and context. To make customized models for automation worthwhile, we leverage the concept of active learning to reduce manual workload by optimizing training data selection. In a comparison across three participation processes, we show that transformer-based active learning can significantly reduce manual classification efforts for process sizes starting at a few hundred contributions while maintaining high accuracy and affordable runtimes. We then turn to the criteria of practical applicability that conventional evaluation does not encompass. By proposing measures that reflect class-related demands users place on data acquisition, we provide insights into the behavior of different active learning strategies on class-imbalanced datasets, which is a common characteristic in collections of public input.

Afterward, we shift the focus to the analysis of citizens' reasoning. Our first contribution lies in the development of a robust model for the detection of argumentative structures across different processes of public participation. Our approach improves upon previous techniques in the application domain for the recognition of argumentative sentences and, in particular, their classification as argument components. Following that, we explore the machine prediction of argument concreteness. In this context, we account for the subjective nature of argumentation by presenting a first approach to model different perspectives in the input representation of machine learning in argumentation mining.

Zusammenfassung

Es ist ein weit verbreitetes demokratisches Instrument, die Öffentlichkeit in politische Entscheidungsprozesse einzubinden. Sogenannte Bürger*innenbeteiligungsverfahren haben zum Ziel, Entscheidungen durch Informationen zu stützen und ihre Akzeptanz zu erhöhen. Bei der manuellen Auswertung der oft großen Anzahl von Beiträgen stehen die Behörden aufgrund begrenzter Ressourcen jedoch regelmäßig vor Herausforderungen. Im Hinblick auf Beiträge, die in textueller Form vorliegen, können Methoden des Natural Language Processings die Auswertung automatisch unterstützen, doch noch sind diese oft unzureichend für den praktischen Einsatz. In dieser Dissertation wird erforscht, wie bestehende Forschungslücken mithilfe von Textklassifikationsmethoden überwunden werden können. Ein besonderer Fokus liegt dabei auf den Aufgaben der thematischen Strukturierung von Beiträgen und der Argumentationsanalyse.

Zu Beginn wird ein systematischer Literaturüberblick über bisherige Ansätze zur maschinengestützten Auswertung von Textbeiträgen gegeben. Angesichts des identifizierten Mangels an Sprachressourcen wird ein Datenkorpus für die Entwicklung von Textklassifikationsmodellen für deutschsprachige Öffentlichkeitsbeteiligung erarbeitet.

Nachdem die Grundlagen geschaffen sind, steht zunächst die thematische Strukturierung mit Fokus auf die inhaltliche und kontextuelle Einzigartigkeit von Verfahren im Mittelpunkt. Um den Einsatz individuell angepasster Machine Learning-Modelle lohnenswert zu gestalten, wird das Konzept des Active Learnings eingesetzt, um den manuellen Klassifikationsaufwand durch eine optimierte Trainingsdatenauswahl zu verringern. In einem Vergleich über drei Beteiligungsprozesse hinweg zeigt sich, dass die Kombination von Active Learning mit Transformer-basierten Architekturen den manuellen Aufwand bereits ab einigen hundert Beiträgen signifikant reduzieren kann, bei guter Vorhersagegenauigkeit und geringen Laufzeiten. Anschließend entwickeln wir Maße, um weitere praxisrelevante Anforderungen der Einsetzbarkeit zu evaluieren. Diese geben Einblick in das Verhalten verschiedener Active Learning-Strategien hinsichtlich klassenbezogener Eigenschaften auf den häufig imbalancierten Datensätzen.

Danach wird der Schwerpunkt auf die Analyse der Argumentation der Bürger*innen verlagert. Der erste Beitrag ist ein robustes Modell zur Erkennung von Argumentationsstrukturen über verschiedene Prozesse der öffentlichen Beteiligung hinweg. Unser Ansatz verbessert die zuvor in der Anwendungsdomäne eingesetzten Techniken zur Erkennung von argumentativen Sätzen und insbesondere zur Klassifikation von Argumentkomponenten. Zudem wird die maschinelle Vorhersage der Konkretheit von Argumenten untersucht. Hierbei tragen wir der subjektiven Natur von Argumentation Rechnung, indem wir einen ersten Ansatz zur direkten Modellierung verschiedener Perspektiven als Teil des maschinellen Lernprozesses des Argumentation Minings vorstellen.

Contents

1

Introduction

1.1 Motivation

An important instrument of democracy is the representation of citizens' attitudes and beliefs in political decision-making processes. Political participation can be defined as the behaviors citizens undertake voluntarily, alone or with others, with the goal of influencing political decisions (Kaase, 2000). The forms of political participation are multifaceted and range from elections and engagement in political parties to civic initiatives and dialogue-oriented formats (Theocharis and Van Deth, 2018).

In this thesis, we focus on the particular mechanism of *public participation processes*. Such processes are initiated and carried out by government agencies to consult the public on political issues (Bock and Reimann, 2021), and can be realized in various forms. Involvement can take place either through online platforms or through offline alternatives, such as on-site events or postal surveys. There are methods that emphasize deliberation and those that do not provide an avenue for discussion among participants. The target audience may include the general public, but may also be tailored to specific focus groups. Eventually, the numerous application scenarios of public participation cover diverse matters, ranging from urban planning (Damer and Hague, 1971) to water resource management (Priscoli, 2004) and even constitution drafting (Árnason and Dupré, 2020).

Empirical studies have shown that involving the public can indeed influence decision-making processes. In a study of local transportation planning in the city of Palo Alto, Chen and Aitamurto (2019) found that 46% of the more than 250 comments actually resulted in a change in policy. Another example is Iceland's 2011 crowdsourced constitution-drafting process, in which about 10% of the citizens' textual suggestions led to a modification of the draft constitution (Hudson, 2018). The impact not only pertains to the process itself, in which more interests are represented, but is also reflected in improved process outcomes. For instance, public participation in environmental decision-making can yield decisions of better environmental quality and adhering to higher environmental standards (Jager et al., 2019; Dietz and Stern, 2008).

Bobbio (2018) refers to three main reasons that drive policymakers to involve the public. First, participation as a learning process can enable more informed decisions based on the knowledge of the people. Second, participation can improve the legitimization of a process by helping people accept the outcome, even if it does not reflect their personal opinion. And third, participation can be utilized to empower the people in accordance with the literal meaning of democracy.

Achieving each of these goals requires a thorough analysis of the input gathered to tap into collective intelligence and provide feedback on ideas and process outcomes. This means that once the public has been consulted, the various statements made by the public must be collected and analyzed to make sense of them. Originally a purely manual task, the *evaluation of public participation contributions* consists of a number of steps that are either performed within public administrations themselves or outsourced to external service providers (Shulman et al., 2004; Maragoudakis et al., 2011; Romberg and Escher, 2020; Jasim et al., 2021; Simonofski et al., 2021).

First, human analysts read each contribution at least once, but often several times. During this phase, efforts are made to identify and collate contributions that are substantially identical. This allows mass campaigns to be detected to avoid undue influence on the process by individual stakeholders (Livermore et al., 2017; Shulman, 2009). What is more, summarizing congruent contributions also improves the clarity of the collection and thus simplifies further evaluation. In parallel, the indispensable step of organizing the contributions thematically takes place. Often, these thematic groups are aligned with administrative units, which then examine the contributions that are relevant to them. In this way, public administration also gains an initial overview of the issues that are of concern to the citizens. After the phase of data cleaning and pre-structuring, the detailed content analysis of the citizens' ideas follows. Attention is paid to the individual opinions on the topics of discussion, which arguments are put forward for or against planned policies, and which suggestions for improvement are made. Once the content has been evaluated in detail, conclusions can be drawn from the input and policy recommendations can be formulated.

As part of the overarching political decision-making process, the evaluation of input received must follow democratic norms to prevent negative effects such as anger and mistrust among the public (Innes and Booher, 2004). These norms include ensuring a fair and transparent decision-making in which all opinions are given equal treatment (Dahl, 1989). What is more, public perceptions of legitimacy are directly impacted by how public authorities assess citizen input (Schmidt, 2013). As a consequence, perceived non-compliance with the above criteria may cause the resulting policies to be seen as less legitimate (Esaiasson, 2010; Strebel et al., 2019). Therefore, the evaluation process as well as the decisions derived from it need to be comprehensible and justified. At the same time, however, policy-making efforts usually have to keep up with a prescheduled agenda. This means that contributions must not only be evaluated according to high democratic standards, but also in a timely manner to meet deadlines.

Keeping up with these stringent requirements poses significant challenges to the manual analysis. This is particularly due to the fact that public participation processes – offline and online – have the potential to generate large amounts of input. For instance, the participatory phase of Chile's 2016 constitutional process collected over 200,000 arguments from questionnaires, local deliberative on-site events, and provincial and regional councils (General Secretariat, Presidency of Chile, 2017). Electronic rulemaking initiatives also repeatedly encounter a tremendous response. Examples include the United States Department of Agriculture's National Organic Program, which received over 277,000 responses (Shulman, 2003), and the United States Environmental Protection Agency, which had over four million comments to process under the Clean Air Act (Livermore et al., 2017). Such volumes of data can be an insurmountable hurdle, as administrations often only have limited resources (both financial and personnel). And, more than that, regularly the evaluation of citizens' input becomes an overload even in processes with much lower participation, such as a few hundred contributions (Mahyar et al., 2019).

One way out of this dilemma is offered by *machine learning*: Automating sub-tasks in the evaluation process can support the success of public participation (Organisation for Economic Co-operation and Development, 2003; Livermore et al., 2017; Mahyar et al., 2019; Arana-Catania et al., 2021a; Jasim et al., 2021; Reynante et al., 2021; Simonofski et al., 2021). Machine learning can be described as "the technique that improves system performance by learning from experience via computational methods" (Zhou, 2021). Its concept is inspired by human learning behavior, which is based on continuously accumulated experience. Machine learning, for its part, gains its knowledge through information from data collections. This thesis specifically deals with textual data, as this is the most common way in which citizens contribute. To this end, we make use of *natural language processing*, a field of research in which computational methods are used to model human language with the aim to learn, understand, and also generate it. The most effective strategies of natural language processing nowadays have their roots in machine learning, such as the famous transformer-based language models (Devlin et al., 2019; Radford et al., 2019).

Literature has mainly proposed three starting points for the machine-assisted analysis of citizens' textual content. Yang and Callan (2005) suggested to automatize the *detection of duplicates*. Furthermore, *thematic structuring* of the unorganized collection was identified as a potential opportunity for machine support (Kwon et al., 2006; Cardie et al., 2008a; Yang and Callan, 2009; Teufl et al., 2009). Lastly, there was interest in using machine learning to assist the more in-depth content analysis of citizens' ideas. The main focus was on *analyzing the arguments* made by the public on the issue at hand (Kwon et al., 2006; Park and Cardie, 2014; Lawrence et al., 2017).

Unfortunately, many of these natural language processing approaches to public participation are not yet practical because they either cannot provide sufficient accuracy, do not perform robustly across datasets, or neglect important aspects of practical application. For this reason, there is an urgent need to develop and improve machine learning methods to help public administration evaluate citizen participation efforts for policy-making. In the following section, we establish the research focus of this thesis by discussing the shortcomings of current approaches in more detail.

1.2 Research Goal

This thesis focuses on supervised machine learning, where methods gain their experience from annotated training data. More precisely, we investigate *classification algorithms* to support the evaluation of textual public contributions. Given a fixed set of classes (hereinafter also referred to as categories) and a set of (manually) labeled training documents, a classification function is learned that maps documents to classes using a learning algorithm. It is referred to as *text classification* when the documents are composed of written text, such as books, news articles, as well as more compact units like paragraphs or single sentences (Manning et al., 2008). If every document is assigned to exactly one class, we speak of *single-label classification*, whereas in *multi-label classification*, a document can be assigned to any subset of the classes.

Text classification algorithms are evaluated in different ways depending on the problem at hand. In this work, we consider two well-known measures. The *F_1 score* is the harmonic mean of precision, the proportion of correct predictions out of the total number of documents that truly belong to a class, and recall, the proportion of correct predictions out of the total number of documents that a learned model assigns to a class. It judges algorithmic performance at a class-wise level and provides important insights especially for imbalanced datasets where classes are represented with different frequencies, a property that holds for all datasets considered in this thesis. To condense the individual class scores into a global indicator of performance, we use the macro-averaged F_1 score which computes the arithmetic mean. The second measure we rely on is *accuracy*, which per se evaluates the performance of a learning algorithm globally by giving the overall percentage of correct class predictions. While the original definition refers to single-label classification tasks, several definitions have been proposed to simulate accuracy in the multi-label case. We opt for the micro-averaged F_1 score, a global average which matches the definition of accuracy when predicting exactly one class.

The main goal of this thesis is to advance the two most prominent sub-tasks of evaluation whose automation can benefit public participation. These are the thematic structuring of public participation data and the more in-depth analysis of citizens' arguments. While the third sub-task, the detection of duplicates, was already researched with considerable success (Yang and Callan, 2005; Yang et al., 2006; Yang and Callan, 2006), prior work in both thematic structuring and argument analysis reveals gaps that need to be closed for their beneficial practical application. The need for viable machine support in both sub-tasks is also reflected in interviews we conducted in 2020 with government agencies, external service providers, and planning officers (Romberg and Escher, 2020), further underscoring the relevance of this thesis.

1.2.1 Thematic Structuring of Citizen Contributions

Structuring citizen contributions thematically can be framed as a text classification problem[1]. We will refer to this task as *topic classification* in the following. Multiple works have investigated the performance of text classification algorithms for the topical categorization of public input (e.g., Kwon et al., 2006; Cardie et al., 2008a,b; Purpura et al., 2008; Aitamurto et al., 2016; Fierro et al., 2017; Balta et al., 2019; Giannakopoulos et al., 2019; Kim et al., 2021). The results are mixed because there are many influencing variables, such as the classification schemes used, the number of classes, and how imbalanced these classes are. Giannakopoulos et al. (2019), for example, achieved an accuracy of 0.75 using a combination of different deep neural networks, while in the application of Balta et al. (2019) vanilla BERT (Devlin et al., 2019), a modern language model, achieved an accuracy of 0.68. The still consider-

[1] An alternative approach is the use of unsupervised machine learning, see for example Yang and Callan (2009), Teufl et al. (2009), Hagen et al. (2015), Hagen (2018), or Arana-Catania et al. (2021a).

able proportion of classification errors illustrates that the topic classification of public participation contributions is a challenge even for state-of-the-art methods.

Regardless of the reported classification performance, all these works share the assumption that a sufficient pool of training data is available. Across all works, the algorithms have been trained with a large number of manually pre-labeled contributions, sentences or arguments. At the same time, however, there is a consensus that the topics discussed are highly dependent on the individual public participation processes and therefore require specifically tailored classification systems.

For practical application, this implies that a substantial share of process input usually needs to be manually classified before a suitable machine learning model can be trained to support the human analysts with the remaining contributions. Depending on the uniqueness of the process, the size of the collection, the number and distribution of classes, and the additional effort needed to train a model, in the worst case, therefore, machine support may not lead to any time savings or work relief in the sub-task of thematic structuring. It was only Purpura et al. (2008) who explicitly acknowledged this problem and proposed the use of active learning as a solution.

Active learning (Cohn et al., 1996; Settles, 2009) is a collaborative process between human and machine. It approximately solves the optimization problem of identifying a minimal subset of training data that is capable of learning a best classification function for maximizing the prediction performance. Such small but informative sets of training examples are identified through targeted *query strategies*, also called acquisition functions. Active learning proceeds according to the following scheme: Using some query strategy, a batch of examples is selected from the accessible quantity of unlabeled data. These examples are then labeled by an oracle (e.g., a human annotator) and moved to the pool of labeled training data. Finally, a model is fit to the training dataset. This process is repeated until a predefined stop criterion is met (e.g., a given annotation budget is exhausted or a satisfactory model accuracy is reached).

Purpura et al. (2008)'s proposal to optimize the training data selection while maintaining the advantages of automation was effective. Experiments demonstrated a noticeable reduction in manual effort for topic classification in the dataset under consideration. However, it became apparent that the learning algorithms still needed nearly a thousand training examples to reach the maximum accuracy of 0.70 (a classification accuracy comparable to many of the studies introduced above).

Despite this strong push into a promising direction of research, to the best of our knowledge no follow-up work has been published since with the goal of supporting the evaluation of public participation processes. Yet, the practical usefulness of topic classification with active learning has not been conclusively clarified due to the still considerable manual labeling efforts reported as well as the notable number of misclassifications. In this thesis we therefore pursue and advance the active learning approach for topic classification of public input to promote practical applicability. Our focus is on the performance of current state-of-the-art methods with respect to training data reduction, classification quality, and practice-oriented evaluation criteria.

1.2.2 Analysis of Citizen Arguments

It is critical to identify not only the topics being discussed, but also the thoughts and considerations the public is having about them. The computational approach to reasoning is *argumentation mining* (also referred to as argument mining), a field of research in natural language processing that strives to automatically identify and classify argumentative structures in natural language data. In recent years, the focus has widened from discovering basic argument structures to more challenging tasks such as evaluating the quality of arguments and synthesizing argumentative texts. Citizens' reasoning is a key indicator of public opinion, and for this reason a significant body of research supporting the evaluation of public participation through machine learning has looked at argumentation mining.

Works such as Kwon et al. (2006), Liebeck et al. (2016), and Morio and Fujita (2018b) were concerned with finding and classifying the basic building blocks of arguments. To this end, they followed variations of the claim-premise model of argumentation (based on Freeman, 1991). In this model, an argument consists of a controversial statement (i.e., a claim) and a set of reasons supporting that statement (i.e., premises) (Stab and Gurevych, 2014; Besnard and Hunter, 2008). The detection of stances, i.e., whether someone is taking a position in favor or against a statement, received some attention as well (Konat et al., 2016; Lawrence et al., 2017; Liebeck, 2017).

There has also been a push to classify claims according to certain characteristics (Park et al., 2015; Niculae et al., 2017; Fierro et al., 2017). These studies mostly build on a three-part scheme of facts (claims that are verifiable with objective evidence), values (claims that indicate preferences, interpretations, or judgments), and policies (claims that propose a course of action) (Hollihan and Baaske, 2022; Snider and Schnurer, 2002; Branham, 2013). Furthermore, the verifiability of citizens' propositions has been researched in more detail (Park and Cardie, 2014).

The literature demonstrates encouraging results for the identification and classification of argumentation structures and characteristics. While many of these approaches were not yet mature for real-world usage, the consistently improving field of argumentation mining offers promising advancements. As part of this thesis, we will complement previous research efforts by exploring aspects that constitute a further step towards the practical application of argumentation mining in public participation.

Specifically, we will research the generalizability of argumentation mining models across datasets. This essential characteristic has received little attention so far. However, trained models are only usable for public authorities and service providers if they produce reliable results for new use cases.[2] What is more, we will focus on how concrete citizens are in formulating their propositions. Such quality of argumentation can be of interest to analysts and is already carried out manually in some cases of public participation evaluation. Sorting ideas according to how concrete they are can, for instance, help to process more ideas in a shorter period of time as it is easier to derive actions or policies to be implemented from more specific input. A special challenge arises from the subjective perception of concreteness, which we will also address.

1.3 Contributions

This thesis presents a number of contributions. In the following, we list these and explain their value to the domain of public participation, as well as their relevance to

[2] In contrast to the specifics of topic classification outlined earlier, the classes to be predicted remain the same in argumentation mining. The goal is therefore to create a universal model to eliminate the difficult and tedious manual labeling of argumentation in the future.

research in active learning, argumentation mining, and text classification.

i) **Systematic review and development of a research agenda.** We conducted a systematic literature review of the field of machine-assisted evaluation of textual contributions to public participation processes (Chapter 2, publication R5).

   In doing so, we addressed the lack of an overview of the current state of research in this interdisciplinary field that links the two disciplines of computational linguistics and policy informatics. We set out why supporting the evaluation of public contributions through natural language processing should be recognized as a research field in its own right and we outlined the established sub-tasks. Based on the review, we identified research gaps that need to be filled in order to successfully apply natural language processing in the highlighted field. Finally, we developed a practice-oriented research agenda that provides recommendations for future work.

ii) **Multidimensional corpus annotation to facilitate the development of models for evaluating German-language contributions.** We created the *CIMT PartEval Corpus*, a new publicly-available German-language corpus that comprises several thousand citizen contributions from six mobility-related planning processes in five German municipalities (Chapter 3, publication R6). It was released under the Creative Commons CC BY-SA License and can be downloaded from GitHub.[3]

   The building of models through supervised machine learning relies on annotated data. Such language resources are scarce for our application domain, especially in languages other than English. The CIMT PartEval Corpus therefore provides annotations for approaching the following tasks of evaluation: i) the recognition of argument components and their classification, ii) the assessment of the concreteness of arguments, iii) the detection of textual descriptions of locations in order to assign citizens' ideas to a spatial location, and iv) the thematic categorization of contributions according to a generic schema of mobility. We added to solving the four tasks as follows: Our dataset for task i) contains seven times more sentences than other existing German corpora. In contrast to prior work, we included multiple public participation processes that differ in format and process subject to help evaluating how robustly machine learning models generalize to new data. Regarding task ii), we were the first to provide annotations for machine learning the concreteness of arguments. The created dataset for task iii) established a new application domain for text-based document geo-location[4] that differs from previously targeted genres in document length, text quality, and prevalence of location. For solving task iv), we developed a comprehensive categorization schema of mobility. The annotated documents can serve as the basis for training topic classification models that may be universally applied to a variety of mobility-related planning processes.

[3] The four sub-corpora can be downloaded from the following GitHub repositories: The CIMT Argument Components sub-corpus is available at `https://github.com/juliaromberg/cimt-argument-mining-dataset`, the CIMT Argument Concreteness sub-corpus is available at `https://github.com/juliaromberg/cimt-argument-concreteness-dataset`, the CIMT Geographic Location sub-corpus is available at `https://github.com/juliaromberg/cimt-geographic-location-dataset`, and the CIMT Thematic Categorization sub-corpus is available at `https://github.com/juliaromberg/cimt-thematic-categorization-dataset`.

[4] Text-based document geo-location is the task of determining the geographic coordinates of a document's associated location by its textual content.

iii) **Case study of topic classification with active learning.** In a comparison of current approaches to text classification with active learning on three datasets from online participation processes in German municipalities, we answered three practice-relevant research questions, namely what classification accuracy can be achieved, how much manual labeling effort can be saved through active learning, and how time-efficient the different approaches are (Section 4.1, publication R4).

While text classification and active learning have evolved greatly in recent years, this research strand has not received attention in our application domain since the work of Purpura et al. (2008). Considering more recent approaches, we showed that combining a BERT classifier with the active learning strategies Contrastive Active Learning and Maximum Expected Entropy improves the classification accuracy of previous approaches by 7% on average while saving up to 80% of the training data volume. In the best case, this translated into a need for only 120 training documents. Moreover, the models operated within an efficient runtime. Our approach dramatically cuts the time required for evaluation from which in particular processes with a larger number of contributions benefit. However, it also allows the application of automated topic classification to processes that only generate a few hundred contributions, a recurrent use case that was previously intractable.

iv) **Practice-relevant measures for active learning in topic classification scenarios.** We developed four measures that reflect class-related demands users may place on data acquisition when using active learning for topic classification (Section 4.2, publication R2).

Typically, active learning strategies in text classification tasks are evaluated and compared based on their accuracy or F_1 performance. However, this lab scenario neglects further criteria that are relevant for a successful transfer to practice. In public participation datasets characterized by an often increased number of imbalanced topic categories, these include, in particular, class-related characteristics.

Applying our measures on a range of text classification datasets, we demonstrated that pure reliance on accuracy and F_1 score in selecting a best query strategy cannot account for the requirement of full class coverage that is crucial for practical deployment. Furthermore, we were able to analyze the potentially desirable behaviors of favoring minority classes, covering the topic classes as quickly as possible, and class diversity in the selected annotation batches across various strategies of active learning. Our measures offer a promising starting point for refining existing techniques to better fulfill practical requirements in topic classification scenarios.

v) **Robust argument component mining for public participation.** We conducted a comprehensive evaluation of machine learning methods across five public participation processes in German municipalities that differ in format (online participation platforms and questionnaires) and process subject in order to build models for the identification and classification of argument components that generalize across datasets (Section 5.1, publication R3).

We first showed that fine-tuned BERT models surpass previously applied argumentation mining approaches for public participation processes on German data for both tasks, reaching macro F_1 scores of between 0.76 and 0.80 for the identification of argumentative units and macro F_1 scores of between 0.86 and 0.93 for their classification as premise or major position. In a cross-dataset evaluation, we then highlighted the robustness of our models: If trained on only one of the five public participation processes under consideration, they could recognize argument structures in the remaining datasets with comparable goodness of fit despite differing formats and process subjects. Such model robustness constitutes an important step towards the practical application of argumentation mining in municipalities.

vi) **Classification of argument concreteness.** We introduced the first study on the automated classification of argument concreteness (Section 5.2, publication R1).

Aspects of argument quality have received increasing attention in recent years. However, the level of concreteness in argument components' content remained understudied, despite it being an important characteristic in different applications of argumentation mining. One example is the evaluation of public participation where imprecise ideas are more laborious to evaluate. Automatically predicting how concrete citizens' premises and conclusions are can thus assist the human analyst to prioritize such contributions that are more easy to process.

We proposed a classification according to three levels of concreteness (low, intermediate, and high). By comparing a number of algorithms for text classification and different feature sets, we revealed the challenge of this task. Modern transformer-based models achieved only a macro F_1 score of 0.67 on the heavily imbalanced dataset and an accuracy of 0.79. These first findings form the foundation and indicate the need for further research on argument concreteness.

vii) **A multi-perspectivist approach for subjective classification tasks in argumentation mining.** We introduced the first approach in the field of argumentation mining to represent multiple perspectives in the input of machine learning processes. The novel method adds subjectivity information to the conventional text classification workflows of ground truth prediction (Section 5.2, publication R1).

It is common practice in machine learning to build models on aggregated ground truth. Regarding classification tasks that are considered to be subjective, this approach cannot do justice as it neglects individual perspectives. This also includes many parts of argumentation, especially when it comes to properties of arguments or their quality. Applying our method to the subjective task of argument concreteness, we found that text length is a strong indicator of subjectivity. Moreover, pre-trained language models do not yield a significant advantage over traditional algorithms, namely support vector machines, random forests and logistic regression, in terms of accuracy. We showed that the subjective perception of argument concreteness can be assessed with an accuracy of 0.74 respectively 0.52 (two or four levels of subjectivity) and with an F_1 score of 0.72 respectively 0.42. The results show that machines can, at least to some degree, learn to predict the subjective nature of arguments regarding concreteness.

1.4 Structure of the Thesis

This thesis is organized into chapters, each of which emphasizes a specific aspect and includes one or more publications that are thematically interrelated. In each case, the publications are embedded in the overall context of the dissertation topic, and the author's personal contribution is credited. We draw on research papers that have been accepted to appear in international peer-reviewed conferences, workshops and journals. A complete list of publications that are part of this dissertation, as well as further publications not directly associated, can be found in the Appendix *Publications of the Author*.

In Chapter 2, we start with a thorough survey of the state of research on machine-assisted evaluation of textual contributions in public participation processes by organizing current approaches to sub-tasks of evaluation. We explore the benefits and drawbacks of prior work and provide an agenda for future studies that will help the field move forward. We then introduce the CIMT PartEval Corpus in Chapter 3 in light of the findings from the literature review, which demonstrated the lack of annotated datasets for developing text classification methods, particularly for supporting a number of evaluation sub-tasks in languages other than English. Afterwards, we turn to the two main pillars of this thesis. Chapter 4 covers the topic classification of public participation contributions. In a comprehensive comparison of approaches on German public participation processes, we demonstrate the potential of active learning to reduce the amount of training data, and thus human effort, while maintaining high classification accuracy and efficient runtime. Subsequently, we define evaluation measures that reflect practice-relevant requirements for topic classification in an active learning scenario and provide insights into the behavior of different active learning strategies on participation data by applying our measures. Chapter 5 concentrates on argumentation mining in public input. We develop robust methods to detect and classify argument components across diverse participation processes. We then introduce the first approach to predicting subjective perceptions of the concreteness of arguments. In Chapter 6, we present a summary of the findings from this thesis and draw general conclusions. We close with an outlook on future work.

2

Existing Research Gaps

Starting in the late 1990s, a number of research projects have investigated selected aspects of supporting the evaluation of participation processes with methods of natural language processing. Some notable examples are the multi- and transdisciplinary *Cornell eRulemaking Initiative* (Cardie et al., 2006) and the *eRulemaking Research Group* (Shulman et al., 2005), a task force formed by several U.S. university working groups. In addition to these large-scale research projects, there is a body of relevant work conducted by smaller initiatives and further working groups, such as the PhD programme "Online Participation" funded by the State of North Rhine-Westphalia, Germany[1].

In this chapter, we lay the foundation for the later stages of the thesis by reviewing previously published work. A particular challenge in sifting through prior approaches arises from the fact that the research field itself is not clearly delineated. Relevant work has thus been published scattered across different research areas and can be found in the various publication formats of policy informatics, digital government, computational linguistics, natural language processing, machine learning and artificial intelligence. This makes it difficult to gain a comprehensive overview.

Existing literature reviews mostly looked at the field from a very government-oriented but rather non-technical point of view, trying to develop recommendations for the use of artificial intelligence in the public sector (Suominen and Hajikhani, 2021; Wirtz et al., 2019; Zuiderwijk et al., 2021). The only overview paper we are aware of that specifically focuses on the technological side is outdated and covers solely the task of sentiment analysis (Maragoudakis et al., 2011). Our first step in this thesis is therefore to provide an up-to-date summary of existing natural language processing methods for the evaluation of public participation, which has been lacking so far.

[1] https://www.fkop.de/en/

Paper: Julia Romberg and Tobias Escher. Making Sense of Citizens' Input through Artificial Intelligence: A Review of Methods for Computational Text Analysis to Support the Evaluation of Contributions in Public Participation. *Digital Government: Research and Practice*, just accepted, 2023.

Personal Contribution: The systematic literature search, as well as the analysis and assessment of the methods discovered, were carried out by Julia Romberg alone, making her the primary contributor. Tobias Escher and she jointly drafted the manuscript and together, they derived the research gaps from the results of the literature review and developed a research agenda.

Status: published

Based on a systematic literature search in two popular databases on computational linguistics and digital government, in this paper we detail the state of research on supporting the evaluation of public participation contributions through natural language processing. We observe that the approaches developed so far pursue three core objectives. These are the detection of duplicate contributions, grouping the contributions by topic, and gaining deeper insights into the individual contributions through the analysis of arguments, sentiment, and discourse, as well as comment summarization.

Most of the literature dealt with the grouping of contributions by topic, as well as with the analysis of arguments and opinions. Although there are several promising approaches, we reveal that there are still significant obstacles to overcome before most of these could provide any reliable support in practice. In many cases, the performance of the algorithms was not yet convincing. What is more, there was a very strong focus on English datasets and the development of monolingual models, while many other languages were left out of the equation. Finally, the full development cycle starting from the development of natural language processing methods and ending with ready-to-use user applications for the public sector often remained incomplete. Consequently, the results of research mostly did not find use in actual public participation processes and thus had no practical relevance.

We identify a number of directions for future research that could eventually result in practical answers. Besides the creation of non-English language resources and putting effort into the integration of methods into everyday work of experts, this entails developing methods based on state-of-the-art transformer architectures (which have had little application to date) along with more robust models that work reliably and consistently across datasets. We further infer that the most promising approaches incorporate the expertise of human evaluators, such as active learning in topic classification.

Making Sense of Citizens’ Input through Artificial Intelligence

Making Sense of Citizens’ Input through Artificial Intelligence

A Review of Methods for Computational Text Analysis to Support the Evaluation of Contributions in Public Participation

Julia Romberg[*]

Heinrich Heine University Düsseldorf, julia.romberg@hhu.de

Tobias Escher

Heinrich Heine University Düsseldorf, tobias.escher@hhu.de

Public sector institutions that consult citizens to inform decision-making face the challenge of evaluating the contributions made by citizens. This evaluation has important democratic implications but at the same time, consumes substantial human resources. However, until now the use of artificial intelligence such as computer-supported text analysis has remained an under-studied solution to this problem. We identify three generic tasks in the evaluation process that could benefit from natural language processing (NLP). Based on a systematic literature search in two databases on computational linguistics and digital government, we provide a detailed review of existing methods and their performance. While some promising approaches exist, for instance to group data thematically and to detect arguments and opinions, we show that there remain important challenges before these could offer any reliable support in practice. These include the quality of results, the applicability to non-English language corpora and making algorithmic models available to practitioners through software. We discuss a number of avenues that future research should pursue that can ultimately lead to solutions for practice. The most promising of these bring in the expertise of human evaluators, for example through active learning approaches or interactive topic modelling.

CCS CONCEPTS • Computing methodologies→Artificial intelligence→Natural language processing • Applied Computing -> Computing in Government •

Additional Keywords and Phrases: policy analytics, citizen participation, computational linguistics

[*] Corresponding author

1 THE ROLE OF PUBLIC PARTICIPATION FOR POLICY-MAKING

Democratic governments around the world rely increasingly on public participation of citizens in order to inform policy processes. In such public participation processes citizens are invited to make contributions on particular issues which subsequently need to be evaluated in order to derive specific measures. These procedures can take various forms such as written statements to planning procedures, oral statements during a public hearing on a proposed development, or proposals located on a digital map through an interactive online platform. In contrast to citizen-led initiatives such as petitions, expressions of political opinions (through discussions and demonstrations) or political consumerism, these top-down consultations allow authorities substantial control of the process through determining the design and organizational framework. What is more, they have a specific (even if often weak) link to decision-making processes that is regularly codified in law. Nevertheless, given that contemporary large-scale democracies are representative in nature with only limited opportunities for citizens to engage in decision-making directly, the role of public participation remains largely consultative. Public participation acts primarily as one of many sources of input (albeit a particularly important one) for those people who are legitimized (e.g. through elections) to take final decisions. Public authorities may utilize public participation to elicit input for different stages of the policy-making process, most regularly for agenda-setting, policy formulation and decision-making [82]. Generally, they pursue two distinct but related aims [77]: On the one hand, through the additional information acquired by such procedures, the resulting policies should be better informed and provide better adapted solutions, therefore ideally resulting in more effective policies. On the other hand, enabling citizens to provide knowledge, voice their concerns and (to some degree) shape the final policies, are expected to achieve higher acceptance if not satisfaction with the decisions, hence ideally resulting in higher legitimacy of the policies. Especially in response to heightened concerns about citizens' (dis)satisfaction with the way democracy works, such public participation has been increasingly used by authorities around the world and at all levels of government, taking various shapes, from simple invitations to comment, to large-scale deliberative events [22].

Policy-makers that aim to incorporate the knowledge and attitudes of citizens to inform their policy decisions face a number of challenges, such as whom to include in such consultations, how to design the process in order to achieve the desired outcomes and how much control citizens should wield over the process and its results – many of which have not yet conclusive answers. We focus on one particular challenge, which is the processing of the collected data by the authority responsible. Policy-makers and their administrations regularly face the problem of how to make sense of the diversity of statements that the public provides [1,2,37,52,54,71,82]. It involves both identifying overarching patterns and individual statements requiring further action to ultimately prepare conclusions from the input [52]. We call this process the *evaluation of public participation contributions*.

The relevance of this evaluation process can hardly be overstated. For example, basic democratic norms require that all citizens and their contributions are treated equally and that the process of decision-making is fair and transparent [20]. The way in which public authorities evaluate the input from citizens has direct consequences for public perceptions of legitimacy [77]. Empirical research has shown that if the public believes that the evaluation fails these criteria, this translates into lower legitimacy perceptions of the resulting policies [24,58,87]. What is more, in more formal participation procedures, public sector authorities may face costly litigation if they fail to identify and respond to substantive input by the public [52].

Hence public authorities have to dedicate care to the evaluation process in order to ensure that these normative criteria are satisfied and all contributions by citizens receive equal scrutiny. However, authorities are faced with the problem that evaluation takes considerable effort. It is regularly time-consuming, often requires substantial resources in terms of staff and money, and can lead to information overload [2,16,52,63]. When authorities do not have sufficient resources to engage in these efforts, they might choose evaluation strategies that do not satisfy democratic norms or decide to refrain from engaging the public altogether. Therefore, finding ways to support this evaluation process is of crucial importance, not least because it can be the decisive factor for authorities to engage or not to engage the public at all.

While there are different potential solutions to this problem such as increasing staff or using more structured participation formats, we focus on technological solutions in the form of computer-supported analysis procedures. While we believe that due to the often contested nature of public participation and its potentially far-reaching consequences, evaluation always requires some form of human assessment [56], the question is to what degree these human evaluators can be supported in their work. Technical means have long been proposed as one potential solution to this problem [63] and in the meantime, natural language processing (NLP) has made huge advances. These artificial intelligence (AI)-based techniques could be applied to the evaluation as the majority of contributions in public participation are in the form of textual data. However, despite early research efforts dating back almost 20 years, so far we lack an overview of which of the available computational methods have already been applied to the evaluation of public participation and how these have performed. What is more, within the burgeoning field of AI and public policy, supporting the evaluation of contributions by the public through NLP is not yet recognized as a research field in its own right and relevant research is widely dispersed across different fields and disciplines.

Therefore, the key objective of this paper is i) to identify generic tasks in the evaluation process and how these could be supported through the use of AI, ii) to summarize which approaches relying on computational text analysis have been used so far and to provide an assessment of their performance, and iii) to identify remaining gaps to inform future research efforts that could ultimately lead to solutions that offer reliable support in practice and hence make democratic participation possible. While we rely on a systematic literature review, our aim is not to conduct a detailed census but to provide an overview of the state of the field along with its strength and weaknesses.

The remainder of this paper is structured as follows. After briefly reviewing the state of the field (2), we describe our research methodology (3) and identify the tasks involved in the evaluation process and how these might be supported through automated procedures (4). The main body of this study focuses on reviewing approaches to topical grouping of content (5) and to extraction of arguments and opinions (6). We then summarize and discuss the main findings of the review and identify gaps that should be addressed by future research (7) before we draw a resume and offer reasons for the existing gaps in research (8).

2 SUPPORTING EVALUATION THROUGH COMPUTER-SUPPORTED TEXT ANALYSIS: RELEVANCE AND STATE OF THE FIELD

The task of evaluation is to make sense of citizens' contributions. These contributions derive from different sources and can take different forms. In offline public participation procedures citizens are asked for their opinion within on-site events or with tools such as questionnaires. Another source of contributions are online public participation procedures in which citizens have the opportunity to communicate their viewpoints via internet platforms. While citizen contributions can take many different forms, we focus our attention on textual data that might be derived from written statements from citizens, either created digitally or later digitized. Although by no means the only format, we believe these to be those most regularly used.

When public agencies have collected input from citizens, this needs to be analyzed. The overarching aim of such an analysis is to get to know which issues are raised and to decide if this input should trigger further action, such as a response or a change of the proposed plan. This requires reading each contribution, often several times, and as a result, the process of evaluating citizen contributions can take a significant amount of effort. How much effort depends on the number of citizens who participate as well as the amount and the length of contributions. Historically, there are numerous instances of offline participation that have resulted in large amounts of data. For example, when in 1997 the United States Department of Agriculture (USDA) launched its public comment period on standards for the marketing of organic agricultural products, the majority of the more than 277,000 comments were received via paper mail [79]. However, the ease and velocity afforded by information and communication technologies (ICT) has enabled more people to submit more statements in shorter time. Coupled with increasing relevance of public participation, there are now more instances of public participation, that each tend to receive more comments than in the pre-digital era. Livermore et al. [52] provide an overview of this development for the particular case of US e-rulemaking that eventually resulted in "megaparticipation" such as the US Environmental Protection Agency receiving more than 4 million comments for the proposed Clean Air Act. This development has increased the administrative burden and hence the urgency of the problem [52,80].

From early on, ICTs have not just been perceived as one cause of the problem but also as a possible solution to tackle the evaluation problem. For example, in 2003 an OECD [63] report highlighted the analysis of e-contributions as a challenge that might be solved through the use of content analysis techniques that help to structure contributions. Already in the late 1990s, in response to a growing number of comments on regulatory rules [19,79] the National Science Foundation among others had funded research such as the Cornell eRulemaking Initiative (CeRI) or the Penn Program on Regulation that investigated the potential to use text-processing techniques to sort through public comments [79,81,97]. This has sparked a remarkable research activity that resulted in the development of functionalities for searching similar content [79], duplicate detection [97], categorizing comments [12] and relating these to regulations [47]. Yet, as far as we know, these have not moved beyond the stage of prototypes and they have never experienced sustained use in public administration.

Since then, not only have government consultations and other instances of public participation increased, also the technology in the form of AI has made vast improvements. In particular the progress of machine learning algorithms has increased the capabilities of NLP, an "area of computer science that deals with methods to analyze, model, and understand human language" [91:4] and as such is of relevance to the evaluation task. As a matter of fact, the public sector now regularly applies AI to large amounts of data in order to derive insights for different stages of the policy-making process [95,104]. However, so far, we lack an overview of which specific technologies have been used to analyze citizen contributions and how these perform in comparison to established human evaluation. The only review that we know of that focuses on the technology is outdated and incomprehensive [55].

While the more recent advances in NLP techniques such as pre-trained language models have shown remarkable results on a variety of application tasks such as text translation and conversational agents (most recently through the release of ChatGPT) and in different application domains, they have yet to demonstrate their value for the input from public participation as these texts exhibit a number of differences from other domains. For example, tweets or other social media contents are not only shorter than citizen proposals but have also been shown to use a different vocabulary and syntax [31], not least demonstrated by the fact that specially trained models exist for this particular domain [e.g. 61]. Also, the contributions from public participation usually revolve around making proposals and deliberating about different possible solutions. As such their content differs from the contributions in comment sections on news portals, product reviews or online discussion groups which are primarily used to voice opinions and sentiments. The specific properties of public participation data lead us to believe that existing breakthrough are not necessarily delivering the same results in this domain.

Given the need for support of the evaluation process, we believe it is urgently required to take stock of this field by reviewing the strengths and weaknesses of those approaches that have been used and by offering guidance for further research. Here we focus mainly on the technological basis to offer an assessment of whether NLP technologies *could* be a support to the public authorities to reduce burden on human resources or achieve more accurate results. Clearly, whether such technologies actually *should* be used depends on additional normative considerations given that evaluation has important implications for the democratic process as outlined earlier. The increasing use of AI in the public sector raises fundamental questions about transparency (e.g. what goes on inside the black box of the algorithm), accountability (e.g. who is responsible for decisions derived from AI), fairness (e.g. is the algorithm biased) and how these impact on the legitimacy of decisions, among others. There is now an established debate that focuses on these implications that are different for governments than for businesses [21,42,86,95].

However, we believe that questions about the ethics of AI use in government cannot be answered without a better understanding of the value that the technology could actually provide: If existing technologies cannot support the evaluation process, their implications would remain irrelevant. Conversely, if AI would be able to support evaluation, it is necessary to assess the degree of efficiency gains and the risks involved (such as mislabeling) to weigh these up against normative requirements such as fairness and accountability. This review can offer the basis for such a normative judgement. Therefore, in contrast to current reviews of AI use in government [88,95,104], we focus explicitly on the technology used and its performance instead of more general implications of the use of AI in the public sector.

3 METHODOLOGY

We have conducted a systematic literature review, following the basic steps as suggested by Kitchenham [39] including i) identifying relevant research, ii) selection of relevant studies, and iii) quality assessment of the selected studies, followed by the actual analysis and synthesis of the data.

The major challenge to the *identification of relevant research* has been that the task of evaluating citizen contributions has so far not been recognized as a research problem in its own right, but that relevant research occurs in different research areas. The research area that focuses on the development as well as the implications of using AI for policy-making has been termed policy analytics [30,82] but relevant research has also been undertaken under the heading of big data [29,88], data science [8], artificial intelligence in government [104] or policy informatics [102].

We have addressed this challenge by combining two search strategies, namely a search of publication databases and a snowballing approach to identify additional studies of interest. We started by searching two publication databases that complemented each other as one focused on the technology of interest, while the other focused on the application area of interest. On the one hand, we used the Association for Computational Linguistics Anthology[1] as it offers a large collection of more than 80,000 papers from the field of computational linguistics. On the other hand, we drew on the almost 18,000 documents from the Digital Government Reference Library [78] to find peer-reviewed papers in the domain of digital government and democracy[2]. Including all articles up until early 2023, the search resulted in 285 documents that were subsequently screened to *select studies of relevance* to the goal of this literature review. Papers were not only required to use NLP techniques but also had to rely on datasets from the field of public participation or to present the application of these techniques specifically to this domain. What is more, as the focus of this survey is explicitly on contributions generated directly by citizen participation processes, papers were excluded that related only to citizen contributions in a broader sense (such as citizen posts on Twitter about municipal issues). We further requested that the studies either critically evaluated the results of the applied NLP techniques, or proposed particular software solutions for practitioners that used NLP for the analysis of contributions in citizen participation processes. This left a total of 27 studies. Because of

[1] https://aclanthology.org/

[2] See Appendix F for the search terms that were employed.

this small number and the fact that these had all been peer-reviewed, no further *assessment of the study quality* was necessary.

As a second strategy to identify additional relevant literature, we employed a snowballing approach as defined by Wohlin [96]. Using these 27 publications as a starting set, we conducted backward snowballing by accessing the references cited in these publications, as well as forward snowballing by using Google Scholar to find more recent publications citing any of the publications in the starting set. To complement the snowballing approach, we followed the suggestion by Wohlin [96] and screened the entire list of publications of all authors that had (co-)authored several of the papers in our list of relevant documents. This strategy resulted in 28 additional studies.

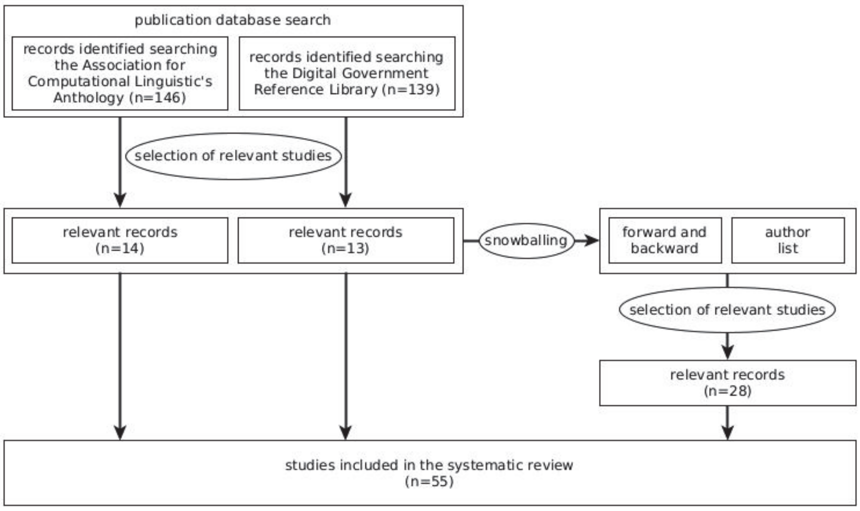

publication database search
records identified searching the Association for Computational Linguistic's Anthology (n=146)
records identified searching the Digital Government Reference Library (n=139)
selection of relevant studies
relevant records (n=14)
relevant records (n=13)
snowballing
forward and backward
author list
selection of relevant studies
relevant records (n=28)
studies included in the systematic review (n=55)

Figure 1: Study identification and selection process

Through this combination of strategies that is visualized in Figure 1 we identified 55 studies. These offer a comprehensive overview of the diversity of existing approaches that have been in use for the particular domain of evaluating public participation contributions, and allow us to identify gaps that we will discuss in the next sections. Given the dispersed state of the field, it is almost impossible to provide a complete overview of all existing studies, but our strategy should allow us to offer a rather comprehensive overview of the state of the field.

4 TASKS IN THE PROCESS OF EVALUATING PUBLIC PARTICIPATION CONTRIBUTIONS

While consultation processes initiated by public authorities differ in the format of contributions citizens provide, the type of information the receiving authority is looking for, and the formal requirements for processing submissions, it is possible to recognize two broad evaluation requirements that are common across all of these types of processes. These are identifying substantive contributions on the individual level, and gaining insights into common themes and trends on the aggregate level. Livermore et al. [52:1015] term these the “haystack problem”, i.e. to find signal in the noise of mass contributions, and the “forest problem”, i.e. to derive information from the whole corpus of contributions. While the analytical perspectives are different, the tasks necessary to achieve these insights are largely similar.

Based on the literature reviewed here [37,55,81,82] and confirmed by our own interviews with practitioners [74], we can identify a number of generic tasks that need to be performed: i) detecting (near) duplicates, ii) grouping of contributions by topic and iii) analyzing the individual contributions in depth, e.g. to identify arguments or other content of relevance. Each of these tasks can help to find the individual comment of relevance among a mass of comments, for example by removing duplicates, by grouping those with a particular content in one group (and disregarding others) or by providing a sentiment score for individual comments. In the same way, these tasks support identification of themes on the aggregate level, by identifying different topics or providing sentiment distributions.

Figure 2 details these three tasks along with their specific subtasks that we introduce in this section. Tasks highlighted in green are those that have received most attention in the literature and which we subsequently focus on in this review.

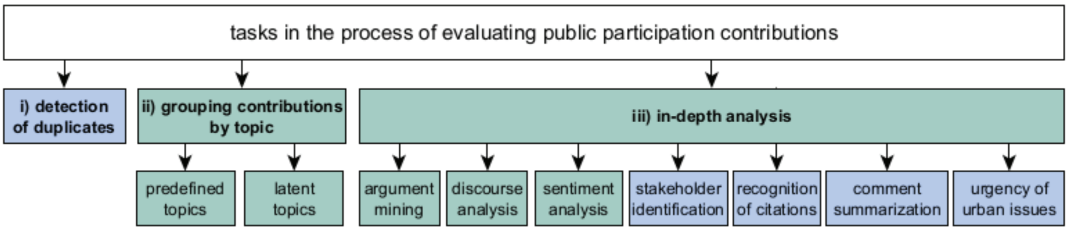


Figure 2: Overview of tasks in the evaluation of public participation contributions

The evaluation process often starts with the **detection of exact duplicates or substantially identical proposals** even though this filtering can also occur in later stages of the process. Given that in particular online comments can be easily submitted, and often campaigns might invite the public to make use of preformulated statements, authorities might receive many comments that are identical or nearly identical. For example, Livermore et al. [52] assume that 99% of the 4m comments to the EPA's Clean Power Plan were actually duplicates or near duplicates. For an earlier rulemaking, Shulman [80] reported that three quarters of comments related to copy-and-paste letters and not individually crafted statements. Identifying duplicate contributions is important for analysts to save time during the evaluation and to avoid undue influence on the process by individual stakeholder groups. At the same time, in the case of near-duplicates, care must be taken to ensure that no substantial information is lost.

The detection of (near) duplicates in the domain of online citizen participation has already been studied by Yang and colleagues [97,98,100] who released the DURIAN (DUplicate Removal In lArge collectioNs) system. Applying DURIAN to 3,000 English-language public comments from U.S. rulemaking showed that the system recognizes duplicates well and with an acceptable runtime. In particular, the high agreement with human ratings of near-duplicates is remarkable. The language-independent structure of the algorithm suggests that duplicates can be detected similarly well in other languages.

Notwithstanding the relevance of duplicate detection, more important for the analysis of citizen input are actually the two remaining tasks. The second task that occurs regularly is that the mass of **contributions needs to be grouped thematically**. This global structuring of all contributions provides the analyst with a quick overview of the topics which arose and in which contributions these can be found. We will provide a detailed overview of the approaches to grouping by topic in Section 5.

As a third task, contributions are **analyzed in further depth**, mainly for **arguments or opinions**. The analysis of arguments and certain aspects of discourse can support a more detailed assessment and indicate how certain issues are perceived by the public. Approaches to solving these tasks form the largest portion of the literature reviewed and will be discussed in Section 6. In addition, there are a number of other aspects for which automated solutions were considered useful in the evaluation of citizen participation processes. These include stakeholder identification [4], the recognition of citations in public comments [5], the estimation of the urgency of urban issues [57], a relatedness analysis of provisions in drafted regulations and public comments [47] and the summarization of comments [3].

5 GROUPING THE DATA COLLECTION BY TOPIC

There are two ways of addressing the task of sorting citizen contributions into topic groups that are shown in Figure 3. In *supervised machine learning*, the goal is to predict the true label(s) for a given data point out of a set of predefined topics. To build such a machine learning model, labeled training data is required to fit a model to the task. In contrast, *unsupervised machine learning* does not need training data. The goal of these models is to find latent topics in the data to form clusters of topically similar data points. We review in turn how both approaches have been used to categorize contributions from citizens.

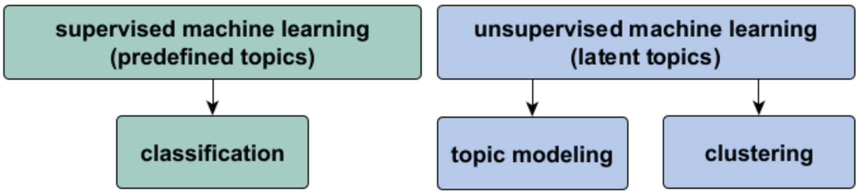

supervised machine learning (predefined topics)
unsupervised machine learning (latent topics)
classification
topic modeling
clustering

Figure 3: Approaches to grouping the data collection by topic

5.1 Supervised Approaches: Classification by Thematic Categories

We first concentrate on the *classification* (hereinafter also referred to as *categorization*) of textual content into appropriate content-based categories. This approach relies on a predefined set of (thematic) categories and uses supervised learning to train an algorithm which can then subsequently classify citizen contributions and assign these to the pre-defined topic groups. Administrative staff or service providers usually categorize contributions according to various aspects when evaluating them. By assigning the contributions to the appropriate categories, it is easier to grasp and summarize the essential issues raised within each of the individual categories. It also allows to focus on particular topics in order to identify individual contribution of relevance. Table 1 in the Appendix provides a systematic overview of the literature covered here, including information on the datasets, the categorization schemes, and the algorithms that have been applied in the studies.

The evaluation datasets range from formal processes such as U.S. eRulemaking to more informal civic participation projects (online and on-site) from Chile, Germany and South Korea. Thematically, the processes focused on transportation and environment, as well as on urban issues and a constitutional process. A variety of categorization schemes is used, which differ in the number and subject of categories as well as between hierarchical and non-hierarchical structures. Categorization is furthermore conducted on different levels of granularity: either on contributions in their entirety or smaller units of analysis, e.g. sentences, ideas, or arguments.

Categorizing contributions [6,38,44,45] yielded good results for the categories that occur frequently in the training datasets, while most categories with little support could only be recognized moderately to poorly. Balta et al. [6] faced a further difficulty when working with a category that represents a collection of miscellaneous topics. In contrast to the more specific categories, it is difficult to find class-typical indicators for such a group (i.e. “other”).

Cardie and her co-authors [12,13] focus on sentence-level categorization. They compare a flat categorization approach with a hierarchical attempt that leads from main categories to more detailed subcategories. Surprisingly, the hierarchical approach cannot surpass the flat one. At the same time, however, none of the approaches can really convince. Aitamurto et al. [1] also categorize hierarchically and achieve good results for the main categories. At the lower levels of the hierarchy the performance is significantly weaker.

Fierro et al. [26] predict matching constitutional concepts for arguments with moderately good results. Interestingly, in addition to exact match performance, the authors also consider whether the correct concept is among the five most likely concepts identified by the algorithms. This is indeed almost always the case for the best performing algorithm fastText. Especially with regard to a software solution in which human and machine work together, these are promising results because human coders could be supported by restricting their choices from a large number of categories to a few most likely ones. Giannakopoulos et al. [28] enhance the exact classification performance with a neural network but the algorithm takes more than seven hours to train.

Regardless of the classification quality of the approaches presented so far, in all works a substantial amount of data was used for training purposes, e.g. several to over a hundred thousand sentences, arguments or documents. At the same time, all works use categorization schemes that are tailored to the corpus in question and hence a customized model must be trained for each dataset. This creates a tension because in order to support an analyst's work, the additional workload caused by manual annotation of data must be kept low.

To address these problems and to provide a feasible solution, Purpura et al. [70] suggest the use of *active learning*. Active learning takes place in close collaboration with the user and consists of two steps: First, a fixed number of unlabeled data points are selected that are assumed to bring the highest gain for the training of a (classification) algorithm. Second, the selected unlabeled data points are manually labeled with the appropriate topic and the classifier is re-trained with all already labeled data. Both steps are repeated until the classifier is reliable. As expected, the evaluation shows that active learning tends to achieve good precision faster than non-active learning, but a closer look at the results highlights that the tested algorithms (Support Vector Machines (SVM), Naïve Bayes, and Maximum Entropy) must still be trained with about 1,000 data points to achieve good results. In a more recent paper, however, Romberg & Escher [75] were able to show that the amount of training data can be significantly reduced to a few hundred data points when active learning is combined with current state-of-the-art approaches for text classification (i.e., pre-trained language models).

5.2 Unsupervised Approaches: Topic Modeling and Clustering

In contrast to supervised procedures, unsupervised approaches that assume no prior knowledge of the data can be applied. Basically, there are two types of approaches which are both unsupervised learning strategies: In *topic modeling* the latent topics of a collection of texts are explored and for each document the degree of membership to each topic is determined. In *clustering*, documents are grouped by similarities. If the similarities are determined on the basis of the content of the texts, the clusters can represent topics as in topic models. In the following we will provide an overview of those works in which these algorithms are not only applied but also analyzed and evaluated. The existing works applied unsupervised approaches to eRulemaking processes as well as e-participation and e-partitioning data from the U.S., Austria, China, Spain and Belgium. The detailed list of works and their characteristics can be found in Table 2 in the Appendix.

In contrast to supervised learning, the evaluation of unsupervised learning algorithms is more complex because there is no labelled ground truth to which the results can be compared. In the works reviewed here, either manual qualitative analysis or measurement of the agreement between algorithmic and human topic assignment are used to rate the algorithms' quality. Most works relied on the topic modeling method Latent Dirichlet Allocation (LDA) to find clusters of thematically similar contributions, which presupposes a fixed number of topics. Levy & Franklin [49] algorithmically detect eight topic clusters of which seven are confirmed by human review. Hagen et al.'s [36] best model, determined by experimenting with different values for the number of topics, consisted of 30 topics of which 21 had a coherent theme. Manual judgement also showed that labeling the topic clusters with the most probable topic term worked well for high-quality topics. Similar findings were reported by Arana-Catania et al. [2], but for the respective dataset the alternative method Non-Negative Matrix Factorization (NMF) was able to detect a higher number of relevant topics than LDA. In contrast to the manual analysis used in these studies, in Ma et al. [53] the best number of topics is estimated with the perplexity metric. In a user study, the LDA model outperformed a common public management search method.

An alternative approach to LDA is the use of associative networks, in which topically related concepts can be clustered based on activation patterns [89]. Manual comparison showed that the emerging clusters resemble the categories that are used by the citizens on the participation platform, e.g. environment, health or education. Simonofski et al. [82] proposed the use of *k*-means clustering which (similar to LDA) requires a predefined number (*k*) of clusters to be found. To overcome this limitation, the authors proposed the so-called elbow method to computationally determine an optimal value. In a manual analysis with two practitioners, the limitations become clear: both believed that the clusters must be checked manually. Nevertheless, they also acknowledged the helpfulness of the algorithm to avoid manual clustering.

The abovementioned works show that unsupervised learning can identify topics, but with serious limitations. To address the challenges of interpretability and validity of LDA for content analyses, Hagen [34] has three recommendations for the application of topic modeling: (1) Word stemming can enhance results but further preprocessing of the data should be kept to a minimum. (2) The number of topics should be determined with a combination of the perplexity metric and human judgement. (3) The generated topics should always be validated (e.g. for topic quality, external validity and internal coherence).

Topic models without strong human supervision tend to produce topics that have no clear meaning to analysts which can be caused by inappropriate model parameter choices, or the deviation of the statistically meaningful model outcome from the outcome expectations of an analyst. To overcome the mismatching of topic models, Cai, Sun & Sha [10] propose the use of interactive topic modeling. Similar to the active learning approach for supervised learning, in interactive learning, the human user is directly involved in the model building process. In the first step, topics are discovered unsupervised. Then, the user investigates the clusters and refines them by merging or splitting topic clusters. The resulting topic model can be qualitatively inspected to decide whether further refinement is necessary. Evaluation on some example cases showed that the manual refinement operations improved the clustering and led to higher overall topic coherence. Yang & Callan [99] also use an interactive approach, based on clustering, and introduce the software OntoCop to construct topic ontologies in collaboration with a user. Human evaluation showed that the interactive setting can reduce the time needed to receive a satisfactory topic clustering and that interactively constructed ontologies resemble manually constructed ones.

6 MINING ARGUMENTS AND OPINIONS IN CITIZEN CONTRIBUTIONS

After reviewing approaches to the second task of topical grouping, we now turn to technical solutions to support the third evaluation task, namely an in-depth analysis of individual contributions. While these include different tasks as outlined in Section 4, here we focus on the analysis of argumentation components, of discourse and of sentiments as these are the tasks that have been most often addressed in the studies we review here.

6.1 Argument Mining

Public participation often takes place in a discursive format. Citizens can express their opinions and ideas on certain topics, have the possibility to refer to the contributions of others in their comments and to argue for or against stances. In the evaluation, the analysis of arguments is important in order to make the different citizen opinions visible. The term *argument mining* refers to the automated identification and extraction of arguments from natural language data. Judging from the results of our literature review, it is one of the most prominent parts of research in the field of citizens' participation. Table 3 in the Appendix provides the details on the individual studies which we summarize in the following subsections. Like in topic grouping, many of the datasets originate from U.S. eRulemaking initiatives. Further data sources that have been used derive from German-language citizen participation on the restructuring of a former airport site, as well as on transportation-related spatial planning processes, a Japanese-language online citizen discussion on the city of Nagoya, and citizen contributions from the 2016 Chilean constitutional process.

According to Peldszus & Stede [68], argument mining can be systematized as three consecutive subtasks: (1) segmentation, (2) segment classification, and (3) relation identification. While some of the reported approaches tackle multiple steps at once, where possible we nevertheless address the results separately in the three steps.

6.1.1Segmentation

In the segmentation step, citizen contributions are divided into units of argumentative content[3]. All papers that we review here use sentences as units of information and classify them as either argumentative or not[4]. A direct comparison of the results is hardly possible due to the differences in the datasets (i.e. language, specific properties of the processes analyzed, share of argumentative content). While Eidelman & Grom [23] work on a dataset consisting of almost 90 percent non-argumentative sentences, argumentative content prevails in the datasets introduced by Liebeck, Esau & Conrad [51], Morio & Fujita [59] and Romberg & Conrad [73]. This class distribution strongly influences the performance of the algorithms. So do similar algorithms (such as SVM) produce divergent results on the different datasets. Overall fastText and logistic regression with embedding features [23], SVMs with a combination of unigrams and grammatical features [51], BERT [73] and parallel constrained pointer architectures (PCPA) [60] lead to the best but not yet sufficient results in classifying argumentative sentences on the respective datasets.

6.1.2Segment Classification

Following segmentation, the identified argument units need to be mapped to their function in the argument. The schemas used to capture the different functions of argumentative discourse units vary widely. Most works focus on recognizing the contextual function of the components of an argument. Additionally, there are a number of works that focus on intrinsic properties, i.e. the verifiability, evaluability, and concreteness of arguments.

Morio and Fujita [59,60] use a straightforward scheme of *claim* and *premise*. Claims are defined as the core component of an argument and consist of controversial statements. Premises are reasons supporting or opposing a claim. A related two-fold division is used by Kwon and co-authors [44,46] who distinguish *main claims* from *sub-claims and main-supporting/opposing* reasons of a main claim. Liebeck et al. [51] introduce *major positions* ("options for actions or decisions that occur in the discussion") as an additional component type for processes in which citizens can submit their own proposals for discussion. Romberg & Conrad [73] likewise differentiate between premises and major positions. Some

[3] Peldszus and Stede (2013) originally assume that relevant (i.e. argumentative) text passages have previously been detected. We also consider the distinction between argumentative and non-argumentative content in the segmentation step.
[4] The task of sentence splitting is well studied and usually provides reliable results.

works further differentiate into supporting or opposing arguments [23,51]. Another argumentation scheme [26] divides arguments according to whether a *policy* is being proposed, a *fact* is being stated, or a *value*-based statement is being made.

In addition to differing concepts of argument components, the various works approach the classification process differently. While some use a sequential approach in which several subtasks (e.g. the identification of claims and classification of claim types) are solved successively [44,45,50,51,73], others attempt to solve the segment classification in a single step [26,28,59,60]. Eidelman & Grom [23] are the only ones who compare the results of a flat classification using all argument types and a sequential strategy combining stance (opposition, support) classification with a more precise classification into specific stance types.

How do these approaches perform? All evaluated approaches for argument component classification in Kwon et al. [45] and Kwon & Hovy [44] perform poorly. Liebeck et al.'s [51] best approach, a SVM with unigram and grammatical features, shows encouraging results but still leads to frequent misclassifications. In claim type classification, SVMs with character embeddings and Random Forests (RF) with unigrams show good results. Promising results are also shown by Morio & Fujita [60] using Pointer Networks (PN) and their own approach PCPA, and by Eidelman & Grom [23], who reported the best performance with logistic regression and word embeddings. Likewise, the approaches still need to be improved. The results obtained by Fierro et al. [26] and Giannakopoulos et al. [28] are strong, although the class distribution of the data is very imbalanced. It turns out that neural networks (convolutional and recurrent layers, attention mechanism) can outperform classical approaches and fastText on this dataset. Encouragingly, Romberg & Conrad [73] were able to show that BERT can consistently provide a very good distinction between premises and major positions across a variety of processes.

One of the problems with the interpretation of arguments from citizens' contributions is that they often lack a justification or a supporting component that substantiates the statement. A number of studies [33,64,67] concentrate on developing NLP models to classify the verifiability of propositions. The comparison of their approaches shows that although networks with Long Short-Term Memory (LSTM) exceed other approaches in that unverifiable propositions could be identified with high quality, the prediction of the different types of verifiability (i.e. non-experiential and verifiable experiential propositions) seems more difficult. Niculae et al. [62] and Galassi et al. [27] focus on a more comprehensive argumentation model to assess the evaluability of citizen's contributions for eRulemaking within the *Cornell eRulemaking Corpus – CDCP* [65]. Promising results suggest the use of structured learning approaches, which cannot be surpassed by residual networks. Falk & Lapesa [25] highlight the role of personal experiences and stories in grounding arguments in political discourse. They show that BERT models can reliably find contributions that contain such a form of justification.

Another aspect that can aid evaluators to efficiently process contributions is to assess the concreteness of proposals by citizens as it is easier to derive measures for implantation from more specific proposals. Looking at a transport-related spatial planning process, Romberg [72] proposes a ranking based on three levels of concreteness and the results of the best-performing method BERT show that the prediction of concreteness is possible but needs to be improved.

6.1.3Relation Identification

In order to understand arguments in their entirety, it is also important to investigate the relationship between the previously identified components. Most related works focus on *support* relations [18,27,48,62]. The first three tested on the CDCP corpus, which makes the results directly comparable. Unlike the other works, Cocarascu et al. [18] trained their models on further argument mining datasets that are not from the public participation domain. This has the advantage that a larger amount of training documents is available to build the classification model. While most approaches perform weakly, the use of additional training datasets shows strong results for all models evaluated. Surprisingly, simple RF and SVM approaches can compete with deep learning models if an appropriate training dataset is used. However, the results vary considerably depending on the choice of the training dataset.

In addition to *support* relations, Morio & Fujita [59,60] define an argumentation scheme specifically for discussion thread analysis and thus to the discursive reply-to structure that can be found in (online) citizen participations. In particular, they distinguish between inner-post relationships of different argumentative components within a post, and inter-post relationships that link two distinct posts in a discussion thread that relate to each other. PCPA, an algorithm specifically designed for thread structures, clearly outperforms state-of-the-art baselines for identifying such relations.

6.2 Discourse and Sentiment Analysis

Based on argumentation structures, a discussion can be analyzed for certain characteristics such as the controversy, divisiveness, popularity and centrality of discussion points. Analyzing discursive elements allows tracking of how consensus decisions develop or where great disagreement between citizens exists. This information can support an analyst in the more in-depth analysis of data and in summarizing the important points of a debate. Approaches to determine controversial points in online discussions are presented in two works. Konat et al. [41] rely on argument graphs and apply two measures for divisiveness defined on graph properties. Cantador, Cortés-Cediel & Fernández [11] propose a theory-based metric to measure controversy. The authors' review of selected examples suggests this is a reasonable approach. To determine the centrality of discussion points, Lawrence et al. [48] apply the mathematical concept of eigenvector centrality on an argument graph. A comparison of the results with human annotation shows a strong overlap, suggesting eigenvector centrality to be a suitable way to predict centrality.

Sentiment Analysis, also referred to as Opinion Mining, is the process of detecting and categorizing opinions in order to determine the writer's attitude regarding a certain subject. This can be relevant for the evaluation of citizens' contributions as it enables officials to get a sense not only of what the key issues are, but how (positively or negatively) these are perceived by citizens. Maragoudakis et al. [55] provide a general overview of existing opinion mining techniques and make assumptions on if and how they can be transferred to analyzing citizens' contributions. They formulate a basic framework for the use of opinion mining methods in e-participation and provide recommendations for use. In addition, there are various works that develop or apply sentiment analysis methods to public participation contributions that we summarize here and which are listed in more detail in Table 4 in the Appendix.

Research focused on the analysis of citizens' subjective claims and the public opinion in large data collections to support rule-writers, the impact of the sentiments in public input on a policymaking process, and the analysis and visualization of the public opinion of open-ended survey questions and free texts from e-consultations. Except for one Greek-language dataset, all works rely on English datasets from the field of civic participation and eRulemaking.

Methods have been developed to analyze public opinion on different levels of granularity (single claims, comments/contributions, or topics) and with varying tonality scales. While most papers use *discrete tonality scales*, e.g. a distinction into negative or positive polarity of a comment or a distinction into supporting or opposing stance towards some claim, Aitamurto et al. [1] use a *continuous scale* in the range of values from -1 to 1, where -1 describes an all negative and +1 an all positive attitude.

The best results for classifying supporting or opposing opinions achieved by Kwon and colleagues [44–46] come from a boosting algorithm and provide almost human-like results. Although it is difficult to predict whether the approach can provide similar outcomes on other datasets, the results seem promising for the automated determination of stance positions. The additional distinction of neutral opinions, on the other hand, was harder and significantly lowered the prediction quality. The approach of Soulis et al. [85] scored worse, but considering the number of sentiment classes (four) and the small training dataset, these results are likewise positive. The results of the only approach with a continuous tonality scale seem to be more limited.

In contrast to previous work analyzing citizens' attitudes via sentiment (from positive to negative), Jasim et al. [37] propose analyzing the emotions behind them. This was prompted by findings from interviews with human evaluators in which a division into positive and negative attitudes was considered insufficient. Rather, they expressed a desire to learn whether the citizens were excited, happy, neutral, concerned, or angry regarding an issue. In a comparison of different classification algorithms, BERT was found to perform best, predicting the five emotions very well.

7 DISCUSSION

Based on the presented NLP approaches, we can assess how well the three generic evaluation tasks identified in Section 4 are already supported and what issues remain that should be addressed in further research.

7.1 Summary of the Current State of Research on the Evaluation of Public Participation Contributions

Much to our surprise, with DURIAN we found only one approach that has been specifically developed for (near) *duplicate detection* in the domain of public participation [97,98,100]. However, the developed solution achieves good results. There is considerably more research on the task of *topical grouping*. Overall, the different supervised learning approaches, varying in granularity of analysis and in categorization schemes, showed moderate to good results. However, so far identifying rarely occurring categories poses a challenge to all these efforts. What is more, the usability of these supervised learning approaches is hampered by categorization schemes tailored to individual datasets and the resulting additional effort required to manually categorize a considerable amount of contributions for the training of customized classification models. According to the reviewed papers, this implies several thousand data points (e.g. sentences, arguments, or contributions). Clearly, especially for small datasets, categorization approaches that need to be trained on such large datasets are not a relief, but rather an additional burden for authorities. It should also be noted that participation processes with less than a thousand contributions do occur regularly. As a solution the use of active learning was proposed to keep the required amount of training data as low as possible [70], and recent work has confirmed that combining it with modern language models can meaningfully support participation processes consisting of a small number of contributions [75]. Still, a manual labeling effort is required. What is more, in active learning the classification algorithms must be constantly retrained, posing limits to the use of complex (i.e. time-consuming) models.

Unsupervised models avoid the manual effort of labeling training data. Most research projects rely on the topic modeling technique LDA and have achieved some promising results. However, the studies have shown that the quality of the resulting topic clusters strongly depends on case-specific model settings, such as the initial choice of the number of topic clusters. In the reviewed articles, parameter selection is either approached by human judgement or by using metrics, but it is understood that the model outcome needs human validation. Therefore again, a strong involvement of the analyst is needed. A further problem in the application of topic modeling methods is rooted in the statistical model itself: Although a resulting topic model might be correct from a mathematical point of view, it does not necessarily correspond to the perception of topics by a human evaluator. The only solution to control the emerging topics and to approximate them to those desired by the user seems to be the direct involvement of the user through interactive topic modeling [10,99]. For the practical application of topic modeling in the public sector, this development is very promising but in need of further research. What is more, only in a few papers has an attempt been made to automatically provide labels for discovered topic clusters.

Most of the literature in this review focused on the *automated recognition and analysis of arguments,* one particular aspect of the task of in-depth analysis of contributions. Overall, although promising approaches exist for each of the three consecutive subtasks (segmentation, segment classification, and relation identification), none of them has been solved satisfactorily. Good approaches for classifying argument components have relied on PN and PCPA [60] or BERT [25,73]. In addition to arguments, the *analysis of discourse structures as well as sentiments* has produced good results already.

7.2 Research Agenda

Considering the field as a whole, since the beginnings of using NLP to support the evaluation of public participation contributions, the technical possibilities in machine learning have steadily developed. In particular, the rise of pre-trained language models (PLM) in recent years has brought an unprecedented boost. Above all, models based on the transformer architecture such as BERT and GPT-3, have been able to achieve considerable improvements over earlier algorithms in many supervised machine learning tasks [93], including topic classification, sentiment analysis and argument mining. However, despite this encouraging development, it remains to be tested whether these successful applications are transferable to our domain. This literature review has revealed that PLMs have rarely been applied to the evaluation of citizens' input from participation processes. So far, PLMs (i.e., BERT) were only used in grouping input by topic [6,75], in the analysis of arguments for the detection and classification of argument components and properties [25,72,73], as well as for the prediction of relations between the argument components [18] and for emotion analysis [37]. These initial efforts are promising but need more systematic application and evaluation. In particular, the focus should be on the development of robust PLMs that perform reliably and consistently across different participation processes. Such important properties have so far remained a challenge for the practical applicability of algorithms [94], but are essential to ensure the value of automation and thus the benefit for practitioners.

Turning to the individual tasks discussed in this study, we identify the following promising avenues for further research. *Duplicate and near-duplicate detection* is a well-known task in data science for which a multitude of approaches are available [e.g. 90] but so far these have not been studied in detail beyond the early DURIAN approach. This obvious gap is waiting to be addressed in future work. Regarding *topic classification* as the supervised approach to thematic grouping, more recent work has shown the benefits of PLMs. Given the trade-offs between training and automation outlined above, active learning that combines human feedback in the training process offers the possibility to reduce training efforts. What is more, is has also the potential to increase trust in the AI-based classification process as it brings human and machine closer together. While existing efforts seem promising [70,75], the field of active learning constantly evolves from which further research efforts should benefit [103]. An alternative to active learning could be provided by the development of categorization schemes that are universally applicable to particular types of content such as different issues that are regularly subject to participation (e.g. infrastructure planning or regulation drafting). This would allow one-time training of arbitrary classification models, which could then be used directly in practice.

Research has also progressed for unsupervised machine learning tasks such as *topic modeling*. Since the introduction of LDA, other topic modeling approaches have been introduced, such as word-embedding based topic models or topic modeling with BERT [17]. Again, these novel techniques offer great potential for the automatic support for the evaluation of public participation data, especially when applied in interactive settings. A starting point for this is offered by various works on the support of content analysis by human-in-the-loop topic models in recent years that focused on user needs and perceptions [e.g. 84] and on technical advancements [e.g. 43,101]. What is more, only in a few papers has an attempt been made to automatically provide labels for discovered topic clusters. These efforts should be pursued in order to aid the interpretation of the output from unsupervised methods, because having a label can be extremely helpful for an analyst to understand the content of individual topics faster.

Two gaps have been revealed in the research on *argument mining*. First, further work is needed on techniques for identifying argument components and their relationships for participation data. After all, the mining of arguments is a very complex area that has developed rapidly in recent years [e.g. 83]. Second, the lack of standardized argumentation models became obvious. What should be prioritized is the (theoretical) development of uniform argumentation models for citizen participation procedures. For example, Liebeck et al. [51] and Fierro et al. [26] have tailored argumentation models to informal participation procedures. These models do not necessarily have to be highly complex. Simply recognizing proposals and the respective rationales can already provide great support in the evaluation. Worth further exploring is also the idea to use additional argument mining training datasets from domains other than participation processes in order to improve the classification as has been demonstrated for relation identification [18]. Regarding *discourse and sentiment analysis*, the application of PLMs has so far been neglected despite its obvious potential, not least illustrated by the successful analysis of emotions in citizen contributions [37]. An open question remains whether analysts are better supported by discrete tonality classes or via continuous values and, when choosing discrete categories, how many categories the polarity spectrum should comprise. We suspect that the use of a few meaningful categories, such as agreement and disagreement, might be better suited to quickly convey the essential points of the content to the analyst.

Apart from these specific gaps, there are a number of other *broader challenges* that exist across all evaluation tasks. A large part of the research concentrates exclusively on English language data. There is little research that focuses on other languages. As languages differ in their syntactic and semantic properties, more coded datasets in other languages are required to apply, adapt and test existing algorithms. Currently, only few non-English language resources are publicly available [2,51,76].

Another overarching challenge is that in order to reap the benefits of such automated procedures, it is not enough to identify suitable mechanisms and algorithms but such procedures need to be made available in ways that public officials can apply them to their data. For example, as multiple reviews highlight [92,95,104], there remains a significant lack of technological expertise in the public sector and among those tasked with implementing and using the technologies reviewed here. Hence, it is necessary to provide end-user software. This review has found that only little work has been devoted to the dedicated development of tools that implement such analysis technologies. These are listed in Table 5 in the Appendix. Given that most of these tools are not available or supported any more[5], cover only specific aspects (e.g. language, functionalities), and are restricted either by the underlying techniques or the visualization methods, we identify a clear need for (preferably open source) applications that make these algorithmic approaches accessible to public administration. A promising step in this direction if offered by CommunityPulse [37]. However, the development of suitable solutions and their integration into the everyday work of experts poses a number of challenges, as Hagen et al. [35] highlight.

8 CONCLUSION

While public authorities are routinely consulting citizens to inform decision-making processes, these procedures come with the challenge of evaluating the contributions made by citizens. This evaluation has important consequences for the effectiveness and legitimacy of policies deriving from public participation but it is a resource-intensive process, so far requiring substantial human effort. We have argued that AI in the form of NLP could be one possibility to support this human evaluation process and eventually be a decisive factor for the public sector to engage or not to engage the public at

[5] In fact, we found only publicly-available implementations for CivicCrowd Analytics (https://github.com/ParticipaPY/civic-crowdanalytics) and Consul (https://github.com/consul).

all. While the use of automated procedures in decision-making processes raises normative concerns such as transparency or accountability, here we have focused on assessing the state of the technology and its potential benefits to inform the debate on these important questions.

Overall, public authorities are still largely lacking reliable tools that could be used in practice to support their work. What is more, despite the fact that NLP has seen major advances in recent years, research on computer-supported text analysis to support the evaluation of citizen contributions is sparse and dispersed across different fields and disciplines. Therefore, this study set out to take stock of this field by reviewing the strengths and weaknesses of existing approaches to offer guidance for further research. Despite a number of promising approaches, we established that most of them are not yet ready for practical use. It remains to be seen whether this situation improves once current state-of-the-art NLP techniques are applied more frequently to this domain.

Among the approaches that are proposed as possible solutions to the problems identified, many draw upon the expertise of humans, for example through active learning or interactive topic modelling. While this suggests that human expertise can never be fully replaced as for example asserted by Grimmer & Stewart [32], it has yet to be established whether such approaches would eventually still require less time for evaluation than human-only evaluation. Finally, it became clear that there remains a significant lack of non-English language datasets and models as well as software that would allow the application of the models in practice.

Taken together, this leads us to conclude that the evaluation of citizen contributions - despite the significance outlined above - has not received the scholarly attention that it deserves. We hypothesize a number of reasons for this lack of interest. First, while interest in the utility of big data for policy has been large, citizen contributions do not fulfil the definition of big data: Despite their occasional large number, they usually remain in the hundreds or thousands. What is more, compared to traffic or sensor data, instances of public participation are sporadic and not continuous and hence might attract less interest for automation. Second, natural language data remains highly unstandardized which makes automatic analysis more challenging. Third, further challenges arise from the exceptionally high requirements for transparency and due process for public participation that we outlined earlier, as failures in the evaluation process such as omitting a relevant statement can have important consequences that might also prevent the adoption of automation. Fourth, the lack of technology expertise and capacity in public administration is a barrier to the utilization of advanced technologies [29,69]. At the same time, despite these difficulties, the field of government technologies has been a profitable ground for technology companies and consultancies who offer their technologies to support service provision including dealing with citizen contributions. Due to their business model, these have few incentives to publicly share their technologies, making it more difficult to assess the state of the field.

Although we believe that an open source solution is preferable (e.g. to facilitate deployment in communities or countries with low budgets), the lack of access to commercial solutions is one limitation of this study. Further limitations arise from the fact that the evaluation of citizen contributions is not a clearly demarcated field. As outlined earlier, this makes it possible that our review has missed individual studies. While we have justified our focus on studies that deal with contributions from participatory processes, this has excluded research that could potentially also provide relevant insights, e.g. in relation to social media data. Consequently, further research should try to use the lessons learned from these approaches and test whether they perform well also on public participation data despite the differences in domains. Similarly, in contrast to the top-down public participation that is the focus of this article, bottom-up participation such as petitions or more broadly online discussions (e.g. on social media) are more difficult to incorporate into the formalized decision-making process of public authorities. Nevertheless, increasingly efforts are made to analyze such exchanges to gauge public opinion outside of such formal arenas as these could supplement the input from consultations [see for example

7,15,82]. Such studies can offer further insights on how to provide additional information for decision-making. Finally, we have focused on textual data only, but contributions might also include images, audio or even videos. These would also benefit from automated support and supplement the analysis of citizen contributions but were beyond the scope of this review. Supporting the evaluation of contributions in public participation with computational text analysis is an exciting area of research. Still, more work is needed to turn approaches from research into fruitful approaches to practice. With the rapid progress in the fields of AI, NLP, and policy analytics, these gaps can hopefully be bridged in the near future.

Acknowledgements

We are indebted to previous research efforts within Heinrich Heine University Düsseldorf on the automated support of evaluation of participation processes, conducted by Matthias Liebeck in collaboration with Katharina Esau, as well as the student work of Markus Brenneis and Philipp Grawe.

Funding Statement

This research has been supported by the Federal Ministry of Education and Research within the programme "Social-Ecological Research" (FONA) under grant number 01UU1904. Responsibility for the results rests solely with the authors.

REFERENCES

[1] Tanja Aitamurto, Kaiping Chen, Ahmed Cherif, Jorge Saldivar Galli, and Luis Santana. 2016. Civic CrowdAnalytics: Making sense of crowdsourced civic input with big data tools. In *Proceedings of the 20th International Academic Mindtrek Conference*, 86–94. DOI:https://doi.org/10.1145/2994310.2994366

[2] 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 (July 2021), 1–22. DOI:https://doi.org/10.1145/3452118

[3] Miguel Arana-Catania, Rob Procter, Yulan He, and Maria Liakata. 2021. Evaluation of Abstractive Summarisation Models with Machine Translation in Deliberative Processes. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, Association for Computational Linguistics, Stroudsburg, PA, USA, 57–64. DOI:https://doi.org/10.18653/v1/2021.newsum-1.7

[4] Jaime Arguello and Jamie Callan. 2007. A bootstrapping approach for identifying stakeholders in public-comment corpora. In *Proceedings of the 8th annual international conference on Digital government research: bridging disciplines & domains*, 92–101.

[5] Jaime Arguello, Jamie Callen, and Stuart Shulman. 2008. Recognizing Citations in Public Comments. *J. Inf. Technol. Polit.* 5, 1 (2008), 49–71. DOI:https://doi.org/10.1080/19331680802153683

[6] Dian Balta, Peter Kuhn, Mahdi Sellami, Daniel Kulus, Claudius Lieven, and Helmut Krcmar. 2019. How to Streamline AI Application in Government? A Case Study on Citizen Participation in Germany. In *International Conference on Electronic Government*, 233–247. DOI:https://doi.org/10.1007/978-3-030-27325-5_18

[7] Olfa Belkahla Driss, Sehl Mellouli, and Zeineb Trabelsi. 2019. From citizens to government policy-makers: Social media data analysis. *Gov. Inf. Q.* 36, 3 (July 2019), 560–570. DOI:https://doi.org/10.1016/j.giq.2019.05.002

[8] Jonathan Bright, Bharath Ganesh, Cathrine Seidelin, and Thomas M. Vogl. 2019. Data Science for Local Government. *SSRN Electron. J.* (2019). DOI:https://doi.org/10.2139/ssrn.3370217

[9] Thomas R Bruce, Claire Cardie, Cynthia R Farina, and Stephen Purpura. 2008. Facilitating Issue Categorization & Analysis in Rulemaking. (2008).

[10] Guoray Cai, Feng Sun, and Yongzhong Sha. 2018. Interactive visualization for topic model curation. *CEUR Workshop Proc.* 2068, (2018).

[11] Iván Cantador, María E. Cortés-Cediel, and Miriam Fernández. 2020. Exploiting Open Data to analyze discussion and controversy in online citizen participation. *Inf. Process. Manag.* 57, 5 (2020). DOI:https://doi.org/10.1016/j.ipm.2020.102301

[12] Claire Cardie, Cynthia Farina, Adil Aijaz, Matt Rawding, and Stephen Purpura. 2008. A Study in Rule-Specific Issue Categorization for e-Rulemaking. In *9th Annual International Conference on Digital Government Research (dg.o 2008)*, Montreal, Canada, 244–253.

[13] Claire Cardie, Cynthia Farina, Matt Rawding, and Adil Aijaz. 2008. An eRulemaking Corpus: Identifying Substantive Issues in Public Comments. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation ({LREC}'08)*, European Language Resources Association (ELRA), Marrakech, Morocco. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/699_paper.pdf

[14] Tommaso Caselli, Giovanni Moretti, Rachele Sprugnoli, Sara Tonelli, Damien Lanfrey, and Donatella Solda Kutzmann. 2016. NLP and public engagement: The case of the Italian school reform. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 401–406.

[15] Yannis Charalabidis, Euripidis N. Loukis, Aggeliki Androutsopoulou, Vangelis Karkaletsis, and Anna Triantafillou. 2014. Passive crowdsourcing in government using social media. *Transform. Gov. People, Process Policy* 8, 2 (May 2014), 283–308. DOI:https://doi.org/10.1108/TG-09-2013-0035

[16] Kaiping Chen and Tanja Aitamurto. 2019. Barriers for Crowd's Impact in Crowdsourced Policymaking: Civic Data Overload and Filter Hierarchy. *Int. Public Manag. J.* 22, 1 (January 2019), 99–126. DOI:https://doi.org/10.1080/10967494.2018.1488780

[17] Rob Churchill and Lisa Singh. 2022. The Evolution of Topic Modeling. *ACM Comput. Surv.* 54, 10s (January 2022), 1–35. DOI:https://doi.org/10.1145/3507900

[18] Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. Dataset independent baselines for relation prediction in argument mining. *Front. Artif. Intell. Appl.* 326, (2020), 45–52. DOI:https://doi.org/10.3233/FAIA200490

[19] Cary Coglianese. 2006. Citizen Participation in Rulemaking: Past, Present, and Future. *Duke Law J.* 55, 5 (2006), 943–968. DOI:https://doi.org/10.2139/ssrn.912660

[20] Robert Alan Dahl. 1989. *Democracy and Its Critics*. Yale University Press, Yale.

[21] Katherine A. Daniell, Alec Morton, and David Ríos Insua. 2016. Policy analysis and policy analytics. *Ann. Oper. Res.* 236, 1 (January 2016), 1–13. DOI:https://doi.org/10.1007/s10479-015-1902-9

[22] 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 (80-.).* 363, 6432 (2019), 1144–1146. DOI:https://doi.org/10.1126/scienceaaw2694

[23] 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* (ICAIL '19), Association for Computing Machinery, New York, NY, USA, 199–203. DOI:https://doi.org/10.1145/3322640.3326714

[24] Peter Esaiasson. 2010. Will citizens take no for an answer? What government officials can do to enhance decision acceptance. *Eur. Polit. Sci. Rev.* 2, 03 (2010), 351–371. DOI:https://doi.org/doi:10.1017/S1755773910000238

[25] Neele Falk and Gabriella Lapesa. 2022. Reports of personal experiences and stories in argumentation: datasets and analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 5530–5553. DOI:https://doi.org/10.18653/v1/2022.acl-long.379

[26] 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*, Association for Computational Linguistics, Stroudsburg, PA, USA, 1–10. DOI:https://doi.org/10.18653/v1/W17-5101

[27] 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*, Association for Computational Linguistics, Stroudsburg, PA, USA, 1–10. DOI:https://doi.org/10.18653/v1/W18-5201

[28] 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 *2019 6th Swiss Conference on Data*

Science (SDS), IEEE, 57–62. DOI:https://doi.org/10.1109/SDS.2019.000-7

[29] Sarah Giest. 2017. Big data for policymaking: fad or fasttrack? *Policy Sci.* 50, 3 (September 2017), 367–382. DOI:https://doi.org/10.1007/s11077-017-9293-1

[30] J. Ramon Gil-Garcia, Theresa A. Pardo, and Luis F. Luna-Reyes. 2018. *Policy Analytics, Modelling, and Informatics*. Springer International Publishing, Cham. DOI:https://doi.org/10.1007/978-3-319-61762-6

[31] Kristina Gligorić, Ashton Anderson, and Robert West. 2018. How Constraints Affect Content: The Case of Twitter's Switch from 140 to 280 Characters. *Proc. Int. AAAI Conf. Web Soc. Media* 12, 1 (June 2018). DOI:https://doi.org/10.1609/icwsm.v12i1.15079

[32] Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Polit. Anal.* 21, 3 (January 2013), 267–297. DOI:https://doi.org/10.1093/pan/mps028

[33] Chinnappa Guggilla, Tristan Miller, and Iryna Gurevych. 2016. CNN- and LSTM-based claim classification in online user comments. In *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, 2740–2751.

[34] Loni Hagen. 2018. Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Inf. Process. Manag.* 54, 6 (November 2018), 1292–1307. DOI:https://doi.org/10.1016/j.ipm.2018.05.006

[35] Loni Hagen, Thomas E. Keller, Xiaoyi Yerden, and Luis Felipe Luna-Reyes. 2019. Open data visualizations and analytics as tools for policy-making. *Gov. Inf. Q.* 36, 4 (October 2019), 101387. DOI:https://doi.org/10.1016/j.giq.2019.06.004

[36] Loni Hagen, Ozlem Uzuner, Christopher Kotfila, Teresa M Harrison, and Dan Lamanna. 2015. Understanding Citizens' Direct Policy Suggestions to the Federal Government: A Natural Language Processing and Topic Modeling Approach. In *2015 48th Hawaii International Conference on System Sciences*, IEEE, 2134–2143. DOI:https://doi.org/10.1109/HICSS.2015.257

[37] Mahmood Jasim, Enamul Hoque, Ali Sarvghad, and Narges Mahyar. 2021. CommunityPulse: Facilitating Community Input Analysis by Surfacing Hidden Insights, Reflections, and Priorities. In *Designing Interactive Systems Conference 2021*, ACM, New York, NY, USA, 846–863. DOI:https://doi.org/10.1145/3461778.3462132

[38] Byungjun Kim, Minjoo Yoo, Keon Chul Park, Kyeo Re Lee, and Jang Hyun Kim. 2021. A value of civic voices for smart city: A big data analysis of civic queries posed by Seoul citizens. *Cities* 108, (January 2021), 102941. DOI:https://doi.org/10.1016/j.cities.2020.102941

[39] Barbara Kitchenham. 2004. *Procedures for Performing Systematic Reviews*. Keele.

[40] Roman Klinger, Philipp Senger, Sumit Madan, and Michal Jacovi. 2012. Online Communities Support Policy-Making: The Need for Data Analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 132–143. DOI:https://doi.org/10.1007/978-3-642-33250-0_12

[41] 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)*, 3899–3906.

[42] Pascal D. König and Georg Wenzelburger. 2020. Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Gov. Inf. Q.* 37, 3 (July 2020), 101489. DOI:https://doi.org/10.1016/j.giq.2020.101489

[43] Varun Kumar, Alison Smith-Renner, Leah Findlater, Kevin Seppi, and Jordan Boyd-Graber. 2019. Why Didn't You Listen to Me? Comparing User Control of Human-in-the-Loop Topic Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 6323–6330. DOI:https://doi.org/10.18653/v1/P19-1637

[44] Namhee Kwon and Eduard Hovy. 2007. Information acquisition using multiple classifications. In *Proceedings of the 4th international conference on Knowledge capture - K-CAP '07* (K-CAP '07), ACM Press, New York, New York, USA, 111–118. DOI:https://doi.org/10.1145/1298406.1298427

[45] Namhee Kwon, Stuart W. Shulman, and Eduard Hovy. 2006. Multidimensional text analysis for eRulemaking. In *Proceedings of the 2006 national conference on Digital government research - dg.o '06*, ACM Press, New York, New York, USA, 157–166. DOI:https://doi.org/10.1145/1146598.1146649

[46] 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), Digital Government Society of North America, 76–81.

[47] Gloria T Lau, Kincho H Law, and Gio Wiederhold. 2005. A Relatedness Analysis Tool for Comparing Drafted Regulations and Associated Public Comments. *I/S A J. Law Policy Inf. Soc.* 1, 1 (2005), 95–110.

[48] 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 (July 2017), 1–22. DOI:https://doi.org/10.1145/3032989

[49] Karen E C Levy and Michael Franklin. 2014. Driving Regulation. *Soc. Sci. Comput. Rev.* 32, 2 (April 2014), 182–194. DOI:https://doi.org/10.1177/0894439313506847

[50] Matthias Liebeck. 2017. *Automated Discussion Analysis in Online Participation Projects*. PhD thesis, Heinrich-Heine-Universität Düsseldorf.

[51] 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)*, 144–153.

[52] Michael A. Livermore, Vladimir Eidelman, and Brian Grom. 2018. Computationally assisted regulatory participation. *Notre Dame Law Rev.* 93, 3 (2018), 977–1034.

[53] Baojun Ma, Nan Zhang, Guannan Liu, Liangqiang Li, and Hua Yuan. 2016. Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. *Inf. Process. Manag.* 52, 3 (May 2016), 430–445. DOI:https://doi.org/10.1016/j.ipm.2015.10.004

[54] Narges Mahyar, Diana V. Nguyen, Maggie Chan, Jiayi Zheng, and Steven P. Dow. 2019. The Civic Data Deluge. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, ACM, New York, NY, USA, 1171–1181. DOI:https://doi.org/10.1145/3322276.3322354

[55] Manolis Maragoudakis, Euripidis Loukis, and Yannis Charalabidis. 2011. A Review of Opinion Mining Methods for Analyzing Citizens' Contributions in Public Policy Debate. In *Electronic Participation: Proceedings of the 3rd IFIP WG 8.5 International Conference, ePart 2011*, Efthimios Tambouris, Ann Macintosh and Hans Bruijn (eds.). Delft, The Netherlands, 298–313. DOI:https://doi.org/10.1007/978-3-642-23333-3_26

[56] Giada De Marchi, Giulia Lucertini, and Alexis Tsoukiàs. 2016. From evidence-based policy making to policy analytics. *Ann. Oper. Res.* 236, 1 (January 2016), 15–38. DOI:https://doi.org/10.1007/s10479-014-1578-6

[57] Christian Masdeval and Adriano Veloso. 2015. Mining citizen emotions to estimate the urgency of urban issues. *Inf. Syst.* 54, (December 2015), 147–155. DOI:https://doi.org/10.1016/j.is.2015.06.008

[58] Nina A Mendelson. 2012. Should Mass Comments Count? *Michigan J. Environ. Adm. Law* 2, 1 (2012), 173–183. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208234

[59] Gaku Morio and Katsuhide Fujita. 2018. Annotating Online Civic Discussion Threads for Argument Mining. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 546–553. DOI:https://doi.org/10.1109/WI.2018.00-3

[60] Gaku Morio and Katsuhide Fujita. 2018. End-to-End Argument Mining for Discussion Threads Based on Parallel Constrained Pointer Architecture. In *Proceedings of the 5th Workshop on Argument Mining*, Association for Computational Linguistics, Stroudsburg, PA, USA, 11–21. DOI:https://doi.org/10.18653/v1/W18-5202

[61] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Stroudsburg, PA, USA, 9–14. DOI:https://doi.org/10.18653/v1/2020.emnlp-demos.2

[62] 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)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 985–995. DOI:https://doi.org/10.18653/v1/P17-1091

[63] OECD. 2003. *Promise and Problems of E-Democracy*. OECD. DOI:https://doi.org/10.1787/9789264019492-en

[64] Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the first workshop on argumentation mining*, 29–38.

[65] 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)*.

[66] 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)*, European Language Resources Association (ELRA), Miyazaki, Japan. Retrieved from https://aclanthology.org/L18-1257.pdf

[67] 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*, Association for Computational Linguistics, Denver, CO, 39–44. DOI:https://doi.org/10.3115/v1/W15-0506

[68] Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. International Journal of Cognitive Informatics and Natural Intelligence 7, 1–31. DOI:http://doi.org/10.4018/jcini.2013010101

[69] Martijn Poel, Eric T Meyer, and Ralph Schroeder. 2018. Big Data for Policymaking: Great Expectations, but with Limited Progress? *Policy & Internet* 10, 3 (September 2018), 347–367. DOI:https://doi.org/10.1002/poi3.176

[70] Stephen Purpura, Claire Cardie, and Jesse Simons. 2008. Active Learning for e-Rulemaking: Public Comment Categorization. In *9th Annual International Conference on Digital Government Research (dg.o 2008)*, Montreal, Canada, 234–243.

[71] Brandon Reynante, Steven P. Dow, and Narges Mahyar. 2021. A Framework for Open Civic Design: Integrating Public Participation, Crowdsourcing, and Design Thinking. *Digit. Gov. Res. Pract.* 2, 4 (October 2021), 1–22. DOI:https://doi.org/10.1145/3487607

[72] Julia Romberg. 2022. Is Your Perspective Also My Perspective? Enriching Prediction with Subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 115–125. Retrieved from https://aclanthology.org/2022.argmining-1.11

[73] 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*, ACL, Punta Cana, 88–99. Retrieved from https://aclanthology.org/2021.argmining-1.9

[74] Julia Romberg and Tobias Escher. 2020. *Analyse der Anforderungen an eine Software zur (teil-)automatisierten Unterstützung bei der Auswertung von Beteiligungsverfahren*. Düsseldorf. Retrieved from https://www.cimt-hhu.de/en/2020/cimt-practical-workshop-i/

[75] Julia Romberg and Tobias Escher. 2022. Automated Topic Categorisation of Citizens' Contributions: Reducing Manual Labelling Efforts Through Active Learning. 369–385. DOI:https://doi.org/10.1007/978-3-031-15086-9_24

[76] Julia Romberg, Laura Mark, and Tobias Escher. 2022. A Corpus of German Citizen Contributions in Mobility Planning: Supporting Evaluation Through Multidimensional Classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2874–2883. Retrieved from https://aclanthology.org/2022.lrec-1.308

[77] Vivien A. Schmidt. 2013. Democracy and Legitimacy in the European Union Revisited: Input, Output and "Throughput." *Polit. Stud.* 61, 1 (2013), 2–22. DOI:https://doi.org/10.1111/j.1467-9248.2012.00962.x

[78] Hans Jochen Scholl. 2022. The Digital Government Reference Library (DGRL) 18.5. Retrieved from http://faculty.washington.edu/jscholl/dgrl/

[79] Stuart W Shulman. 2003. An Experiment in Digital Government at the United States National Organic Program. *Agric. Human Values* 20, 3 (2003), 253–265. DOI:https://doi.org/10.1023/A:1026104815057

[80] Stuart W Shulman. 2009. The Case Against Mass E-mails: Perverse Incentives and Low Quality Public Participation in U.S. Federal Rulemaking. *Policy & Internet* 1, 1 (January 2009), 23–53. DOI:https://doi.org/10.2202/1948-4682.1010

[81] Stuart W Shulman, Eduard Hovy, and Stephen Zavestoski. 2004. SGER Collaborative: A Testbed for eRulemaking Data. *J. E-Government* 1, 1 (2004), 123–127. DOI:https://doi.org/10.1300/J399v01n01

[82] Anthony Simonofski, Jerôme Fink, and Corentin Burnay. 2021. Supporting policy-making with social media and e-participation platforms data: A policy analytics framework. *Gov. Inf. Q.* 38, 3 (July 2021), 101590. DOI:https://doi.org/10.1016/j.giq.2021.101590

[83] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon,

Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An autonomous debating system. *Nature* 591, 7850 (March 2021), 379–384. DOI:https://doi.org/10.1038/s41586-021-03215-w

[84] Alison Smith-Renner, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2020. Digging into user control. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, ACM, New York, NY, USA, 519–530. DOI:https://doi.org/10.1145/3377325.3377491

[85] Konstantinos Soulis, Iraklis Varlamis, Andreas Giannakoulopoulos, and Filippos Charatsev. 2013. A tool for the visualisation of public opinion. *Int. J. Electron. Gov.* 6, 3 (2013), 218. DOI:https://doi.org/10.1504/IJEG.2013.058404

[86] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data Soc.* 9, 2 (July 2022), 205395172211151. DOI:https://doi.org/10.1177/20539517221115189

[87] Michael Andrea Strebel, Daniel Kübler, and Frank Marcinkowski. 2019. The importance of input and output legitimacy in democratic governance: Evidence from a population-based survey experiment in four West European countries. *Eur. J. Polit. Res.* 58, 2 (May 2019), 488–513. DOI:https://doi.org/10.1111/1475-6765.12293

[88] Arho Suominen and Arash Hajikhani. 2021. Research themes in big data analytics for policymaking: Insights from a mixed-methods systematic literature review. *Policy & Internet* (June 2021), poi3.258. DOI:https://doi.org/10.1002/poi3.258

[89] Peter Teufl, Udo Payer, and Peter Parycek. 2009. Automated Analysis of e-Participation Data by Utilizing Associative Networks, Spreading Activation and Unsupervised Learning. In *In International Conference on Electronic Participation*, Ann Macintosh and Efthimios Tambouris (eds.). Springer, Berlin, Heidelberg, 139–150. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-03781-8_13

[90] Martin Theobald, Jonathan Siddharth, and Andreas Paepcke. 2008. SpotSigs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, ACM Press, New York, New York, USA, 563–570. DOI:https://doi.org/10.1145/1390334.1390431

[91] Harshit Vajjala, Sowmya; Majumder; Bodhisattwa; Gupta, Anuj; Surana. 2020. *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media.

[92] Stefaan G. Verhulst, Zeynep Engin, and Jon Crowcroft. 2019. Data & Policy : A new venue to study and explore policy–data interaction. *Data Policy* 1, (June 2019), e1. DOI:https://doi.org/10.1017/dap.2019.2

[93] Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. Pre-Trained Language Models and Their Applications. *Engineering* (September 2022). DOI:https://doi.org/10.1016/j.eng.2022.04.024

[94] Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and Improve Robustness in NLP Models: A Survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Stroudsburg, PA, USA, 4569–4586. DOI:https://doi.org/10.18653/v1/2022.naacl-main.339

[95] Bernd W. Wirtz, Jan C. Weyerer, and Carolin Geyer. 2019. Artificial Intelligence and the Public Sector—Applications and Challenges. *Int. J. Public Adm.* 42, 7 (May 2019), 596–615. DOI:https://doi.org/10.1080/01900692.2018.1498103

[96] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14*, ACM Press, New York, New York, USA, 1–10. DOI:https://doi.org/10.1145/2601248.2601268

[97] Hui Yang and Jamie Callan. 2005. Near-duplicate detection for eRulemaking. In *Proceedings of the 2005 national conference on Digital government research*, 78–86.

[98] Hui Yang and Jamie Callan. 2006. Near-duplicate detection by instance-level constrained clustering. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, ACM Press, New York, New York, USA, 421. DOI:https://doi.org/10.1145/1148170.1148243

[99] Hui Yang and Jamie Callan. 2009. OntoCop: Constructing Ontologies for Public Comments. *IEEE Intell. Syst.* 24, 5 (2009), 70–75.

[100] 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*, 239–248.

[101] Michelle Yuan, Benjamin Van Durme, and Jordan Boyd-Graber. 2018. Multilingual anchoring: Interactive topic modeling and alignment across languages. In *Advances in Neural Information Processing (NeurIPS)*, Montreal, Canada. Retrieved from https://proceedings.neurips.cc/paper/2018/file/28b9f8aa9f07db88404721af4a5b6c11-Paper.pdf

[102] Daniel Zeng. 2015. Policy Informatics for Smart Policy-Making. *IEEE Intell. Syst.* 30, 6 (November 2015), 2–3. DOI:https://doi.org/10.1109/MIS.2015.106

[103] Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A Survey of Active Learning for Natural Language Processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 6166–6190. Retrieved from https://aclanthology.org/2022.emnlp-main.414

[104] Anneke Zuiderwijk, Yu-Che Chen, and Fadi Salem. 2021. Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Gov. Inf. Q.* 38, 3 (July 2021), 101577. DOI:https://doi.org/10.1016/j.giq.2021.101577

APPENDICES

A: Overview of supervised approaches for thematic classification

Table 1: Overview of supervised approaches for thematic classification

Dataset	Language	Classification granularity	Categorization schema	Research article	Input features	Algorithms
Rulemaking process by the U.S. Environmental Protection Agency (online and offline submissions)	English	Sentence	9 categories (economic, environment, government responsibility, health, legal, policy, pollution, science, technology)	Kwon et al. [45]	Word *n*-grams, named entities, WordNet synonyms	SVM
				Kwon & Hovy [44]	Word *n*-grams, named entities, WordNet synonyms	SVM
CeRI FTA Grant Circular Corpus: Transportation rulemaking process (online and offline submissions)	English	Sentence	39 categories hierarchically ordered with 17 top level issues (e.g. funding, JARC program issues, planning, procedural, evaluation, none)[6]	Cardie, Farina, Rawding, et al. [13]	Word *n*-grams	SVM
				Cardie, Farina, Aijaz, et al. [12]	Word *n*-grams	SVM[7]
				Purpura et al. [70]	Word *n*-grams	SVM, NB, Maximum Entropy (active learning)

[6] For the complete list, please refer to the paper.

[7] The authors state that NB and CRF were also evaluated. However, the results are not further reported.

Crowdsourced urban planning process with focus on transportation (online)	English	Unique ideas (extracted from citizens' contributions)	Categories hierarchically ordered with 6 top level issues (big picture infrastructure, public transit, private transit, non-motor powered transit, special needs, other) and more detailed subtopics (e.g. traffic calming and road safety)[8]	Aitamurto et al. [1]	Concept extraction tool (heavenondemand)[9]	
Citizen contributions of the 2016 Chilean constitutional process (local on-site events)	Spanish	Arguments	114 predefined constitutional concepts + additional concepts defined by participants, hierarchically ordered with 4 main topics (values, rights, duties and institutions) and more detailed subtopics (e.g. dignity and gender equality)	Fierro et al. [26]	Word *n*-grams, part-of-speech information, Word embeddings	SVM, LR, RF, fastText, deep averaging networks
				Giannakopoulos et al. [28]	Word embeddings	CNN, BiGRU, Attention
Nine different online participation projects for urban development of the city of Hamburg	German	Contributions	8 categories (traffic and mobility, living and work, green and recreation, sports and leisure, climate and environment, other, urban development and urban space, social and culture)	Balta et al. [6]	Character *n*-grams	Language models, BERT
Citizen contributions from the "Oasis of 10 Million Imagination" civic online participation platform for Seoul	Korean	Civic Queries	10 different complaint types (health, economy, traffic, culture, welfare, taxes, safety, female, housing, environment, public)	Kim et al. [38]	Word embeddings	RF

[8] We could not find the total number of topics or an overview of all subtopics in the paper.
[9] The authors do not specify the algorithm and refer to the tool's website for further information. As this webpage is not available anymore, we are unfortunately unable to provide more detailed information about the type of classifier.

Participation processes on cycling infrastructure in three German municipalities	German	Contributions	8 categories (cycling traffic management, signage, obstacles, cycle path quality, traffic lights, lighting, bicycle parking, misc)	Romberg & Escher [75]	Word *n*-grams	SVM, NB, Maximum Entropy, Ensemble Classifier, BERT (active learning)

B: Overview of topic modeling and clustering approaches

Table 2: Overview of topic modeling and clustering approaches

Research article	Dataset	Language	Algorithms
Yang & Callan [99]	Three U.S. Notice and Comment Rulemaking processes on wildlife and environment protection	English	OntoCop (interactive approach) with *k*-medoids clustering
Teufl et al. [89]	Discussion from Austrian youth e-participation platform *mitmachen.at* about future issues	German	Associative Networks with Robust Growing Neural Gas algorithm
Levy & Franklin [49]	Comment data of regulatory debates about electronic monitoring in the U.S. trucking industry from *regulations.gov*	English	Latent Dirichlet Allocation (LDA)
Hagen et al. [36]	Data from U.S. e-petitioning platform *We the People* (WtP)	English	LDA
Hagen [34]	Data from U.S. e-petitioning platform *We the People* (WtP)	English	LDA
Cai et al. [10]	Data from U.S. e-petitioning platform *We the People* (WtP)	English	Correlation Explanation topic modeling (interactive approach)
Ma et al. [53]	Online citizen opinions from a platform for urban public affairs issues in Peking	Chinese	LDA
Simonofski et al. [82]	eParticipation data of four different cities of Belgium	French	*k*-means clustering
Arana-Catania et al. [2]	Public participation datasets from the Consul platform instance of the Madrid City Council	Spanish	Non-negative matrix factorization (NMF), LDA

C: Overview of argument mining approaches

Table 3: Overview of argument mining approaches

Dataset	Language	Argument mining schema	Research article	subtask[10]	Input features	Algorithms
Rulemaking process by the U.S. Environmental Protection Agency (online and offline submissions)	English	argument components: *main root (claim)* and *subroot* (sub-claim or main-support) relations: *support, opposition* and *restate*	Kwon et al. [45]	(1) + (2) + (3)	*n*-grams, subjectivity score, structural properties, cue phrases, named entities, sentiment features, topic information	SVM
			Kwon & Hovy [44]	(1) + (2) + (3)	*n*-grams, subjectivity score, structural properties, cue phrases, topic information	SVM, boosting
User comments from U.S. eRulemaking online platform *RegulationRoom.org* (two rules: Airline Passenger Rights and Home Mortgage Consumer Protection)	English	proposition types: *unverifiable, verifiable experiential* and *verifiable non-experiential*	Park & Cardie [64]	(2)	*n*-grams, core clause tags, part-of-speech information, sentiment and emotion cues, speech events, imperative expressions, tense, pronouns	SVM
			Park et al. [67]	(2)	*n*-grams, lexicon-based features, part-of-speech information, emotion cues, tense, pronouns	CRF
			Guggilla et al. [33]	(2)	embeddings (word2vec, dependency, factual)	CNN, LSTM

[10] Peldszus and Stede (2013) systematize argument mining as three consecutive subtasks: (1) segmentation, (2) segment classification, (3) relation identification.

Cornell eRulemaking Corpus – CDCP [66]: User comments on Consumer Debt Collection Practices rule from *RegulationRoom.org*	English	proposition types: *fact*, *testimony*, *value*, *policy* and *reference* relations: *evidence* and *reason*	Niculae et al. [62]	Joint model for (2) and (3)	lexical information (e.g. *n*-grams, word embeddings and dependency information), lexicon-based features, structural properties, context information, syntactic properties (e.g. part-of-speech and tense), discourse properties	SVM, RNN, linear structured SVM, structured RNN
			Galassi et al. [27]	Joint model for (2) and (3)	word embeddings, structural properties	deep network without residual network block, deep residual network
			Cocarascu et al. [18]	(3)	word embeddings, sentiment features, syntactic features, textual entailment	SVM, RF, GRU, Attention, BERT
			Falk & Lapesa [25]	(2); focus on testimony	*n*-grams, surface features, syntactic features, textual complexity features, sentiment/polarity features	RF, FeedforwardNN, BERT
Regulation Room Divisiveness Corpus – User comments on Airline Passenger Rights rule from *RegulationRoom.org*	English	relations: *pro-arguments, con-arguments* and *rephrases of argument*	Konat, Lawrence et al. [41]	(3)	-	Two graph theoretical measures for divisiveness

eRulemaking_Controversy Corpus – User comments on Airline Passenger Rights rule from *RegulationRoom.org*	English	relations: *pro-arguments* and *con-arguments*	Lawrence et al. [48]	(3)	word features and grammatical features, e.g. discourse indicators and syntactic structure of an argument	semantic similarity, SVM, NB, rule-based classifier, a graph theoretical measure for centrality
Various user comments from U.S. eRulemaking online platform *regulations.gov* (annotated semi-automatically)	English	4 generic argument types: *opposition (explicit, likely), support (explicit, likely)* + 12 specific argument types: *burdensome*, *not sufficient type*, *lacks flexibility*, *conflicting interest*, *disputed information*, *legal challenge*, *overreach*, *requests clarification*, *seeks exclusion*, *lacks clarity*, *too broad, too narrow*	Eidelman & Grom [23]	(1) + (2)	*n*-grams, word embeddings	LR, fastText
THF Airport ArgMining Corpus - German language dataset of a citizen online participation in the restructuring of a former airport site	German	Argument components: *Claim*, *premise* and *major position*	Liebeck et al. [51]	(1) +(2)	*n*-grams, part-of-speech information, dependency information, structural properties	SVM, RF, *k*-NN
		Argument components: *Claim (pro/contra)*, *premise* and *major position*	Liebeck [50]	(1) +(2)	*n*-grams, word embeddings, part-of-speech information, dependency information, named entities, structural properties, topic information, sentiment features	SVM, RF, *k*-NN, CNN, LSTM, BiLSTM

Multiple transportation-related public participation processes (online platforms and survey data)	German	Argument components: *Premise* and *major position*	Romberg & Conrad [73]	(1) +(2)	*n*-grams, word embeddings, part-of-speech information, dependency information	SVM, fastText, ECGA, BERT
		Argument concreteness: *high, intermediate* and *low*	Romberg [72]	(2)	*n*-grams, text length (in tokens)	LR, SVM, RF, BERT
Online civic discussion data about the city of Nagoya	Japanese	Argument components: *claim* and *premise* relations: *inner-post* and *inter-post*	Morio & Fujita [59]	(1), (2) and (3)	*n*-grams, part-of-speech information, structural properties	SVM
			Morio & Fujita [60]	Joint model for (1), (2) and (3)	sequence of sentence representations	SVM, RF, LR, STagBiLSTMs. PN, PCPA
Citizen contributions of the 2016 Chilean constitutional process (local on-site events)	Spanish	Argument components: *policy, fact* and *value*	Fierro et al. [26]	(2)	*n*-grams, word embeddings, part-of-speech information	SVM, RF, LR, fastText, deep averaging networks
			Giannakopoulos et al. [28]	(2)	word embeddings	CNN, BiGRUs, Attention

D: Overview of sentiment analysis approaches

Table 4: Overview of sentiment analysis approaches

Dataset	Language	Granularity level	Tonality scale	Research article	Algorithms
Rulemaking process by the U.S. Environmental Protection Agency (online and offline submissions)	English	claim	discrete (claim attitude: *support*, *oppose*, *neutral/propose a new idea*)	Kwon et al. [45]	NB, heuristic decision rules
				Kwon & Hovy [44]	boosting
				Kwon et al. [46]	boosting
Contributions from the online consultation website *opengov.gr* and from a consultation held by the European Commission	Greek	comment	discrete (*strong disagreement, disagreement, agreement, strong agreement*)	Soulis et al. [85]	SVM
Crowdsourced urban planning process with focus on transportation (online)	English	suggestion	continuous (-1 to +1)	Aitamurto et al. [1]	Sentiment analysis tool (heavenondemand)
Urban planning project for the redesign of a major street	English	comment	five emotions (*excitement, happiness, neutral, concerned,* and *angry*)	Jasim et al. [37]	SVM, RF, CNN, LSTM, BERT

E: Overview of software solutions for practitioners

Table 5: Overview of software solutions for practitioners

Name of the tool / Description	Research article
system for information acquisition	Kwon & Hovy [44]
WICA (Workspace for Issue Categorization and Analysis)	Bruce et al. [9]
OntoCop	Yang & Callan [99]
system to support policy-making in online communities	Klinger et al. [40]
information visualization tool for surveys	Soulis et al. [85]
PIERINO (PIattaforma per l’Estrazione e il Recupero di INformazione Online)	Caselli et al. [14]
Civic CrowdAnalytics	Aitamurto et al. [1]
system for interactive topic modeling	Cai et al. [10]
interactive dashboard for the analysis of social media and e-participation data	Simonofski et al. [82]
information extraction and visualization modules for the open source platform Consul	Arana-Catania et al. [2]
CommunityPulse	Jasim et al. [37]

F: Literature database search

In order to identify those studies in the Association for Computational Linguistics Anthology that applied NLP to the relevant application area, the anthology was searched with multiple search terms: “public participation”, “online participation”, “political participation”, “civic participation”, “e-participation”, “public engagement”, “online engagement”, “political engagement”, “civic engagement”, “e-engagement”, “e-government”, “public consultation” and “e-rulemaking”.

The documents for the Digital Government Reference Library were narrowed to the application area of interest by using the search terms “participation”, “engagement”, “consultation” and “rulemaking”, and subsequently searched for studies that utilized the relevant technology by searching for the terms “natural language processing”, “nlp”, “text mining”, “text analysis”, “machine learning” and the more specific machine learning tasks “topic modeling”, “document categorization”, “classification”, “clustering”, “argument mining” and “sentiment analysis”.

3

Datasets

In the previous chapter, we examined prior research for its shortcomings. One finding was the strong focus on processing English-language participation contributions. As a result, supervised machine learning models have generally been developed and tested in a monolingual fashion drawing mostly on processes from U.S. rulemaking (Kwon et al., 2006; Cardie et al., 2008b; Park and Cardie, 2014; Konat et al., 2016; Niculae et al., 2017; Lawrence et al., 2017; Park and Cardie, 2018; Eidelman and Grom, 2019).

In terms of other languages, Fierro et al. (2017) developed models on a Spanish-language dataset of citizen contributions to the 2016 Chilean constitutional process annotated with argument types and constitutional concepts. Kim et al. (2021) dealt with the topic classification of Korean input on the Seoul City "Oasis of 10 Million Imagination" civic online participation platform. Morio and Fujita (2018a,b) had a look at Japanese online civic discussion data about the city of Nagoya, where they predicted argument structures, and Soulis et al. (2013) addressed sentiment analysis of Greek contributions in an online consultation held by the European Commission. There is also little annotated data on public participation in German that we came across during our search. Liebeck et al. (2016) shared the THF Airport ArgMining Corpus, a German-language dataset of an online participation for the restructuring of a former airport site annotated with arguments, and Balta et al. (2019) trained topic classification models on different online participation projects for urban development that were run in the city of Hamburg.

This dissertation is embedded in a research project that specifically aims to support German-language public participation processes. For this reason, the focus of this work is on the development of classification models suitable for German texts. One step along this path is the creation of new language resources to complement the few existing ones. While corpora from other application domains may already exist for related tasks, we chose to create domain-specific resources. In this way, we believe we can best meet the goals of the project by developing methods on actual participation data and testing their performance as close to the application as possible.

Paper: Julia Romberg, Laura Mark, and Tobias Escher. A Corpus of German Citizen Contributions in Mobility Planning: Supporting Evaluation Through Multidimensional Classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2874–2883. European Language Resources Association, 2022.

Personal Contribution: The coding tasks were developed and designed by Julia Romberg in collaboration with Laura Mark and Tobias Escher. Julia Romberg prepared the various datasets for annotation and also implemented the annotation process, which she supervised together with Laura Mark. She solely conducted all analyses included in the paper, namely calculation of the annotator agreement, error analysis, and corpus statistics. The manuscript was written jointly by the three co-authors.

Status: published

In interviews with practitioners from public administrations, participation service providers, and planning consultants, we identified four frequent tasks whose automation would aid in the evaluation of public participation, with particular emphasis on spatial planning and mobility (Romberg and Escher, 2020). Based on this knowledge, we developed the *CIMT PartEval Corpus*, which includes several thousand German-language citizen contributions from six planning processes in five municipalities.

The first task we consider is the detection and classification of argument components. We have a total of $17,852$ sentences annotated, using a scheme that categorizes sentences as non-argumentative, premise, or containing a major position. Unlike previous datasets, the corpus builds on a selection of processes that differ in format and process subject matter to specifically support the creation of robust classification models. As a second task, we focus on the concreteness of the argumentative components. To address this previously unstudied dimension of argument quality, we have $1,127$ arguments annotated according to a scheme of low, intermediate and high concreteness.

We then look at some of the special characteristics of public participation for spatial and mobility-related planning. Given that evaluating contributions for spatial planning requires these to be linked to the places addressed, we choose text-based document geo-location (i.e., determining the geographic coordinates of the associated location of a document based on its textual content) as a third task. To this end, $2,529$ contributions are tagged with $4,830$ location phrases and GPS coordinates. As a fourth task, we decide on the thematic classification of contributions into a category scheme for transportation that is not limited to individual processes but can be used for structuring all kinds of mobility-related planning processes. This results in 679 annotated contributions.

All annotation processes are performed by a total of five trained annotators and two supervising researchers using carefully developed guidelines. Solid inter-annotator agreement in all four tasks underscores the quality of the CIMT PartEval Corpus.

A Corpus of German Citizen Contributions in Mobility Planning: Supporting Evaluation Through Multidimensional Classification

Julia Romberg, Laura Mark, Tobias Escher
Heinrich Heine University Düsseldorf
Universitätsstraße 1, 40225 Düsseldorf
{julia.romberg, laura.mark, tobias.escher}@hhu.de

Abstract

Political authorities in democratic countries regularly consult the public in order to allow citizens to voice their ideas and concerns on specific issues. When trying to evaluate the (often large number of) contributions by the public in order to inform decision-making, authorities regularly face challenges due to restricted resources. We identify several tasks whose automated support can help in the evaluation of public participation. These are i) the recognition of arguments, more precisely premises and their conclusions, ii) the assessment of the concreteness of arguments, iii) the detection of textual descriptions of locations in order to assign citizens' ideas to a spatial location, and iv) the thematic categorization of contributions. To enable future research efforts to develop techniques addressing these four tasks, we introduce the *CIMT PartEval Corpus*, a new publicly-available German-language corpus that includes several thousand citizen contributions from six mobility-related planning processes in five German municipalities. The corpus provides annotations for each of these tasks which have not been available in German for the domain of public participation before either at all or in this scope and variety.

Keywords: public participation, argument mining, thematic categorization, location detection, spatial planning, mobility

1. Introduction

Public participation is the "practice of consulting and involving members of the public in the agenda-setting, decision-making, and policy-forming activities" (Rowe and Frewer (2004), p. 512). By enabling citizens to communicate their preferences on specific issues, it is an important element of representative democracies to improve responsiveness between the electorate and their representatives. While there is a debate about what role such consultative procedures can or indeed should play (Parry and Moyser, 1994), here we focus on the more practical issue of how to process and evaluate the input of citizens once public authorities have chosen to engage in such consultations. This has become a more pressing issue because of concerns about declining public support for democratic actors and institutions (Norris, 2011) as well as the easy availability of online forms of participation which has led to widespread use of public participation, regularly resulting in large numbers of contributions from citizens.

Processing the contributions from citizens poses significant challenges for public authorities because norms of democratic equality and administrative justice demand that every single contribution is carefully evaluated. While it is desirable that people participate in large numbers for increasing the acceptance and possibly the usefulness of the output, public administration (or the private companies tasked with evaluation) often lack personnel and time to deal with large quantities of unstructured citizen input (Arana-Catania et al., 2021; Aitamurto et al., 2016; Simonofski et al., 2021). As a result, the evaluation process often takes a long time, which can lead to delays in the planning process and to discontinuities in public communication, with all the associated negative consequences for efficiency, transparency and public acceptance.

Given that evaluation usually means categorizing input from citizens into different dimensions (e.g. according to topic, urgency or responsibility) before taking a decision on the individual contribution, one opportunity to support this manual evaluation process to make it more efficient is pre-structuring citizens' input. While some approaches focus on user-generated structuring, i.e. by letting citizens classify their contributions themselves, these allow only to categorize a limited number of dimensions (in order not to overburden users), and are limited by the lack of expertise of the lay public. Instead, here we focus on utilizing Natural Language Processing (NLP) that has been suggested as an alternative (OECD, 2003). Despite the relevance for democratic participation as well as significant progress in NLP techniques, automated classification of citizen contributions has yet to be advanced to a level sufficient to offer reliable support for practice.

Therefore, in this paper we propose four classification tasks in order to support the evaluation process. We provide datasets from six public participation processes in five German cities that have been annotated according to all or a part of those four dimensions to enable the training of supervised models for these tasks. Table 1 gives an overview of the *CIMT PartEval Corpus*.

In dialogue with practitioners from public administrations, participation service providers and planning consultants, we identified four common tasks whose support through automation would benefit the evaluation of participation processes (Romberg and Escher, 2020).

			datasets						
task	unit level	total units	CD_B	CD_C	CD_M	CQ_B	MC_K	MC_O	language resource reference
i) argument components	sentences	17,852	10,442	1,704	2,193	1,505	2,008		(Romberg et al., 2022a)
ii) argument concreteness	sentence spans	1,127	679	92	110	55	191		(Romberg et al., 2022b)
iii) geographic location	token spans	4,830	4,087	743					(Romberg et al., 2022c)
iv) thematic categorization	documents	697						697	(Romberg et al., 2022d)

Table 1: Overview of the coded units for the different tasks and datasets included in the CIMT PartEval Corpus.

These are i) the detection of arguments, ii) the assessment of the concreteness of arguments, iii) the recognition of locations that contributions refer to, and iv) structuring according to topics.

Individually and in combination with each other, these tasks can help to structure the data and thus facilitate the analysis in the following ways: The distinction into different **argument components** is important because it allows practitioners to get a quick overview of the relevant parts of the contributions. The recognition of **concreteness** enables practitioners to filter the most specific contributions, e.g. as a possible starting point for evaluation. The **recognition of locations** is helpful for processes without user-generated geo-referencing because it allows clustering contributions in spatial entities, e.g. to detect hot spots or assign responsibilities based on geographical jurisdictions. Finally, the **thematic categorization** helps to obtain a content-related overview fast and makes it possible to analyse contributions with similar topics together and therefore find patterns and contradictions more easily. What is more, it is the basis for delegation to those administrative units responsible for dealing with the contributions.

We have chosen to focus on one specific type of such participatory processes, namely those concerned with mobility such as the redesign of streets or the development of strategic mobility plans. Mobility planning is an important area within spatial planning in which consultations are regularly utilized. Structurally, these contributions are not different from participation processes on other issues but the focus on mobility allows us to provide a topic-specific categorization.

Our contributions are: We release a new annotated corpus (available under a Creative Commons License) for the development of supervised models to support the multidimensional evaluation of German-language public participation processes, consisting of six processes that differ in participation format and process focus. We provide annotations for the four described classification tasks. To the best of our knowledge, for some of the tasks, this is the first German-language (iii) or first-ever (ii, iv) annotated corpus from the domain of public participation. Particularly noteworthy are the new quality criterion for arguments (concreteness) and the thematic categorization scheme that is universally applicable to transport-related processes.

The remainder of this paper is as follows: In the next section, we review the existing language resources from the domain of public participation. We then present the public participation processes included in our corpus in Section 3. The four classification tasks are subsequently addressed in Section 4 (argument components), Section 5 (argument concreteness), Section 6 (geographic location), and Section 7 (thematic categorization). In each section, the task is introduced, followed by an overview of relevant work, a description of the annotation process and a presentation of the resulting dataset. Section 8 concludes with a summary and an outlook on future work.

2. Language Resources from Public Participation

In recent years, citizen contributions from different public participation processes have been annotated to support NLP research tasks, mainly the recognition of arguments and their properties as well as thematic categorization of citizen ideas. Most of these derived from rulemaking processes in the USA (Kwon et al., 2006; Arguello et al., 2008; Cardie et al., 2008; Park and Cardie, 2014; Konat et al., 2016; Aitamurto et al., 2016; Lawrence et al., 2017; Park and Cardie, 2018; Eidelman and Grom, 2019), some from processes in Chile (Fierro et al., 2017), Germany (Liebeck et al., 2016), Japan (Morio and Fujita, 2018) and Korea (Kim et al., 2021).

In the field of argument mining, the focus was especially in recognizing argumentation components and their supporting relations. Lawrence et al. (2017) and Konat et al. (2016) focused on the dialogical relation. Park and Cardie (2018) annotated comments with a more detailed scheme, in which propositions were subdivided into different types and then linked. A rather general argumentation scheme for informal online public participation processes was introduced by Liebeck et al. (2016). More specific is the adaptation to the thread structure of online platforms by Morio and Fujita (2018) who added intra-post and inter-post relationships. Probably the largest dataset was presented by Eidelman and Grom (2019), in which about 1.8 million sentences from various rulemaking efforts were semi-automatically assigned argument claim types.

Further work put the emphasis on the quality of citizens' arguments such as the verifiability of propositions (Park and Cardie, 2014). Arguello et al. (2008) proposed the recognition of citations in citizen comments to value them as factual evidence for claims and opinions.

Moreover, attention was paid to structuring citizens' ideas thematically. Cardie et al. (2008) and Aitamurto et al. (2016) focused on thematic categorization of transportation-related rulemaking processes by developing customized categorization schemes. A somewhat different approach to thematic categorization was taken by Kim et al. (2021) who assigned complaints that were submitted to a civic online participation platform to respective administrative fields.

Only a few datasets were coded according to multiple viewpoints. One is that of Kwon et al. (2006), whose multidimensional coding included thematic categorization and the analysis of argument structure. In Fierro et al. (2017), a large-scale dataset of citizen arguments collected during Chile's 2016 constitutional process was presented. Arguments were categorized according to their function and thematically organized into a hierarchy of constitutional concepts.

In summary, there exists only a single German dataset for the domain of public participation and this focuses only on argument mining within a single process (Liebeck et al., 2016). On thematic categorization of citizen ideas we find only a few corpora even for other languages (mainly English). None address concreteness or geographic location and few offer annotations representing multiple dimensions.

To address this gap, we present a collection of German-language datasets coded according to several dimensions, namely i) argument components, ii) concreteness of arguments, iii) location detection, and iv) thematic categorization, since there are no existing (German-) language resources in our application domain for the latter three tasks.

3. Datasets

We consider six different public participation processes in our data collection, namely three "Raddialoge" ("Cycling Dialogues") in the cities of Bonn, Cologne (district Ehrenfeld) and Moers as well as "Leben in Bonn" ("Living in Bonn"), "Krefeld bewegen" ("Moving Krefeld"), and Hamburg's "freiRaum Ottensen" ("Space for Ottensen"). While these are all related to urban mobility planning, they span different mobility-related issues and participation formats.

In detail, the three "Raddialog" datasets derive from largely identical participation processes conducted in autumn 2017 in which the local authorities invited their citizens to propose measures to improve cycling in the city. A map-based online platform allowed citizens to locate their contributions on a map, resulting in $2,314$ unique contributions consisting on average of 4.83 sentences (standard deviation $\sigma = 2.63$) for **Raddialog Bonn** (henceforth CD_B), 366 contributions (4.66 sentences, $\sigma = 3.00$) for **Raddialog Ehrenfeld** (CD_C) and 459 contributions (4.78 sentences, $\sigma = 2.61$) for **Raddialog Moers** (CD_M). In addition, in Bonn the online platform was supplemented with a representative survey of the population. In total, 761 citizens expressed up to three suggestions for improvement either via the paper-based questionnaire or an online alternative, resulting in $1,386$ contributions (1.09 sentences, $\sigma = 0.37$) for **"Leben in Bonn"** (CQ_B).

Within **"Krefeld bewegen"** (MC_K) the city of Krefeld invited citizen comments on the development of a mobility concept. The first phase in 2020 focused on general aims of the new concept and the second phase invited suggestions for specific measures. This resulted in 337 contributions (5.96 sentences, $\sigma = 5.63$).

The most recent dataset included in the corpus derives from a public participation process by the district of Altona in Hamburg (**"freiRaum Ottensen"**, MC_O). As part of the transformation of its quarter Ottensen into a traffic-calmed neighborhood, the district office implemented a map-based online dialogue that took place in August 2021. In total, it received 697 contributions (4.95 sentences, $\sigma = 2.49$).

All datasets were separately examined by service providers as well as our team and any potentially identifying personal information was removed. The data in the corpus is available under a Creative Commons CC BY licence and may be distributed in accordance with the corresponding conditions. Users of the online participation platforms accepted these conditions via the terms of use of these platforms, while the data originating from the questionnaires was released under this licence by the principal investigator of the survey.

4. Sentence-level Argument Components

A central aspect through which citizens communicate their ideas are arguments. Automated analysis of arguments, known as argument mining, enables practitioners to get a quick overview of relevant text passages. We here focus on two common tasks in argument mining, namely the identification of argument components and the identification of clausal properties (Lawrence and Reed, 2019). Part of our corpus for argument component analysis (described in this section) has previously been introduced in Romberg and Conrad (2021).

4.1. Related Work

Previous work in our application domain either followed the classic claim-premise model (Liebeck et al., 2016; Morio and Fujita, 2018), or had a stronger focus on the intrinsic characteristics of claims (Fierro et al., 2017; Park and Cardie, 2018), e.g. if claims are factual, contain values or propose policies. For more detail on related work, please see Romberg and Conrad (2021).

Our work is closest to that of Liebeck et al. (2016) whose THF Airport ArgMining Corpus is the only German-language public participation dataset for argument mining. However, there are several differences between the corpora: First, we provide seven times more sentences coded with argument components. Second, our focus is not on the dialogue structure within each thread but on the detection of propositions within the initial contributions. Third, our corpus comprises several processes differing in format and

		CD_B	CD_C	CD_M	MC_K	CQ_B	all
	total	10,442	1,704	2,193	2,008	1,505	17,852
	non-arg	1,153 (11.0%)	197 (11.6%)	382 (17.4%)	431 (21.5%)	172 (11.4%)	2,335
arg	mpos	2,851 (27.3%)	603 (35.4%)	404 (18.4%)	961 (47.9%)	1,083 (72.0%)	5,902
	premise	6,700 (64.2%)	951 (55.8%)	1,452 (66.2%)	685 (34.1%)	373 (24.8%)	10,161
	overlap	262 (2.5%)	47 (2.8%)	45 (2.1%)	69 (3.4%)	123 (8.2%)	546

Table 2: Distribution of sentences among the different argument component categories per dataset.

		CD_B	CD_C	CD_M	CQ_B	MC_K	all
	sentences	1,251	191	230	188	376	2,236
kappa	non-arg	0.58	0.68	0.67	0.59	0.69	0.63
	mpos	0.82	0.81	0.81	0.73	0.78	0.81
	premise	0.82	0.87	0.83	0.83	0.77	0.84
	overall	0.76	0.80	0.77	0.72	0.73	0.77

Table 3: Number of sentences under consideration and kappa agreement for argument component annotation.

process subject but all coded with a uniform coding scheme. This enables a comprehensive evaluation of machine learning methods, also with respect to the transferability of trained models to new processes, allowing robust models to be developed. Such cross-dataset evaluation is important to better assess the practical applicability of models.

4.2. Definition of Argument Components

Public participation allows citizens to contribute to a decision-making process by proposing their ideas and voicing concerns. In spatial planning processes, this usually involves describing a problem or condition, from which a proposition is derived. We thus define two types of argument components: *Major positions* (short *mpos*) are options for actions that are being proposed. *Premises* are reasons that attack or support either a major position or another premise. With this, we adopt one part of the argumentation scheme of Liebeck et al. (2016). Sentences without premise or major position are considered as *non-argumentative* (*non-arg*).

4.3. Annotation Study

First, we developed annotation guidelines based on 151 contributions from the dataset CD_B. Subsequently, the remaining contributions, as well as all contributions from CD_C, CD_M, CQ_B and MC_K were coded. Three annotators were instructed to decide for each sentence (titles included) whether it has argumentative content and, if yes, if it is a *major position* or a *premise*. Since some sentences contain components of both types, multi-labeling was allowed.

To assess coder agreement on this task, about ten percent of each dataset was processed by all the coders. This sums up to 585 contributions with 2,236 sentences. The agreement on these sentences was measured using Fleiss (1971)' kappa.

With an overall agreement of 0.77[1], the coding can be considered reliable (see Table 3). However, there was a greater uncertainty in the selection of non-argumentative sentences, while the agreement between the two types of argument components was rather high. In a subsequent curation phase, the sentences with inconsistent coding were reviewed and resolved by two annotation process supervisors. This showed that there were regular misclassifications of whether a sentence was indeed argumentative, with coders being more inclined to classify argumentative sentences as non-argumentative than vice versa. Furthermore, it can be seen within the argumentative sentences that the assignment of premises was more accurate than that of major positions.

Due to the considerable time required for multiple coding and given the high reliability, we decided to have the remaining 4,126 contributions with 15,616 sentences coded only once, evenly distributed among the coders.

4.4. Corpus Statistics and Discussion

The resulting distribution of sentences among the annotation classes is given in Table 2. Overall, the share of sentences without argumentative content is small. Depending on the process, 80 to 90 percent of sentences are argumentative. However, the distribution of argument component types varies greatly between the different processes. Premises clearly predominate in the cycling dialogues, while the other two processes seem to be more conclusion-oriented and favor major positions. This is particularly evident in the survey data where participants had limited space for writing suggestions. For online platforms, few sentences contain both a major position and a premise (overlap). In contrast, in the survey data there is a greater overlap of argument components, which nonetheless affects less than one in ten sentences. The variety of the processes included in the corpus results in very different class distributions, supporting the development of robust machine learning models.

5. Argument Concreteness

We then focus on the concreteness of the argumentative components, the automated evaluation of which can help practitioners filter out arguments that can be evaluated immediately. The less specific citizens' ideas are, the more difficult and hence time-consuming it will be for evaluators to derive measures for implementation.

[1] In the overall calculation, sentences containing both major position and premise constitute an additional category.

5.1. Related Work

The evaluability of public participation contributions has previously been raised by Park and Cardie (2018), Park and Cardie (2014) and Arguello et al. (2008), who saw the lack of reasoning and evidence verifying citizen contributions as the main obstacle to evaluating propositions. However, in the evaluation of spatial planning processes, we consider the level of concreteness of the arguments (i.e. how detailed current conditions and proposed improvements are described) as the most important indicator for evaluability.

To the best of our knowledge, we are the first to provide a resource for this type of concreteness of arguments, while other aspects of the *quality of arguments* have received increasing attention in recent years (e.g. Habernal and Gurevych (2016a), Habernal and Gurevych (2016b), Toledo et al. (2019), Gretz et al. (2020)). A systematic taxonomy of dimensions for argument quality, regarding logic, rhetoric and dialectic aspects, can be viewed in Wachsmuth et al. (2017).

5.2. Definition of Degrees of Concreteness

We propose a distinction between high, intermediate and low concreteness. Argument components are *highly concrete* when they contain details that specify the **what**, **how** and **where**. Such specifications can be colour, surface, measurements, etc. (an example is "cycle paths often in poor condition, tarred surface torn up, bumpy due to roots, mostly only half width because overgrown"). Contributions with an *intermediate concreteness* contain some specifications like location or descriptions of what exactly should be done, but leave some room for interpretation (e.g.: "new cycle lanes without interruptions" - the measure is described and somewhat detailed, but it is not clear how it should look exactly and where it should be located). Contributions with *low concreteness* contain no information on location or specific measures, so that a variety of measures could be deducted (e.g.: "unfavorable traffic lights" - it does not become clear, what exactly the problem is, where it is and what should be done).

Distinction of concreteness was applied only to argumentative components, non-argumentative sentences were excluded. In order to support different use cases, such as searching either for (concrete) major positions or (concrete) premises, we consider the concreteness of the two types of argument components separately.

5.3. Annotation Study

We decided to use the curated documents of the previous annotation task in order to ensure the soundness of the annotation of sentence-level argument components. Determining the concreteness of solitary sentence-level argument components is hardly feasible. Therefore, the coders first interrelated argument components of the same types (i.e. premises or major positions) to form units with coherent sense, and the annotation supervisors resolved inconsistencies. In a second step, we asked coders to rate the resulting units' concreteness using guidelines that were developed on the same data as with argument components.

	CD_B	CD_C	CD_M	MC_K	CQ_B	all
mpos	265	40	40	126	42	513
premise	414	52	70	65	13	614
total	679	92	110	191	55	1,127

Table 4: Units of interrelated argument components.

It turned out that the perception of concreteness is rather subjective, which was also confirmed to us by those responsible for analyzing the contributions. We thus decided to include a total of five annotators to obtain a multitude of individual concreteness ratings. Due to the subjective nature, we dispense with a manual curation step in which an unambiguous assignment of concreteness to units is made, but instead release the five individual codings. While the assessment of concreteness exhibits some subjectivity, it is not arbitrary as is documented by Krippendorff (2013)'s weighted alpha[2], which shows an agreement score of 0.46.

5.4. Corpus Statistics and Discussion

Overall, 513 units of interrelated sentences containing major positions and 614 units of interrelated sentences containing premises were formed and coded by concreteness (see Table 4). To each of the units belong five codings by the different annotators. There is complete agreement among coders in 478 cases, about 42 percent of the units. In the majority of disagreements, coders chose adjacent categories, so while subjective perception differs slightly, there is a consistent trend in whether the unit (460 in total) is rather concrete or vague. Within 189 units, however, a strongly subjective assessment is evident, in which all or the two opposing degrees of concreteness were assigned.

Analysis of the degrees of concreteness reveals that citizens clearly tend to write highly concrete arguments in the processes considered here. Nevertheless, on average about twenty percent of the argument units have intermediate or low concreteness, thus automated recognition will allow highlighting the most relevant (concrete) content.

6. Geographic Location

In spatial planning processes, the geographic location of citizens' contributions is of great importance to the evaluation as it allows geo-referencing of contributions and clustering of ideas by location. Map-based processes on online platforms offer a possibility in which citizens can locate their ideas on a map. However,

[2] We weight using the Euclidean distance to account for the level of deviation between the codings, i.e., whether they are adjacent (e.g., low/high and medium concreteness) or non-adjacent categories (low and high concreteness).

not all public participation in spatial planning is georeferenced as exemplified by the survey-based data (CQ_B) in our corpus. To address this problem we propose the use of text-based geo-location and present a dataset of textual locations and GPS coordinates.

6.1. Related Work

Text-based document geo-location is the task of determining the geographic coordinates of a document's associated location by its textual content. Originally a task from information retrieval, it combines language modelling and geographical information science.

This task was initially approached through clustering. Much of these works relied on named entity recognition to narrow the feature space to geographical indications (e.g. Smith and Crane (2001)). Other approaches relied on more unsupervised vocabulary selection strategies (e.g. Adams and Janowicz (2012), Wing and Baldridge (2014)). Putting a stronger focus on natural language processing and supervised learning, the recognition of textual location phrases was supported by the development of specified annotated corpora. McNamee et al. (2020), for example, concentrated on fine-grained tagging of location phrases that complement named entity mentions with additional words which provide further information to specify a location (e.g. prepositions).

Further work directly combined the recognition of location information with a subsequent geo-coding step to associate the textual locations to GPS coordinates. Application domains were, inter alia, textual narratives from travel blogs (Skoumas et al., 2016) and news articles to map the local news coverage (Gupta and Nishu, 2020).

With public participation processes, we here introduce a new application domain that differs from previously targeted genres in document length, text quality, and prevalence of location, among other factors. Our use case requires a very precise mapping to pinpoint geo-coordinates, with location information as accurate as streets, intersections, and addresses.

6.2. Definition of Location Phrases

We define a textual *location* as a single word or a sequence of words included in a citizen's contribution that refers to the spatial placement of the respective contribution. These can be named entities, such as street names or city districts, but also, beyond that, constructions with more fine-grained location information that can be unambiguously marked on a map. Such phrases usually contain information that specifies the exact location, like the description of a specific angle (e.g. approaching some location from the right-hand side, or in the direction of the main station).

A known problem in determining the geo-positions from textual descriptions are ambiguous locations (e.g. Awamura et al. (2015), Smith and Crane (2001)). This includes, for example, street names, squares, or stations (like main station) without assignment to a city. For our use case, many of these cases are solved by the fact that the context in which the processes take place is usually known. Furthermore, we do not understand a word sequence as a location if it refers to several places in the city ("many/various/all parks in the city") or does not have a spatial reference point that specifies its geo-location (like "in the one-way street").

6.3. GPS Coordinates

The next step following the recognition of textual locations is the assignment to GPS coordinates based on the location phrases.

We chose the cycling dialogues (CD_B, CD_C) for the text-based document geo-location task because an assignment of GPS coordinates had already been part of the map-based online platforms, where each citizen was requested to explicitly indicate the location of their contribution as a point on the map. More complex shapes such as polygons were not allowed. We can assume that the textual location descriptions and the geo-locations given refer to the same entity, since citizens generally adhered to the requirements of point-wise referencing, and that the textual description should belong to the geo-referenced location. GPS coordinates are thus included in our annotated data corpus alongside the location phrases.

6.4. Annotation Study

Three trained annotators were instructed to identify the textual location spans within $2,529$ contributions from CD_B and CD_C. The coding guidelines were previously developed on additional 151 contributions from CD_B. Each location unit could consist of any number of consecutive words, but units could not cross sentence boundaries. 305 contributions, about ten percent of each dataset, served to determine the inter-annotator agreement and the remaining $2,224$ contributions were divided equally among the annotators. After calculating the inter-annotator agreement, documents with multiple annotations were reviewed by two supervisors and conflicts were resolved to obtain a unified coding.

We consider Krippendorff et al. (2016)'s alpha for unitizing textual continua[3] to evaluate the reliability of the coders. The alpha measure of 0.75 proves a high agreement between the coders. We assume that the coders worked as reliably on the contributions that were single-coded.

A look at the contributions with multiple codings shows that disagreements in the handling of prepositions (e.g. along, across, into, left/right of) occurred repeatedly. Another source of disagreement were nouns (e.g. bike lane, one-way street, sidewalk) at the beginning of location units. According to our guidelines, the coders had to decide whether additional words made

[3] We use the modified version of earlier definitions (Krippendorff, 1995; Krippendorff, 2013), which corrects shortcomings for studies with more than two annotators.

the location more precise. It turned out that perceptions did not always coincide on this.

6.5. Corpus Statistics and Discussion

The corpus comprises $2,529$ contributions, each of which is assigned to a GPS coordinate, and these contributions contain $4,830$ location phrases. The length of the location phrases varies from a single to up to 36 tokens, with on average 4.9 tokens ($\sigma = 3.48$). Examples for very short locations are street names or districts (e.g. downtown), while longer units contain more precise descriptions.

Overall, about twelve percent of the tokens included in the contributions are part of a location phrase, a proportion that further illustrates the relevance of automated location of citizen ideas for spatial planning processes.

7. Thematic Categorization

Lastly, we address the thematic categorization of citizen contributions in our data corpus. This makes it possible to analyse contributions with similar topics together and detect patterns as well as to delegate contributions to the responsible administrative units.

7.1. Related Work

Content structuring by thematic categories has been addressed before, including by Kwon et al. (2006) for a mercury rulemaking process and by Fierro et al. (2017) in the context of a constitutional process. Cardie et al. (2008) and Aitamurto et al. (2016), like us, focused on transportation-related processes.

A problem shared by previous work is that the categories were fitted to the individual participation process. Such specification makes the development of supervised classification models for real-world use (i.e. beyond research purposes) impractical. If schema and training data have to be developed from scratch for each new process, the time required may quickly exceed the effort of a purely manual analysis, especially for processes with fewer contributions. This problem has previously been described by Purpura et al. (2008), who proposed active learning to reduce the amount of training data. Still, the amount of training data needed for an adequate prediction quality may remain high.

An alternative solution is to use categorization schemes that are universally applicable to multiple participation processes. These can be used to train models which can subsequently be applied to further processes without the need for additional training. An example is the work of Kim et al. (2021), in which contributions were assigned to the competent administrative fields (e.g. housing, culture, environment) based on a guideline for governments. We follow this example and define a universal scheme of transportation-related categories that is not limited to individual processes but can be used for structuring all kinds of mobility-related planning processes.

7.2. A Categorization Scheme for Mobility

We propose a category scheme that covers modes of transport as well as related aspects and allows multi-labeling.

The categorization scheme was developed based on a variety of sources including existing mobility concepts (e.g. Der Senator für Umwelt, Bau und Verkehr (2014)), categorizations proposed in documentations of participation processes (e.g. Zebralog (2020)), and topic choices currently available to users of online consultations[4]. This draft was then subjected to feedback from experts with practical experience in the evaluation of contributions, namely representatives of participation service providers, planning offices and administration, and subsequently improved.

Figure 1 provides an overview of **modes of transport**, almost always relevant in mobility-related processes, and their **specifications**. Please note that it is also possible for a contribution not to be assigned to any mode.

Regarding modes of transport, it is firstly specified if the contribution deals with *motorized* or *non-motorized transport* (or both). If the contribution explicitly refers to particular modes, these are then further specified: non-motorized modes are *cycling*, *walking* and *scooters*. Motorized modes encompass *local* and *long-distance public transport* as well as *commercial transport* which includes, e.g., delivery and waste disposal. Private cars are not included as a separate sub-category of motorized transport. Instead, relevant contributions will be subsumed under motorized modes because even when contributions refer specifically to "cars", the issues usually concern all motorized modes - even if this is not explicitly stated, e.g. when criticizing traffic signaling. As a matter of fact, there are hardly any issues that refer exclusively to private car traffic[5].

Only if the contribution concerns a mode of transport, it can then be assigned to one or more **specifications** such as the type of traffic (*moving traffic* or *stationary traffic*, i.e. parking). What is more, the categories of *new services* and *inter- and multimodality* can be added as supplementary information to the mode of transport, the first referring to technological advancements like e-mobility or app-based offers, the second referring to the connection of and between different modes of transport, like intermodal booking systems or the design of interchanges.

This nested system of categories allows both a general and a more specific classification of the data. The possibility to assign more than one topic to a contribution is an essential difference to most user-generated structuring approaches in online consultations. This multi-labeling is often necessary because contributions can deal with more than one topic.

[4] E.g., see the participation tool of service provider "tetraeder": www.buergerbeteiligung.de/ beispielhausen/

[5] An exception is residential parking, which can be identified through the specification "stationary traffic".

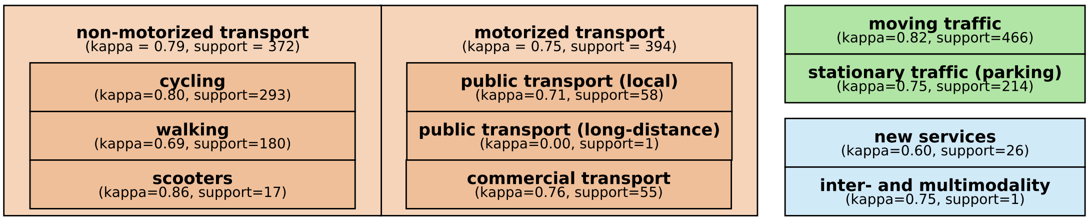


Figure 1: Overview of thematic categorization scheme for mobility. Numbers in parentheses denote inter-annotator agreement (Fleiss' kappa) and class support after solving disagreements.

7.3. Annotation Study

We started annotation with MC_O, a process that aims at a comprehensive mobility concept and therefore includes contributions on various modes of transport. The 697 contributions were coded by three coders according to our hierarchical scheme. Detailed coding guidelines were developed and the coders were trained on contributions from MC_K and further processes not part of the collection presented here. Since it became apparent during the coding process that some categories occurred much less frequently than others, we decided to have each document coded by all coders.

To analyze the reliability of the codings we calculated the Fleiss' kappa agreement for the categories reported in Figure 1. Most categories show a rather high level of agreement of 0.75 and above. Some categories with lower agreement such as long-distance public transport or inter- and multimodality suffer from very few contributions identified as belonging to this category (see next section), which is why the significance of kappa should be viewed with caution here. A subsequent screening and revision of the disagreements by two supervisors, one an urban planner, led to a final unique coding, which is the one presented in the following.

7.4. Corpus Statistics and Discussion

The class support of the final coding is depicted in Figure 1. About 82 percent of the contributions were about motorized or non-motorized transport, with moving traffic prevailing over stationary traffic. The optional categories new services and inter- and multimodality hardly occurred in the process under consideration, just as scooters and long-distance public transport.

These categories remain in the categorization schema as our aim is to provide a comprehensive scheme for all modes of transport, independent of this specific process. In other processes, we expect a different distribution of the classes. In order to provide a sufficient data basis for the development of generally valid classification models, including minority classes, the coding of further processes is scheduled.

18 percent of the documents (126) were assigned to none of the mobility-related categories; those mainly focused on other requirements for public space (e.g. noise, accessibility, quality of stay). Such requirements will be included as additional dimensions in further scheme development.

8. Conclusion and Future Work

When public authorities consult the public, they have to ensure that all contributions are properly considered. In order to support this process that is vital to democratic participation yet costly in terms of resources, we have identified four classification tasks and introduced a new publicly-available German-language corpus.

Our corpus is the first German-language corpus in the domain of public participation that provides annotations of textual and GPS locations, as well as a thematic categorization for modes of transport. Furthermore, it provides annotations to distinguish argument components and their concreteness. In contrast to the previous datasets on argument mining for public participation, this corpus contains six different datasets varying in participation format (online platform vs. questionnaire) and issue. This enables the training of more transferable and robust machine learning methods.

Efforts to develop NLP models to solve the practical application tasks can now rely on this corpus. While it consists of mobility-related processes, its application is not limited to such issues as with the exception of thematic categorization, the tasks are generic to participation processes. The thematic categorization scheme is universally applicable in the mobility section.

Currently we are extending the annotation of the present corpus, as well as adding new datasets in order to increase diversity and representation of minority classes. What is more, we are working on expanding the thematic categorization scheme with additional dimensions (e.g. quality of public space, traffic safety or noise pollution). We have started to develop classification models for these four tasks based on the annotated corpus. A first model for the detection and classification of argument component detection has been introduced in Romberg and Conrad (2021). Our ultimate goal is to provide an open source application that supports public authorities in the evaluation of public participation contributions.

9. Acknowledgements

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.

10. Bibliographical References

Adams, B. and Janowicz, K. (2012). On the geo-indicativeness of non-georeferenced text. In *Proceedings of the Sixth International Conference on Weblogs and Social Media (ICWSM'12)*, pages 375–378. AAAI Press.

Aitamurto, T., Chen, K., Cherif, A., Galli, J. S., and Santana, L. (2016). Civic CrowdAnalytics: Making sense of crowdsourced civic input with big data tools. In *Proceedings of the 20th International Academic Mindtrek Conference*, pages 86–94. Association for Computing Machinery.

Arana-Catania, M., Lier, F.-A. V., Procter, R., Tkachenko, N., He, Y., Zubiaga, A., and Liakata, M. (2021). Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice*, 2(3):1–22.

Arguello, J., Callan, J., and Shulman, S. (2008). Recognizing citations in public comments. *Journal of Information Technology & Politics*, 5(1):49–71.

Awamura, T., Kawahara, D., Aramaki, E., Shibata, T., and Kurohashi, S. (2015). Location name disambiguation exploiting spatial proximity and temporal consistency. In *Proceedings of the Third International Workshop on Natural Language Processing for Social Media*, pages 1–9. Association for Computational Linguistics.

Cardie, C., Farina, C., Rawding, M., and Aijaz, A. (2008). An eRulemaking corpus: Identifying substantive issues in public comments. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 2757–2763. European Language Resources Association.

Der Senator für Umwelt, Bau und Verkehr. (2014). *Verkehrsentwicklungsplan Bremen 2025*. Bremen.

Eidelman, V. and Grom, B. (2019). Argument identification in public comments from eRulemaking. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 199–203. Association for Computing Machinery.

Fierro, C., Fuentes, C., Pérez, J., and Quezada, M. (2017). 200K+ crowdsourced political arguments for a new Chilean constitution. In *Proceedings of the 4th Workshop on Argument Mining*, pages 1–10. Association for Computational Linguistics.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., and Slonim, N. (2020). A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Gupta, S. and Nishu, K. (2020). Mapping local news coverage: Precise location extraction in textual news content using fine-tuned BERT based language model. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 155–162. Association for Computational Linguistics.

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

Habernal, I. and Gurevych, I. (2016b). Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599. Association for Computational Linguistics.

Kim, B., Yoo, M., Park, K. C., Lee, K. R., and Kim, J. H. (2021). A value of civic voices for smart city: A big data analysis of civic queries posed by Seoul citizens. *Cities*, 108:102941.

Konat, B., Lawrence, J., Park, J., Budzynska, K., and Reed, C. (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*, pages 3899–3906. European Language Resources Association.

Krippendorff, K., Mathet, Y., Bouvry, S., and Widlöcher, A. (2016). On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, 50(6):2347–2364.

Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Marsden, P.V. (ed.) Sociological Methodology*, 25:47–76.

Krippendorff, K. (2013). *Content Analysis: An Introduction to Its Methodology*. Sage publications.

Kwon, N., Shulman, S. W., and Hovy, E. (2006). Multidimensional text analysis for eRulemaking. In *Proceedings of the 2006 international conference on Digital government research*, pages 157–166. Digital Government Society of North America.

Lawrence, J. and Reed, C. (2019). Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Lawrence, J., Park, J., Budzynska, K., Cardie, C., Konat, B., and Reed, C. (2017). Using argumentative structure to interpret debates in online deliberative democracy and eRulemaking. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–22.

Liebeck, M., Esau, K., and Conrad, S. (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*, pages 144–153. Association for Computational Linguistics.

McNamee, P., Mayfield, J., Costello, C., Bishop, C., and Anderson, S. (2020). Tagging location phrases in text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4521–4528. European Language Resources Association.

Morio, G. and Fujita, K. (2018). Annotating online civic discussion threads for argument mining. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 546–553. IEEE.

Norris, P. (2011). *Democratic Deficit : Critical Citizens Revisited*. Cambridge University Press, Cambridge, GBR.

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

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

Park, J. and Cardie, C. (2018). A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association.

Parry, G. and Moyser, G. (1994). More participation, more democracy? In David Beetham, editor, *Defining and Measuring Democracy*. Sage, London.

Purpura, S., Cardie, C., and Simons, J. (2008). Active learning for e-rulemaking: Public comment categorization. In *Proceedings of the 9th Annual International Digital Government Research Conference*, pages 34–243. Digital Government Society of North America.

Romberg, J. and Conrad, S. (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. Association for Computational Linguistics.

Romberg, J. and Escher, T. (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.

Rowe, G. and Frewer, L. J. (2004). Evaluating public-participation exercises: A research agenda. *Science, Technology, & Human Values*, 29(4):512–556.

Simonofski, A., Fink, J., and Burnay, C. (2021). Supporting policy-making with social media and e-participation platforms data: A policy analytics framework. *Government Information Quarterly*, 38(3):101590.

Skoumas, G., Pfoser, D., Kyrillidis, A., and Sellis, T. (2016). Location estimation using crowdsourced spatial relations. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 2(2):1–23.

Smith, D. A. and Crane, G. (2001). Disambiguating geographic names in a historical digital library. In Panos Constantopoulos et al., editors, *Research and Advanced Technology for Digital Libraries*, pages 127–136. Springer Berlin Heidelberg.

Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., and Slonim, N. (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*, pages 5625–5635. Association for Computational Linguistics.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics.

Wing, B. and Baldridge, J. (2014). Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 336–348. Association for Computational Linguistics.

Zebralog, (2020). *Integriertes Mobilitätskonzept der Stadt Krefeld: Dokumentation der zweiten Online-Beteiligungsphase (24.01. - 21.02.2020).* Berlin, Bonn.

11. Language Resource References

Romberg, J., Mark, L., and Escher, T., (2022a). *CIMT PartEval Corpus - Argument Components (Subcorpus).* ISLRN 484-558-142-596-7. https://github.com/juliaromberg/cimt-argument-mining-dataset.

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

Romberg, J., Mark, L., and Escher, T., (2022c). *CIMT PartEval Corpus - Geographic Location (Subcorpus).* ISLRN 951-974-499-316-4. https://github.com/juliaromberg/cimt-geographic-location-dataset.

Romberg, J., Mark, L., and Escher, T., (2022d). *CIMT PartEval Corpus - Thematic Categorization (Subcorpus).* ISLRN 441-856-914-941-8. https://github.com/juliaromberg/cimt-thematic-categorization-dataset.

4

Topic Classification

The rise of pre-trained language models has led to significant improvements in the longstanding area of text classification in recent years (Howard and Ruder, 2018; Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020). These language models gain their advantage from initial training with massive quantities of unlabeled text. Supported by the knowledge gathered through pre-training, models can subsequently be adapted more efficiently to downstream tasks using task-specific training data. For a more detailed overview of pre-trained language models and how they work, we refer the interested reader to the surveys of Minaee et al. (2021) and Qiu et al. (2020).

Despite the gains in model fit due to encoded knowledge, fine-tuning still requires a non-negligible amount of data to ensure stable performance (Dodge et al., 2020). Practical applications that are subject to budget constraints may therefore face a significant challenge in applying pre-trained language models (e.g., Purpura et al., 2008; Searle et al., 2019; Schröder et al., 2021). Such budget constraints can be manifold, including scarce financial resources that prevent extensive human annotation (particularly when well-qualified experts are required), shortage of time to annotate large amounts of data, as well as environmental considerations, such as the reduction of computational costs and thus carbon emissions (Strubell et al., 2019).

Active learning offers a solution to meet these requirements and following the introduction of pre-trained language models, a growing body of literature has explored the potential of coupling both methods (e.g., Tamkin et al., 2022; Margatina et al., 2022; Karisani et al., 2022; Hua and Wang, 2022). As one of the first studies in this field, Ein-Dor et al. (2020) evaluated how a variety of query strategies behave in cooperation with the pre-trained language model BERT in balanced and imbalanced real-world scenarios of binary classification. Subsequently, further cases of text classification have been examined, including multi-class (Prabhu et al., 2021), multi-label (Wang and Liu, 2023), and multi-task (Rotman and Reichart, 2022) setups. Moreover, research has focused on the merits and limitations of specific families of query strategies[1] when combined

[1] Zhang et al. (2022b) provide a good overview of types of query strategies in their comprehensive survey of active learning for natural language processing.

with pre-trained language models (Schröder et al., 2022; Snijders et al., 2023), as well as on individual query strategies (Yuan et al., 2020; Margatina et al., 2021; Zhang et al., 2022a).

This progress is also of considerable significance for the evaluation of public participation, where one of the most important steps is to organize citizen contributions according to the issues raised. The motivation behind it is to make the unstructured collection more manageable, to get an initial overview of the issues that citizens are concerned about within the process, and then to forward smaller bundles of comments to the responsible agencies or officers for detailed analysis. Previous work has proposed to support this step with topic classification, a specific application of text classification (e.g., Kwon et al., 2006; Cardie et al., 2008a,b; Balta et al., 2019; Kim et al., 2021).

Most of the studies relied on traditional machine learning algorithms like support vector machines and random forests. The promising advancement brought about by pre-trained language models has received little attention so far (Balta et al., 2019). However, the topic classification of citizen contributions requires solutions that are not only accurate but also cost-effective, given the often limited financial or human resources available in the public sector, and also the need for process-specific classification schemes (as argued in detail in Section 1.2.1). In addition, it must be taken into account that participation processes regularly generate small datasets with hundreds to a few thousand contributions. In these cases, automatic support for topic classification is only worthwhile if the manual annotation of the required training data means a significant reduction in effort compared to a complete manual evaluation. This is where active learning comes in (Purpura et al., 2008).

Like Purpura et al. (2008), we see active learning as a methodological approach that opens the door to the advantageous practical application of topic classification in our use case. As we have outlined above, integrating pre-trained language models into active learning workflows offers the prospect of excelling text classification accuracy along with a significant reduction of required training data. Surprisingly enough, the potential of this symbiotic relationship has not yet been explored for the topic classification of public participation contributions. Therefore, we set the focus of this chapter accordingly. In Section 4.1, we first investigate the benefits of pre-trained language models in combination with active learning to reduce human effort in topic classification of citizen contributions.

In Section 4.2, we delve deeper into the evaluation of active learning by scrutinizing standard performance measures in terms of their meaningfulness for practice. This is motivated by the fact that assessing the practical value of active learning strategies is challenging (Kottke et al., 2017). Researchers have typically focused on overall predictive performance in terms of a task-specific measure (such as accuracy or F_1 measure in text classification tasks) and, in the age of large models, computation times. However, the factors to be considered for the usefulness of active learning in real-world scenarios are far more diverse (Margatina and Aletras, 2023). One of these facets involves the human factor within active learning (Baldridge and Palmer, 2009; Donmez and Carbonell, 2008; Calma and Sick, 2017). In this context, we propose additional performance measures specifically aimed at evaluating to what extent query strategies can satisfy the preferences of human users.

4.1 Reducing Manual Labeling Effort with Active Learning

Paper: Julia Romberg and Tobias Escher. Automated Topic Categorisation of Citizens' Contributions: Reducing Manual Labelling Efforts Through Active Learning. In *Electronic Government*, pages 369–385. Springer, 2022.

Personal Contribution: Julia Romberg mainly developed the concept of the paper, with the support of Tobias Escher. She also designed the experiments, prepared the dataset accordingly, and researched, implemented and evaluated the machine learning approaches. Julia Romberg and Tobias Escher jointly interpreted the results to draw conclusions for the use case and co-authored the manuscript.

Status: published

In this paper, we take a first stab at applying active learning for efficient topic classification of citizen contributions through pre-trained language models. To this end, we assess the performance of several active learning approaches on a case study of three public participation processes to cycling infrastructure in German municipalities. Both single-label and multi-label classification cases are covered in our evaluation.

The focus is on three factors of practical relevance. First, we pay attention to the classification accuracy of the methods. Second, we examine how much manual labeling effort can be saved in order to relieve analysts in their work. Third, we give consideration to the approaches' runtimes, which may limit their practical applicability.

In light of their expected performance and reasonable computing effort, we decide on *Contrastive Active Learning* (Margatina et al., 2021) and *Minimum Expected Entropy* (Holub et al., 2008) as promising query strategies. Representing pre-trained language models as a prominent example, BERT serves as the classification model. What is more, we draw a direct comparison with the methodology introduced by Purpura et al. (2008), which combines the *Query by Committee* query strategy with different traditional classification algorithms, namely support vector machines, maximum entropy, and naive Bayes. This course of action allows us to evaluate what developments research has made since then for our domain of application.

In comparison to the early study of active learning on public participation data by Purpura et al. (2008), we show a remarkable 0.08 point increase in average classification accuracy. Simultaneously, the amount of training data required declines as pre-trained language models exhibit faster learning capabilities than the traditional algorithms. Despite having a significantly longer runtime than the other models, BERT stays within a tolerable range of a few minutes per iteration.

Our results show not only that supervised machine learning models can reliably classify topic categories for public participation contributions, but that active learning significantly reduces the amount of training data required. This has important implications for the practice of public participation because it dramatically cuts the time required for evaluation. We therefore hypothesize that active learning should significantly reduce human efforts in most cases of topic classification of citizen contributions.

Automated Topic Categorisation of Citizens' Contributions: Reducing Manual Labelling Efforts through Active Learning

Julia Romberg*[0000−0003−0033−9963] and Tobias Escher[0000−0002−6607−2088]

Heinrich Heine University, Düsseldorf, Germany
{julia.romberg, tobias.escher}@hhu.de

Abstract. Political authorities in democratic countries regularly consult the public on specific issues but subsequently evaluating the contributions requires substantial human resources, often leading to inefficiencies and delays in the decision-making process. Among the solutions proposed is to support human analysts by thematically grouping the contributions through automated means. While supervised machine learning would naturally lend itself to the task of classifying citizens' proposal according to certain predefined topics, the amount of training data required is often prohibitive given the idiosyncratic nature of most public participation processes. One potential solution to minimise the amount of training data is the use of active learning. While this semi-supervised procedure has proliferated in recent years, these promising approaches have never been applied to the evaluation of participation contributions. Therefore we utilise data from online participation processes in three German cities, provide classification baselines and subsequently assess how different active learning strategies can reduce manual labelling efforts while maintaining a good model performance. Our results show not only that supervised machine learning models can reliably classify topic categories for public participation contributions, but that active learning significantly reduces the amount of training data required. This has important implications for the practice of public participation because it dramatically cuts the time required for evaluation from which in particular processes with a larger number of contributions benefit.

Keywords: Topic Classification · Public Participation · Active Learning · Natural Language Processing

1 Introduction

Democratic authorities are regularly using public participation to consult and involve citizens in order to inform political decisions and increase public support [8]. While their function and effectiveness is open to debate [19], they enjoy considerable popularity among the public that regularly contributes hundreds or even thousands of proposals to such consultations. As a consequence, policymakers and their administrations regularly face the problem of how to make sense of the diversity of statements that the public provides while at the same time maintaining the high standards of transparency and due process required for such important democratic processes. Usually this requires human analysts to read each contribution, detect duplicates, identify common themes, and categorise contributions accordingly before preparing conclusions from the input. This is a time consuming effort that often leads to inefficiencies and delays in the decision-making process [23,2,7].

While human assessment should not be abandoned, given the relevance of citizens' input to the democratic decision-making, technical solutions have long been proposed as a means to reduce the workload of human evaluators [18]. Here we focus on approaches to support analysts by using Natural Language Processing (NLP) techniques to categorise disparate contributions into groups that share certain thematic properties. As we review below, both supervised as well as unsupervised machine learning strategies have been applied to this task with mixed results. Given that categorisation of citizen contributions generally follows certain pre-defined goals such as sorting according to particular topics or administrative responsibilities, categorisation schemes are not arbitrary but constructed before the participation process. As a consequence, we assume that supervised machine learning approaches like classification are better suited to the task than completely unsupervised procedures that aim to detect latent structures in the data. However, these supervised procedures require manually labelled training data, calling into questions any efficiency gains that motivated automation in the first place. This demand would not constitute a barrier if models could be pre-trained and subsequently applied. Yet, regularly public participation processes are distinct and require tailored categorisation schemes. This idiosyncratic nature means models need to be customised for each process, requiring substantial amounts of training data.

A potential solution to minimise the amount of data is the use of active learning, a semi-supervised procedure that (to the best of our knowledge) has been applied to the evaluation of participation contributions only once [20]. While since that study almost 15 years ago, active learning strategies (and NLP in general) have advanced, these promising technologies have not been applied to the analysis of citizen participation. Therefore we systematically assess how different active learning strategies can reduce manual labelling efforts while maintaining a good model performance. To this end we study data from online participation processes in three German cities that consulted citizens on improvements for cycling. Specifically, we investigate different supervised machine learning models in order to establish what classification quality can be achieved without active learning (RQ1). We use this as a baseline to investigate how much manual labelling effort can be saved through active learning (RQ2). However, given that our focus is on enabling a practical application of these models, we also test how time-efficient the different categorisation approaches are to assess whether these could be used in realistic scenarios (RQ3).

We start by discussing previous NLP approaches to structuring contributions thematically (2) before introducing our dataset (3) and the active learning techniques applied (4). We evaluate the results of different query strategies and classifiers (5) and discuss their implications for practical application (6). Finally, the concluding section summarises the results and outlines avenues for further research (7).

2 Approaches to Thematically Structure Contributions

Organising citizens' contributions thematically is a basic step in the evaluation of public participation processes and so far two machine learning strategies have been proposed to support this task. These are unsupervised approaches, mainly topic modeling, on the one hand, and supervised classification algorithms on the other.

Unsupervised machine learning algorithms cluster similar content by discovering hidden patterns in the data. As these rely on unlabelled datasets, they require no previous manual coding which makes them attractive to use. Several such algorithms have been applied in previous work, including k-means and k-medoids clustering [23,25], non-negative matrix factorization [2], associative networks [24] and correlation explanation topic modeling [5]. By far the most popular is topic modeling with Latent Dirichlet Allocation (LDA) (see for example [15,16,2,11,10]).

Much of the work mentioned above shows that the detection of meaningful topics by unsupervised learning is subject to major limitations. To start with, for algorithms such as LDA and k-means the number of topic clusters to be identified must be specified in advance. This risks that the number of topics is somewhat arbitrary. What is more, while an approximate number of topics can be found with strategies such as experimenting with different values using human judgment or statistical measures, this requires considerable manual analysis effort [23,10]. An even more serious limitation are the topic clusters that emerge. Even with an appropriate number of topics to be found, there is still no guarantee that the algorithms will return those topics that are required by the user.

However, human evaluators of participation processes generally already have a good idea of what categories they are interested in. The reason is that such processes are initiated in order to consult the public on a specific topic such as a proposed infrastructure project or a legal text. Therefore, even before the process begins, there are a number of categories on which the analysts expect input and this pre-defined categorisation scheme can then later be refined when contributions are reviewed. As a consequence, we argue that it is more suitable to benefit from this prior knowledge in order to provide clusters of interest rather than to rely on latent structures that might not be relevant to the user. This is exactly the function of supervised machine learning which we therefore consider more appropriate to support categorising contributions thematically [14,6,1,4,13].

Given a set of labelled training data, supervised models are trained to classify citizen contributions into categories that have been previously defined by the user. Most works relied on conventional approaches such as support vector machines, but more recent works also included neural networks and transformer models like BERT. Some promising results have been obtained, but only under the condition that a sufficient amount of previously (usually manually) categorised data is available for training the models. This may be true in certain cases, such as in the use case described by Kim et al. [13] who used a categorisation by administrative unit for a city platform that is available to citizens in the long term. Once trained, the model can support officials by being used to automatically classify new requests that are constantly coming in.

However, many participation processes are singular events that have a specific objective and only run for a short period of time. Therefore, regularly analysts have to adapt the thematic categories of the evaluation to the respective process. This usually makes the transfer of trained models impossible. Rather, the classification models must be trained anew for each process with appropriate data, which requires to label (at least part of) the contributions from the process under consideration. This additional human labelling effort must not be underestimated as the previously introduced studies show that relied on training datasets consisting of several thousand data points. Yet, as is not least documented by our dataset, many of the consultation processes, e.g. in municipalities, do not even generate these large numbers of contributions. While hundreds or a few thousands of contributions pose substantial burden to administration to evaluate, fully supervised machine learning may not remedy the situation when analysts would still have to code a large share of the dataset in order to train a classifier. As a consequence, supervised machine learning might not offer an efficiency benefit for a whole range of practical applications in the area of public participation.

In order to provide a feasible solution also for processes with a lower number of contributions, Purpura et al. [20] motivated a human-in-the-loop approach. *Active learning* aims to reduce the amount of required training data by selecting a minimal subset that provides the greatest performance gain in training a classification model. The algorithm works in close collaboration with the user, who gradually categorises small parts of the dataset until the model performs satisfactorily. The authors were able to confirm that active learning can reduce manual labelling efforts while maintaining a high model performance. Nevertheless, depending on the number of categories (17 or 39), still more than 600 respectively more than 800 sentences had to be labelled manually until an accuracy of 70% was reached - a score which is comparable to the results of many of the works on supervised classification introduced above. In summary, it was thus evident that the use of active learning is promising, but the approaches still need to be improved.

Since the study of Purpura et al., the research on NLP and on active learning has evolved. Our goal is to apply state-of-the-art methods to citizen contributions and to evaluate to what extent the advanced methods can further reduce the amount of training data needed. In addition, we also assess the runtime of these models as another potential barrier for practical application.

3 The Cycling Dialogues Datasets

In this paper we focus on contributions collected from citizens in three nearly identical participation processes in the German municipalities of Bonn, Ehrenfeld (a district of Cologne) and Moers. In each city, the authorities consulted the public in order to identify planning measures that would improve the situation for cyclists. To do so, from September to October 2017 citizens were invited to propose measures for particular locations using a map-based online participation platform. Before the process, the local traffic planning authorities of the three cities that initiated these consultations developed a set of eight categories, representing different aspects for improvement such as cycle path quality or lighting. These would subsequently be used in order to process the proposals from citizens.

Initially, each contribution was assigned to a single (primary) category by the citizens when submitting the contribution. This assignment was checked by the moderators of the online platform and adjusted if necessary. After the online participation phase, an analyst went through the contributions from all three processes again and checked the categorisation. In rare cases this led to reassignment of primary categories. What is more, for those contributions whose content would qualify for more than one category, in addition to the primary category further secondary categories where assigned. The share of multi-labelled contributions regarding the eight main categories amounts to 10% in Bonn and Moers, and 15% in Ehrenfeld. Among these, only few contributions had more than two labels assigned (Bonn: 21, Ehrenfeld: 10, Moers: 3).

We use this categorisation as the basis for our study and investigate how to accurately and efficiently predict the correct label(s) for each contribution. While one could certainly insist that this body of data lacks intersubjectivity, it represents a scenario that regularly occurs in practical applications as individual analysts code large parts or even the entire contributions on their own. Nevertheless, although the categorisation is ultimately based on one individual analyst and may contain a somewhat subjective bias on his part, it is by no means arbitrary because it also incorporates the judgement of different people (citizen and moderators). We thus argue that it is certainly sufficient for most of the use cases where this categorisation is the starting point of further processing of contributions. More important for our study is that the labels reflect a consistent assignment [20] which is certainly the case as all were reviewed by a single person.

The coded dataset comprises a total of $3,139$ contributions. *Cycling Dialogue Bonn* has received the most contributions with $2,314$, whereas *Cycling Dialogue Ehrenfeld* and *Cycling Dialogue Moers* account for 366 and 459 unique contributions respectively. The contributions contain an average of 4.83 (Bonn), 4.66 (Ehrenfeld) and 4.78 (Moers) sentences. Table 1 gives insights into the thematic priorities within the eight categories. Cycling traffic management and cycle path quality attracted the most interest in all datasets, followed by either obstacles or traffic lights. The (larger) differences in the amount of contributions as well as the (smaller) difference in the distribution of categories can be attributed to both

Table 1: Overview of datasets and distribution of topic categories by single labels and multiple labels respectively.

	PRIMARY LABELS			PRIMARY & SECONDARY LABELS		
CATEGORIES	Bonn	Ehrenfeld	Moers	Bonn	Ehrenfeld	Moers
cycling traffic management	1,020 (44.1%)	195 (53.3%)	222 (48.4%)	1,056 (45.6%)	204 (55.7%)	229 (49.9%)
signage	150 (6.5%)	16 (4.4%)	19 (4.1%)	182 (7.9%)	20 (5.5%)	27 (5.9%)
obstacles	319 (13.8%)	35 (9.6%)	31 (6.8%)	364 (15.7%)	45 (12.3%)	33 (7.2%)
cycle path quality	449 (19.4%)	58 (15.8%)	111 (24.2%)	519 (22.4%)	71 (19.4%)	118 (25.7%)
traffic lights	178 (7.7%)	34 (9.3%)	47 (10.2%)	197 (8.5%)	39 (10.7%)	51 (11.1%)
lighting	37 (1.6%)	1 (0.3%)	10 (2.2%)	47 (2.0%)	2 (0.5%)	15 (3.3%)
bicycle parking	108 (4.7%)	22 (0.6%)	9 (2.0%)	112 (4.8%)	26 (7.1%)	9 (2.0%)
misc	53 (2.3%)	5 (1.4%)	10 (2.2%)	84 (3.6%)	25 (6.8%)	27 (5.9%)
total documents	2,314	366	459	2,314	366	459

contextual factors such as city size or local infrastructure, and individual-level factors such as the participating stakeholders.

A noteworthy characteristic of the datasets is that some categories are only rarely represented. For example, lighting occurs only twice in Ehrenfeld and bicycle parking occurs only 9 times in Moers. Although this is likely to make classification more difficult, such uneven distributions by topic are not at all the exception in citizen comments, making the results of the evaluation with regard to the rarely occurring classes of great interest.

In contrast to the work of Purpura et al. [20], here we categorise entire contributions rather than individual sentences within these. This is motivated by the fact that this is also the approach chosen by practitioners in the field of citizen participation (see for example [23,2]). What is more, in our dataset the contributions contain just about five sentences on average and thus are relatively short in comparison to the average length of 41.55 sentences reported by Purpura et al. [20].

4 Methodology

In the following, we introduce the concept of active learning and describe the techniques selected to be part of our study. These are various specific strategies for selecting the data points to be labelled as well as suitable classification algorithms.

We consider two types of classification problems, both of which will be addressed in the evaluation. On the one hand, we want to identify the thematic focus, i.e. the primary category, of the contributions. To do this, we solve a *single-label classification problem* in which a decision function is learned that maps each input vector to exactly one class. Second, we are interested to see to what extent all associated topics of a contribution can be recognised. In such a *multi-label classification problem*, the input vectors can be mapped to one or more classes.

4.1 Active Learning

The goal of active learning is to quickly learn a good decision function for classifying data points to save manual labelling effort. Optimally, the subset of data to be labelled should be minimal while the prediction accuracy is maximised. Being an interactive process, the human expert is sequentially consulted by the computer to (in our case) categorise samples of contributions whose labelling can be of most use in training the model.

In each iteration of the process, the k most informative data points are selected using some query strategy. Subsequently, these samples are manually labelled and added to the pool of so far labelled data points (i.e. from earlier iteration rounds). The classification model is then retrained with all labelled samples and evaluated. If the classification performance is sufficient (according to some stopping criterion), the active learning process terminates.

Specific to each active learning approach is therefore on the one hand the choice of *query strategy* and on the other hand the choice of *classifier*.

4.2 Query Strategies

Active learning attempts to find a minimal training dataset that simultaneously maximises the classification performance. Therefore, the challenge is to select those data points whose labelling provides the greatest benefit for training of the classifier in each iteration. Query strategies attempt to find an approximate solution to this problem and here we investigate four different query strategies.

Random Sampling (RS) is a query strategy that randomly selects data points from a pool of unlabelled samples. In this very basic strategy, there is no prioritisation of samples regarding their value for the training. While we can anticipate that this naive approach will not yield the best results, we are interested in seeing what improvements the more targeted strategies can achieve in comparison.

Query By Committee (QBC) [22] is a query strategy in which the disagreement between a committee of classifiers serves as a measure of information gain. To this end, the classifiers, previously trained on already labelled samples, categorise each unlabelled sample and subsequently the predictions are used to calculate a disagreement score (e.g. 0 if all predictions match). The unlabelled samples are then ranked in descending order based on their disagreement scores, and the top-k (i.e. those that the committee was least confident about) are forwarded to the human annotator.

In our experiments, we use a committee of three classifiers and define the disagreement score of a sample as the number of distinct class predictions minus one. We follow the course of action by Purpura et al. [20], but dispense with the specifications for hierarchical schemas.[1] If assignment to more than one category is allowed, we sum up the class-wise disagreement scores.

[1] We also forgo the computationally expensive additional clustering that has been suggested as an extension because of runtime considerations.

Minimum Expected Entropy (MEE) [12] is a query strategy that tries to minimise the prediction uncertainty of unlabelled data points by selecting those with the largest expected uncertainty to be labelled first. The prediction uncertainty of a data point is estimated with the entropy measure. Given a discrete random variable X, $\mathcal{H}(X)$ takes a value between 0 and 1 depending on the probability distribution over the variable's possible values (e.g. the prediction outcome of the current classification model for the different categories C):

$$\mathcal{H}(X) = -\sum_{c \in C} P(X = c) \log_2 P(X = c)$$

Contrastive Active Learning (CAL) [17] is a recent approach to improve querying by selecting so-called contrastive samples. These are samples that are close to each other in the feature space (e.g. share a similar vocabulary), but for which the current classification model's predictions are very different. Similar samples are found using the k-nearest neighbour algorithm and the difference in prediction probabilities is measured using the Kullback-Leibler divergence. The authors could show that CAL can perform equivalently or even better than a range of query strategies such as entropy for several tasks, including topic classification.

4.3 Classifier

In addition to the choice of a suitable query strategy, the choice of the classifier is crucial for the success of active learning. We therefore compare different classifiers, including both classical and state-of-the-art approaches. Following the setup from [20], we consider *support vector machines* (SVM), the *maximum entropy classifier* (MaxEnt), and the *naive Bayes classifier* (NB), some of which are known to perform well across a range of classification tasks. We also test an ensemble classifier that combines SVM, MaxEnt and NB. The textual contributions were transformed into tf-idf-weighted term vectors to obtain a machine-readable format. Non-word tokens were excluded, the words were lower-cased and lemmatised. To further reduce the dimensionality of the feature vectors, we also removed less discriminative words, i.e. words that occurred only once or in more than 80% of the contributions in the respective dataset. We furthermore include *BERT* (Bidirectional Encoder Representations from Transformers) in the comparison, one of the most popular transformer models. Within the last few years, transformer models have contributed significantly to the improvement of results in various NLP applications, and more recently they have also been considered for use in active learning [9]. In this work, we initialise BERT with the case-sensitive gbert-base model[2], a pre-trained language model for German, and encode the textual contributions accordingly.

[2] Model available at `https://huggingface.co/deepset/gbert-base`.

5 Evaluation

We address three research questions, starting by investigating how well the automated topic classification of citizen contributions already works. Keeping this knowledge of the potential and limitations of topic classification in mind, we turn our attention to the savings in manual labelling efforts through the use of active learning. Finally, we analyse the runtime of the approaches and thus consider a second key aspect for their practical applicability.

We answer the questions for public participation processes on cycling in the cities of Bonn, Ehrenfeld and Moers. This allows us to make a direct comparison between three thematically similar processes that differ, however, in the number of citizen ideas collected and the distribution along the categories. In order to obtain reliable results, especially with the small datasets, the experiments were realised with a 5-fold cross-validation of $80\% - 20\%$ splits for training and testing the classification model. The model score will be reported as the average outcome of the five runs and the standard deviation will be indicated. We measure category-wise performance with the F_1 score, the harmonic mean of model precision and recall for the respective class. For assessing model performance on a global level, we compute the proportion of correct predictions using *accuracy* for single-label classifications and *micro-averaged* F_1 for multi-label classifications. Micro-averaged F_1 is a common measure, and for single-label scenarios, it is equivalent to accuracy.

5.1 RQ1: What Classification Quality Can Be Achieved Without Active Learning?

First of all, we are interested in how well topic classification can work on our datasets in general. Table 2 shows the results for each of the five classifiers presented above, for single-label and for multi-label classification respectively. To improve the model fit on the datasets, we tuned hyperparameters in each cross-validation split (see Appendix A for more details).

The results are encouraging: the primary thematic focus of citizens' contributions could be correctly predicted in 75% to 80% of the cases, depending on the dataset. If all related topic categories were to be found, similarly good outcomes were achieved with between 72% and 80% of the predicted labels matching the human annotation. As expected, BERT can improve the accuracy respectively the micro-averaged F_1 score, in our setting by up to 0.11 compared to MaxEnt, the best performing among the other models. The effects are particularly remarkable for rarely occurring categories, such as bicycle parking in Moers, where only seven to eight matching contributions were available for training the model (the remaining contributions were part of the test set). This clearly emphasises the strengths of the pre-trained language model, which stores previously learned knowledge about semantic relationships between words. Comparing the results for the different classification tasks, i.e. single-labelling and multi-labelling, shows that most classifiers perform similarly well in both appli-

Table 2: Results of single-label and multi-label topic classification.

			Single-Label Classification								
			cycling traffic management	signage	obstacles	cycle path quality	traffic lights	lighting	bicycle parking	misc	accuracy
F_1	*Bonn*	SVM	0.75(0.02)	0.45(0.14)	0.65(0.07)	0.71(0.03)	0.73(0.04)	0.74(0.11)	0.82(0.10)	0.03(0.07)	0.71(0.02)
		MaxEnt	0.76(0.02)	0.44(0.10)	0.65(0.08)	0.72(0.02)	0.72(0.03)	0.77(0.11)	0.84(0.07)	**0.12**(0.13)	0.71(0.02)
		NB	0.68(0.02)	0.05(0.05)	0.39(0.14)	0.57(0.02)	0.30(0.06)	0.00(0.00)	0.15(0.09)	0.00(0.00)	0.56(0.02)
		Ensemble	0.76(0.01)	0.44(0.12)	0.66(0.08)	0.71(0.02)	0.73(0.03)	0.73(0.09)	0.83(0.08)	0.03(0.07)	0.71(0.02)
		BERT	**0.80**(0.03)	**0.58**(0.06)	**0.71**(0.04)	**0.75**(0.04)	**0.80**(0.03)	**0.81**(0.10)	**0.90**(0.04)	0.06(0.13)	**0.76**(0.02)
	Ehrenfeld	SVM	0.76(0.04)	0.10(0.22)	0.66(0.13)	0.34(0.18)	0.68(0.05)	0.00(0.00)*	0.62(0.19)	0.00(0.00)	0.66(0.04)
		MaxEnt	0.75(0.05)	0.20(0.18)	**0.68**(0.11)	0.40(0.21)	0.69(0.07)	0.00(0.00)*	**0.84**(0.12)	0.00(0.00)	0.67(0.03)
		NB	0.66(0.04)	0.00(0.00)	0.19(0.25)	0.06(0.08)	0.04(0.10)	0.00(0.00)*	0.00(0.00)	0.00(0.00)	0.49(0.05)
		Ensemble	0.77(0.03)	0.10(0.22)	0.65(0.14)	0.36(0.18)	0.68(0.05)	0.00(0.00)*	0.78(0.08)	0.00(0.00)	0.68(0.04)
		BERT	**0.83**(0.02)	**0.36**(0.25)	0.66(0.14)	**0.63**(0.10)	**0.73**(0.09)	0.00(0.00)*	**0.84**(0.10)	0.00(0.00)	**0.75**(0.03)
	Moers	SVM	0.78(0.05)	0.25(0.23)	0.46(0.15)	0.66(0.10)	0.74(0.24)	0.33(0.31)	0.27(0.37)	0.00(0.00)	0.70(0.05)
		MaxEnt	0.78(0.04)	0.31(0.17)	0.37(0.13)	0.67(0.09)	0.78(0.07)	0.59(0.38)	0.67(0.41)	0.00(0.00)	0.71(0.03)
		NB	0.72(0.03)	0.00(0.00)	0.00(0.00)	0.67(0.03)	0.44(0.14)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.62(0.03)
		Ensemble	0.77(0.05)	0.25(0.23)	0.40(0.21)	0.67(0.09)	0.74(0.24)	0.37(0.34)	0.13(0.30)	0.00(0.00)	0.70(0.05)
		BERT	**0.84**(0.03)	**0.52**(0.17)	**0.59**(0.09)	**0.81**(0.10)	**0.91**(0.08)	**0.70**(0.45)	**0.73**(0.43)	0.00(0.00)	**0.80**(0.03)

			Multi-Label Classification								
			cycling traffic management	signage	obstacles	cycle path quality	traffic lights	lighting	bicycle parking	misc	micro-avg F_1
F_1	*Bonn*	SVM	0.77(0.02)	0.45(0.10)	0.66(0.03)	0.70(0.01)	0.76(0.05)	0.67(0.23)	0.79(0.07)	0.18(0.14)	0.71(0.01)
		MaxEnt	0.75(0.01)	0.46(0.06)	0.64(0.01)	0.69(0.02)	0.76(0.04)	0.79(0.14)	0.80(0.09)	0.28(0.14)	0.70(0.01)
		NB	0.65(0.01)	0.15(0.05)	0.37(0.05)	0.65(0.02)	0.37(0.06)	0.04(0.09)	0.19(0.12)	0.17(0.13)	0.52(0.01)
		Ensemble	0.75(0.02)	0.45(0.10)	0.64(0.04)	0.69(0.05)	0.73(0.09)	0.59(0.25)	0.76(0.11)	0.24(0.17)	0.69(0.02)
		BERT	**0.81**(0.01)	**0.48**(0.17)	**0.71**(0.02)	**0.78**(0.03)	**0.78**(0.03)	**0.83**(0.09)	**0.89**(0.04)	**0.39**(0.07)	**0.77**(0.01)
	Ehrenfeld	SVM	0.45(0.41)	0.00(0.00)	0.39(0.17)	0.29(0.19)	0.45(0.34)	0.00(0.00)	0.54(0.32)	0.20(0.17)	0.43(0.26)
		MaxEnt	0.73(0.04)	0.25(0.25)	0.50(0.12)	0.45(0.06)	0.68(0.08)	0.18(0.25)	0.62(0.29)	0.15(0.14)	0.61(0.04)
		NB	0.77(0.05)	0.00(0.00)	0.21(0.16)	0.26(0.09)	0.17(0.12)	0.00(0.00)	0.11(0.16)	0.24(0.18)	0.49(0.02)
		Ensemble	0.74(0.02)	0.08(0.18)	0.28(0.27)	0.23(0.16)	0.55(0.19)	0.00(0.00)	0.33(0.41)	0.06(0.13)	0.56(0.07)
		BERT	**0.82**(0.03)	**0.33**(0.21)	**0.65**(0.11)	**0.57**(0.13)	**0.76**(0.07)	**0.20**(0.45)	**0.77**(0.20)	**0.24**(0.15)	**0.72**(0.02)
	Moers	SVM	0.78(0.02)	0.30(0.20)	0.25(0.16)	0.69(0.11)	0.82(0.10)	0.46(0.36)	0.33(0.47)	0.00(0.00)	0.69(0.04)
		MaxEnt	0.79(0.07)	0.23(0.13)	0.29(0.09)	0.68(0.09)	0.82(0.07)	**0.63**(0.18)	0.67(0.41)	0.00(0.00)	0.70(0.04)
		NB	0.75(0.06)	0.05(0.11)	0.08(0.11)	0.62(0.09)	0.49(0.07)	0.00(0.00)	0.00(0.00)	**0.13**(0.12)	0.58(0.03)
		Ensemble	0.78(0.04)	0.24(0.14)	0.28(0.18)	0.71(0.03)	0.81(0.09)	0.58(0.35)	0.40(0.55)	0.11(0.25)	0.70(0.04)
		BERT	**0.88**(0.05)	**0.41**(0.34)	**0.56**(0.24)	**0.82**(0.06)	**0.93**(0.03)	0.55(0.16)	**1.00**(0.00)	0.00(0.00)	**0.80**(0.04)

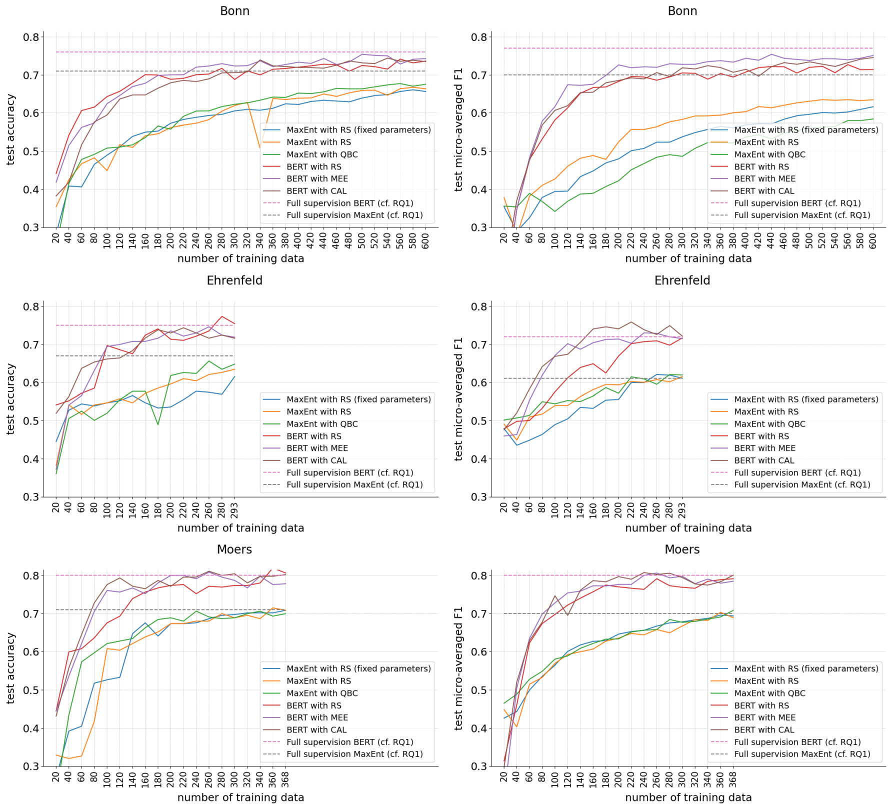
cations. This suggests that predicting all associated labels of a contribution is by no means more difficult than the recognition of the primary topic.

All models had problems with recognising contributions that were grouped in the misc category, which is not surprising due to the missing thematic coherence of the content. It should also be noted that in Ehrenfeld the category lighting occurs too infrequently to allow evaluation in the single-label case.

5.2 RQ2: How Much Manual Labelling Effort Can Be Saved Through Active Learning?

It is evident from the results for RQ1 that even smaller datasets have the potential to provide enough information to train good topic classification models. With the application of active learning, we are now taking a closer look at this potential.

In our experiments, the active learning process (implemented using the small-text library [21]) is initialised with 20 randomly drawn samples (i.e. contributions). Then, in each active learning loop, 20 unlabelled samples are retrieved with the respective query strategy and added to the pool of labelled data. We



(a) Single-Label Classification

(b) Multi-Label Classification

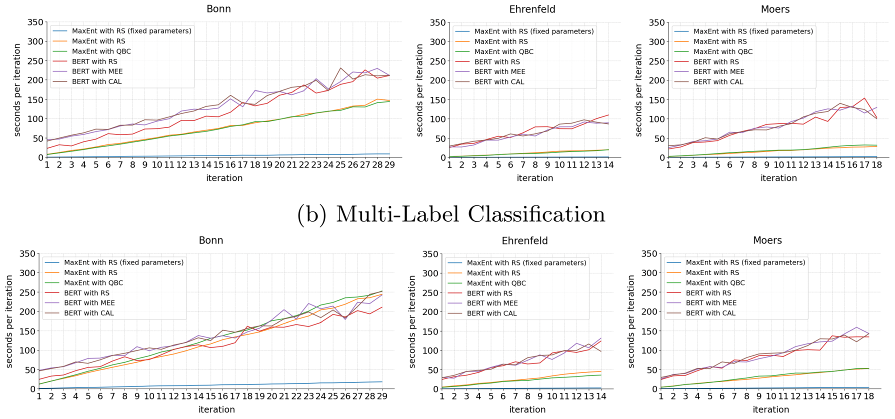
Fig. 1: Accuracy respectively micro-averaged F_1 scores for active learning per iteration.

compare the two best performing classifiers from RQ1 and first evaluate them with RS to have a baseline. QBC and MEE follow a similar strategy of selecting samples (by disagreement of a committee and uncertainty in prediction, respectively). With respect to the work of [20], we combine MaxEnt with QBC. A combination of BERT and QBC, on the other hand, was rejected because of runtime considerations since in addition to the costly transformer model, three further models would have to be trained per active learning iteration. Instead, we use the well-known MEE query strategy with BERT. Furthermore, we explore whether the recently developed query strategy CAL can further improve active learning with BERT. To keep model training time low, hyperparameter tuning for BERT is limited to selecting the best model from 10 training epochs. For MaxEnt, we compare a gridsearch-optimised model against one with fixed hyperparameters.

An overview of the results is provided in Figure 1. Since the learning curve in Bonn levelled off after a few hundred samples, we stopped the time-consuming experiment at this point and only report the results until then.

All BERT variants are superior to MaxEnt, not only because of the accuracy they can achieve but also because they learn faster. While all query strategies work well with BERT, MEE and CAL show an advantage over RS especially in multi-labelling. For single-label classification, the best strategy approximates the maximum accuracy scores from full supervision (averaging 0.77) already with 500 (Bonn), 180 (Ehrenfeld), and 120 (Moers) labelled samples. For multi-label classifications, the pool of labelled data to achieve the best micro-averaged F_1 scores (averaging 0.76) could be reduced to 440 (Bonn), 160 (Ehrenfeld), and 200 (Moers).

5.3 RQ3: How Time-Efficient Are the Different Categorisation Approaches?



(a) Single-Label Classification

(b) Multi-Label Classification

Fig. 2: Time duration of active learning iterations in seconds.

Not only the quality of the results but also the runtime is relevant if such an approach is to be developed for use by practitioners. Figure 2 reports how long the individual iterations, i.e. loops, of active learning take. This reflects the time a user has to wait between coding sessions. BERT-based experiments were run on Google Colab with Tesla P100-PCIE-16GB GPU and 2.2 GHz Intel Xeon CPU processor. The other classifiers were evaluated on a local machine with 1.8 GHz Intel Core i7-8565U CPU processor.

Encouragingly in terms of applicability, no iteration in the observation interval lasts longer than five minutes. Taking into account the findings from RQ2, to achieve these results on average a human analyst would have to wait less than three minutes (Bonn) or even less than one minute (Ehrenfeld, Moers) between the coding sessions. At the same time, however, we can observe that BERT is more computationally intensive than MaxEnt, even though we severely limited hyperparameter tuning in our experiments.

6 Discussion

Based on the evaluation summarised above we can now answer the research questions and discuss their implications.

For the first research question (RQ1), the results show that supervised machine learning can predict the correct label(s) on average for about 77% of the cases. We believe that this accuracy is already sufficient for most of the practical use cases because this categorisation is only the starting point of further manual processing of contributions. During this further processing possible misclassification would be detected and could easily be corrected. A number of issues are particularly noteworthy about this level of accuracy. First of all, the classification works equally well for single and multi-labelling. What is more, BERT as a current state-of-the-art approach offers the best results - not only because it achieves higher accuracy, but also because it works more reliably for categories with few contributions than the other classifiers evaluated. Finally, we test the models on three different datasets that vary in size and we can show that these results can be achieved also on datasets that contain only a few hundred contributions.

These results already show that automated classification through supervised models could be useful in supporting human evaluation of contributions. However, as discussed in the introduction, the main barrier to its practical application is that full supervision requires the manual labelling of large parts of the data. In our evaluation, this accuracy was achieved through coding a share of 80% of the entire dataset, an approach also pursued in several studies that focused on maximising the accuracy of approaches but neglected the drawback of manual labelling effort (e.g. [4]).

To address this shortcoming, as a second research question (RQ2) we investigated the potential of different active learning strategies to reduce manual labelling efforts. Our results show conclusive evidence that active learning can indeed obtain a similar performance while requiring only a fraction of the data to be manually coded. For the three datasets it was sufficient to manually label about 20% (Bonn), 50% (Ehrenfeld), and 30% (Moers) to achieve about the same level of accuracy as with full supervision. Naturally, these efficiency-improvements grow with the size of the dataset. Active learning significantly reduces manual labelling efforts and outperforms the previously used approaches for topic classification of participation contributions [20].

However, this would only offer a useful support for practice if these models can be realistically computed in common administrative settings. Therefore we also investigated the time-efficiency of the different categorisation approaches

(RQ3). As it turns out, all of these require only a few minutes per iteration to compute. However, it should be noted that these time benchmarks depend on specific hardware (e.g. GPU and processor). The implications for practical use will need to be investigated in future work.

To put these figures in perspective and estimate the efficiency gains, we optimistically assume that it would take a human 30 seconds to code a single contribution. Using the dataset of Bonn and the results of the single-labelling experiments, fully manual coding of the entire 2, 314 contributions would thus require 19 hours and 17 minutes of labour. In contrast, training a machine learning model with active learning requires the labelling of only 500 data points (about 22% of the corpus) to achieve a performance that would be comparable to a model with full supervision in training. This would amount to 4 hours and 10 minutes of manual coding time with machine assistance. We might add a human analyst's waiting time in between manual annotation sessions that is required in the active learning process for the computation of the next set of samples to be labelled. However, this only increases total time by 1 hour (on average about 150 seconds for the 24 iterations). What is more, this time can be used to carry out other tasks or to provide the necessary breaks in coding session to the human analyst. This means the time required to label the whole dataset with active learning amounts to 5 hours 10 minutes in contrast to more than 19 hours.

Even if we take into account that the machine learning model would produce a number of misclassifications (based on the results from RQ1 we assume this to be the case for about one in four samples, i.e. 580) which would require manual correction once each result is processed by the human analysts, with about 4 hours and 50 minutes of additional work this still amounts to a substantial reduction in time required: Instead of more than 19 hours, it would take just 10 hours (including one hour of waiting time). Relying on the same assumptions the total time required is reduced by 20% in Ehrenfeld and 50% in Moers through active learning. While the actual efficiency gains will depend on a number of factors (size of corpus, coding time per data point, computing time per iteration, amount of training data required, model accuracy), we believe that in any realistic scenario active learning will always represent a significant reduction in time required from human analysts.

In sum, our results show not only that supervised machine learning models can reliably classify topic categories for public participation contributions, but also that by utilising active learning this can be achieved with manually labelling only a comparatively small part of the data. This has important implications for the practice of public participation because once implemented, these models substantially cut the time required for manual coding.

7 Conclusion and Future Work

Public consultations are popular instruments in democratic policy-making but the subsequent evaluation of the (written) contributions requires considerable human resources. While supervised machine learning offers a way to support analysts in thematically grouping citizen ideas, often the amount of training data required is prohibitive given the idiosyncratic nature of most public participation processes. One possible solution to minimise the manual labelling effort is the use of active learning. However, the merits of this semi-supervised method for evaluating participation data have received little attention so far.

In this study, we researched the application of active learning based on online participation processes in three German cities. We first explored the capabilities of automated topic classification in general. Building on this, we investigated how much manual labelling effort can be saved through active learning and how time-efficient the different approaches are. Our results show that supervised machine learning models can reliably classify topic categories for public participation contributions. When combined with active learning, the amount of training data required can be significantly reduced while keeping algorithmic runtime low. These findings can be of great benefit to the practice of public participation, as they significantly reduce the time required for the thematic pre-sorting of submissions to participation processes.

Despite these exciting findings, some questions remain unanswered that need to be addressed in future work. So far, the coding of our dataset reflects primarily the assessment of a single analyst. Although this is a realistic application scenario, future research should attempt to evaluate predictions based on labels with (higher) intercoder reliability. It could well be that the actual model accuracy is even higher if misclassifications in the training data are avoided. Furthermore, we limited hyperparameter tuning for BERT to reduce computation time. For real-world implementation, we strongly recommend fine-tuning the BERT model to increase model accuracy if a higher runtime is acceptable. Similarly, we would like to evaluate other transformer architectures as well as further query strategies, in particular those specifically designed for deep neural network models (e.g. [3]).

Likewise, we need to address possible limitations of our approaches, such as applicability to long texts and runtime dependency on the GPU. Finally, classes with few contributions deserve a more thorough investigation, examining how effectively they can be found through the various query strategies in active learning and what impact a failure of detection has on the utility in practical application. Eventually, our long-term goal is to make these approaches available as software to make their use feasible for practitioners.[3]

Acknowledgements 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.

[3] The datasets and the code that was used to run the experiments are available at https://github.com/juliaromberg/egov-2022.

Appendix A: Hyperparameter Tuning

For SVM, we apply a gridsearch over the hyperparameters $C \in [0.1, 1, 10, 100]$, $\gamma \in [1, 0.1, 0.01, 0.001]$, and with either the RBF or the linear kernel. For MaxEnt, we search for $C \in [10, 100, 1000]$ in combination with the L1 or the L2 norm for penalty. In the Ensemble classifier, we reduce the number of hyperparameter combinations to keep the duration of the experiments within reasonable limits and thus do not consider $C \in [0.1]$ and $\gamma \in [0.01, 0.001]$ for SVM.

BERT is trained using the AdamW optimizer with a learning rate of $2e-5$ and $\epsilon = 1e-8$. Training runs for 10 epochs, from which the best model is selected using a validation set. In the non-active setup we tested batch sizes of 2, 4 and 8. We found that a batch size of 2 gave the best results (RQ1) and for this reason, we opted for this batch size in the active learning experiments (RQ2).

References

1. Aitamurto, T., Chen, K., Cherif, A., Galli, J.S., Santana, L.: Civic crowdanalytics: Making sense of crowdsourced civic input with big data tools. In: Proceedings of the 20th International Academic Mindtrek Conference. p. 86–94. AcademicMindtrek '16, Association for Computing Machinery, New York, NY, USA (2016)
2. Arana-Catania, M., Lier, F.A.V., Procter, R., Tkachenko, N., He, Y., Zubiaga, A., Liakata, M.: Citizen participation and machine learning for a better democracy. Digit. Gov.: Res. Pract. **2**(3) (jul 2021)
3. Ash, J.T., Chicheng, Z., Akshay, K., John, L., Alekh, A.: Deep Batch Active Learning by Diverse, Uncertain Gradient Lower BoundsDeep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In: International Conference on Learning Representations 2020 (ICLR 2020) (2020)
4. Balta, D., Kuhn, P., Sellami, M., Kulus, D., Lieven, C., Krcmar, H.: How to Streamline AI Application in Government? A Case Study on Citizen Participation in Germany. In: Lindgren, I., Janssen, M., Lee, H., Polini, A., Rodríguez Bolívar, M.P., Scholl, H.J., Tambouris, E. (eds.) Electronic Government. pp. 233–247. Springer International Publishing, Cham (2019)
5. Cai, G., Sun, F., Sha, Y.: Interactive Visualization for Topic Model Curation. In: Proceedings of the ACM IUI 2018 Workshop on Exploratory Search and Interactive Data Analytics (2018)
6. Cardie, C., Farina, C., Aijaz, A., Rawding, M., Purpura, S.: A Study in Rule-Specific Issue Categorization for e-Rulemaking. In: Proceedings of the 9th International Conference on Digital Government Research. pp. 244–253 (2008)
7. Chen, K., Aitamurto, T.: Barriers for Crowd's Impact in Crowdsourced Policymaking: Civic Data Overload and Filter Hierarchy. International Public Management Journal **22**(1), 99–126 (jan 2019)
8. Dryzek, J.S., Bächtiger, A., Chambers, S., Cohen, J., Druckman, J.N., Felicetti, A., Fishkin, J.S., Farrell, D.M., Fung, A., Gutmann, A., Landemore, H., Mansbridge, J., Marien, S., Neblo, M.A., Niemeyer, S., Setälä, M., Slothuus, R., Suiter, J., Thompson, D., Warren, M.E.: The crisis of democracy and the science of deliberation. Science **363**(6432), 1144–1146 (2019)
9. Ein-Dor, L., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y., Slonim, N.: Active Learning for BERT: An Empirical

Study. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7949–7962. Association for Computational Linguistics, Online (Nov 2020)

10. Hagen, L.: Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? Information Processing & Management **54**(6), 1292–1307 (2018)
11. Hagen, L., Uzuner, Ö., Kotfila, C., Harrison, T.M., Lamanna, D.: Understanding Citizens' Direct Policy Suggestions to the Federal Government: A Natural Language Processing and Topic Modeling Approach. In: 2015 48th Hawaii International Conference on System Sciences. pp. 2134–2143 (2015)
12. Holub, A., Perona, P., Burl, M.C.: Entropy-Based Active Learning for Object Recognition. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–8 (2008)
13. Kim, B., Yoo, M., Park, K.C., Lee, K.R., Kim, J.H.: A value of civic voices for smart city: A big data analysis of civic queries posed by seoul citizens. Cities **108**, 102941 (2021)
14. Kwon, N., Shulman, S.W., Hovy, E.: Multidimensional Text Analysis for eRulemaking. In: Proceedings of the 2006 International Conference on Digital Government Research. p. 157–166. dg.o '06, Digital Government Society of North America (2006)
15. Levy, K.E.C., Franklin, M.: Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry. Social Science Computer Review **32**(2), 182–194 (2014)
16. Ma, B., Zhang, N., Liu, G., Li, L., Yuan, H.: Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. Information Processing & Management **52**(3), 430–445 (2016)
17. Margatina, K., Vernikos, G., Barrault, L., Aletras, N.: Active learning by acquiring contrastive examples. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 650–663. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021)
18. OECD: Promise and Problems of E-Democracy. OECD (jan 2003)
19. Parry, G., Moyser, G.: More Participation, More Democracy? In: Beetham, D. (ed.) Defining and Measuring Democracy. Sage, London (1994)
20. Purpura, S., Cardie, C., Simons, J.: Active Learning for e-Rulemaking: Public Comment Categorization. In: Proceedings of the 9th International Conference on Digital Government Research. pp. 234–243 (2008)
21. Schröder, C., Müller, L., Niekler, A., Potthast, M.: Small-text: Active learning for text classification in python (2021)
22. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 287–294 (1992)
23. Simonofski, A., Fink, J., Burnay, C.: Supporting policy-making with social media and e-participation platforms data: A policy analytics framework. Government Information Quarterly **38**(3), 101590 (2021)
24. Teufl, P., Payer, U., Parycek, P.: Automated Analysis of e-Participation Data by Utilizing Associative Networks, Spreading Activation and Unsupervised Learning. In: Macintosh, A., Tambouris, E. (eds.) Electronic Participation. pp. 139–150. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)
25. Yang, H., Callan, J.: OntoCop: Constructing Ontologies for Public Comments. IEEE Intelligent Systems **24**(5), 70–75 (2009)

4.2 User-Centric Measures for Active Learning

Paper: Julia Romberg. Mind the User! Measures to More Accurately Evaluate the Practical Value of Active Learning Strategies. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 996–1006. INCOMA Ltd., 2023.

Contributions: Julia Romberg solely designed and implemented this research, and prepared the manuscript.

Status: published

Conducting user studies of active learning is complex and expensive, and therefore not affordable for mainstream experiments. For this reason, active learning is usually simulated on existing annotated datasets for the development and comparison of methods. However, this has resulted in a disregard of numerous aspects that hold significant practical importance (Margatina and Aletras, 2023). These include out-of-distribution generalization (Longpre et al., 2022), the dependency of dataset and model (Settles, 2009; Tomanek and Morik, 2011), and further typical characteristics of practical datasets, such as imbalanced classes (Ein-Dor et al., 2020) or so-called extreme multi-label scenarios, where a large number of different classes can be assigned (Wertz et al., 2022). Lowell et al. (2019) illustrate the general difficulty of making a-priori forecasts about the practical value of strategies based on experiments.

One important but often overlooked factor is the human annotator, which forms the second pillar of active learning alongside the machine learner. In computational linguistic research, the influence of the users and their requirements on active learning solutions has received little attention so far (Baldridge and Palmer, 2009; Hachey et al., 2005; Donmez and Carbonell, 2008; Tomanek and Hahn, 2010). But these particulars can influence the entire process due to annotation errors and inconsistencies, as well as annotator-specific behaviors and expectations, among other things (Settles, 2011; Calma and Sick, 2017; Margatina and Aletras, 2023).

In this paper, we consider user expectations for active learning and develop four measures regarding class-related requirements in data selection, motivated by our transdisciplinary collaboration with participation practice (Romberg and Escher, 2020). These target how well classes are covered over the course of active learning, how present minority classes are in the acquired batches, and how variant these batches are with respect to available classes. In a comparison of various query strategies coupled with BERT across six datasets for imbalanced multi-class classification, the proposed measures provide important insights that complement existing evaluation approaches. For example, we find that the strongest query strategy in terms of macro F_1 performance is not the one that excels in class coverage. Our measures offer a promising starting point for refining existing strategies to better fulfill practical requirements in text classification scenarios.

Mind the User! Measures to More Accurately Evaluate the Practical Value of Active Learning Strategies

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

Abstract

One solution to limited annotation budgets is *active learning* (AL), a collaborative process of human and machine to strategically select a small but informative set of examples. While current measures optimize AL from a pure machine learning perspective, we argue that for a successful transfer into practice, additional criteria must target the second pillar of AL, the human annotator. In *text classification*, e.g., where practitioners regularly encounter datasets with an increased number of imbalanced classes, measures like macro F_1 fall short when finding all classes or identifying rare cases is required. We therefore introduce four measures that reflect class-related demands that users place on data acquisition. In a comprehensive comparison of uncertainty-based, diversity-based, and hybrid query strategies on six different datasets, we find that strong macro F_1 performance is not necessarily associated with full class coverage. Uncertainty sampling outperforms diversity sampling in selecting minority classes and covering classes more efficiently, while diversity sampling excels in selecting less monotonous batches. Our empirical findings emphasize that a holistic view is essential when evaluating AL approaches to ensure their usefulness in practice – the actual, but often overlooked, goal of development. To this end, standard measures for assessing the performance of text classification need to be complemented by such that more appropriately reflect user needs.

1 Introduction

A well-known problem in supervised machine learning (ML) is scenarios where there are limited resources (e.g., budget or time) to annotate data. One approach to solving this problem is *active learning* (AL; Cohn et al. 1996), a collaborative process between human and machine. Through targeted query strategies, AL aims to find a minimal subset of examples whose labels provide the most information for fitting a model.

In *text classification*, many applications have been found to benefit from AL, such as sentiment analysis, intent or topic detection (e.g., Li et al., 2012; Zhang and Zhang, 2019; Tong and Koller, 2001). In addition to these task-specific studies, increased efforts have been made to systematically evaluate the performance of AL strategies across different use cases (e.g., Settles, 2011; Siddhant and Lipton, 2018; Ein-Dor et al., 2020).

Yet many academic studies ignore crucial real-world factors, leading to flawed assessments of practical utility. Literature has pointed out several limitations, including: the difficulty of making a-priori forecasts about the practical value of strategies (Lowell et al., 2019); the fact that actively acquired datasets are often only effective coupled with the respective model (Lowell et al., 2019; Tomanek and Morik, 2011); the need for out-of-distribution generalization (Longpre et al., 2022); taking into account class imbalance that is regularly encountered in real-world text classification (Ein-Dor et al., 2020); and the consideration of extreme multi-label scenarios (Wertz et al., 2022).

While these works seek to optimize AL from a ML perspective, it has been largely neglected that users themselves can present significant challenges that may impact the success of AL. For instance, it has been found that the effectiveness of AL depends on the expertise of the annotators (Baldridge and Palmer, 2009). Furthermore, examples selected by acquisition functions tend to be more ambiguous in terms of class assignment, leading to an increase in annotation uncertainty (Settles, 2011) and annotation time (Hachey et al., 2005). Such details can affect and even challenge the entire AL process.

We therefore argue that a successful transition from research to practice requires a more holistic evaluation that targets both pillars of AL, the machine learner and the human annotator. In this work, we focus primarily on the requirements that the human annotator places on a successful AL process. More precisely, we introduce evaluation measures that already take this perspective into account during the development phase of AL approaches, further referred to as "user-centric"[1].

Considering the frequent scenario of multi-class text classification with imbalanced classes (Ein-Dor et al., 2020; Wertz et al., 2022), we contribute through four novel measures that capture class-related demands in AL. We compare different query strategies coupled with BERT across six datasets and analyze the results from both a standard ML and a more user-centric perspective. Our findings indicate that the proposed measures can provide important insights into strengths and weaknesses of AL that complement existing approaches.

2 Related Work

In evaluating the performance of AL, predictive accuracy has generally been the main focus (Kottke et al., 2017). Prior work has relied on task-specific measures, such as accuracy and macro F_1. Less commonly, AL-specific measures like deficiency (Yanık and Sezgin, 2015) were used. In addition, several measures have addressed desirable characteristics of query strategies, such as uncertainty of the acquired examples (Yuan et al., 2020; Wang et al., 2022), diversity of the acquired examples (Zhdanov, 2019; Yuan et al., 2020), and representativeness w.r.t the full dataset (Zhu et al., 2008; Ein-Dor et al., 2020). The majority of these measures focus on the input or feature space, but representativeness has also been measured in the output label space (Prabhu et al., 2019; Chaudhary et al., 2021). Another focus besides predictive accuracy has been on the computational effort (Schröder et al., 2022).

With a strong emphasis on ML performance, the current measures tend to overlook the human component in the real-world application of AL. Although user studies have proven helpful in uncovering user-centric pitfalls that can get in the way of practicality (Settles, 2011; Peshterliev et al., 2019), they are expensive and time-consuming, which is why they are often avoided in research. To overcome this hurdle, Calma and Sick (2017) suggested to simulate user factors from real-world applications when evaluating AL in an experimental setup (i.e., benchmarking on an already labeled dataset). They addressed error-proneness in AL and presented a theoretical framework for simulating annotation uncertainty of the user.

Our work follows this lead by incorporating user factors into the laboratory evaluation of AL to provide a simple alternative to costly user studies. However, we focus on the requirements that users place on AL applications in order for them to be considered beneficial in practice. In particular, we address the need for achieving high or full class coverage in a timely manner and covering minority classes. Furthermore, as a solution approach to the annotation uncertainty problem modeled by Calma and Sick (2017), we hypothesize how examples should be acquired to reduce annotation errors and introduce a corresponding measure.

3 Methodology

In this section, we first give a more formal introduction to AL. Then, we motivate and define the four user-centric measures that are central to this work.

3.1 Active Learning

We make use of the pool-based AL scenario (Lewis and Gale, 1994), which assumes that there is a large pool of unlabeled data $\mathcal{U}$ and a small set of labeled data $\mathcal{L}$ at the beginning. We decided to acquire examples in mini-batches, as a practical method.

AL proceeds according to the following scheme: Using some query strategy, a batch $\mathcal{B}$ of examples is selected (and consequently removed) from $\mathcal{U}$. These examples are then labeled by an oracle (e.g., a human annotator) and added to $\mathcal{L}$. Finally, a model is fit to $\mathcal{L}$. This process is repeated until a predefined stop criterion (e.g., a given annotation budget) is met. In the initial run, a default set of labeled examples is used to start the AL process.

3.2 Measures from User-Centric Perspective

In the following, we introduce four measures that reflect demands users may place on AL in practice. The definitions refer to single-label classification.

We draw motivation for the measures from two sources. On the one hand, we refer to the scientific literature, as specified below. On the other hand, we relate directly to the needs of practical users that have been communicated to us in our transdis-

[1] In the following, we will use the terms human annotator and user interchangeably. This terminology is adopted because in certain application scenarios, the human role goes beyond simply annotating data, as AL can simultaneously serve as an analytical tool, e.g., for computational social science.

ciplinary work over several years (among others documented in Romberg and Escher, 2020).

Minority-aware Batch Distribution When "dealing with imbalanced datasets in practice, the rare classes are often the ones that are particularly interesting." as Wertz et al. (2022) state. This is especially true for real-world use cases where AL is used not only for effective dataset creation, but also for efficient dataset analysis (Bonikowski et al., 2022; Yang et al., 2022). In the topic classification of citizens' contributions, e.g., human evaluators are often aware of the common issues in advance (Romberg and Escher, 2022). Thus, from the user's point of view, preference should be given to unexpected classes, which usually corresponds to minority classes. We measure this demand by

$$M(\mathcal{B}) = \frac{1}{n_{\mathcal{B}}} \sum_{c \in C} (1 - \frac{n_{\mathcal{U}_c}}{n_{\mathcal{U}}}) \cdot n_{\mathcal{B}_c} \quad (1)$$

where $n_{\mathcal{B}}$ is the batch size, $n_{\mathcal{U}}$ is the number of examples in $\mathcal{U}$, $n_{\mathcal{U}_c}$ is the number of examples in $\mathcal{U}$ that belong to class c, and $n_{\mathcal{B}_c}$ denotes the number of examples in $\mathcal{B}$ that belong to class c. To give more emphasis to rare classes, we weight all classes by their counter probability of occurring in the initial pool of unlabeled data. $M(\mathcal{B}) \in [0, 1]$, and a higher value indicates more awareness.

Class Coverage It is also of interest to consider how many classes AL can find (Schröder et al., 2021; Wertz et al., 2022). Achieving a high or even full class coverage is desirable for several reasons.

Knowing how query strategies handle the set of classes can be critical to building trust in human-machine collaboration. Indeed, a concern of our practice partners was missing some classes. If there was any potential for incomplete class coverage, this could even be a reason to completely avoid using machine text classification in their use case.

Such needs can relate to task requirements to which the human analyst is also subject. Thus, in these situations, it is not enough to, e.g., simply educate users about the strengths and weaknesses of ML algorithms; ML must meet these requirements.

What is more, with respect to the previously described utilization of AL for data analysis, a timely overview of the collection is an often desired feature, which is given by a fast class coverage.

And overall, having as complete a representation as possible of the classes relevant to the task at hand is generally an important prerequisite for creating reliable datasets.

We measure the class coverage of the examples in $\mathcal{L}$ as

$$K(\mathcal{L}) = \frac{|C_{\mathcal{L}}|}{|C|} \quad (2)$$

where $C_{\mathcal{L}}$ is the set of classes included in $\mathcal{L}$, and C is the total set of classes in the collection.

As a further indicator, we define the full class coverage I_K of an AL experiment as the number of iterations it takes to cover all classes in C.

Variation-aware Batch Distribution The performance of human annotators can be affected by various factors, including declining concentration or fatigue (Calma et al., 2016). One reason for the (more rapid) onset of these factors can be batches that offer little alternation in terms of the classes to be annotated. To reduce error-proneness in annotation caused by monotonous batches, we propose batches to fulfill two conditions: they should represent the available classes (measured by the ratio of acquired to the total number of classes available), and the acquired examples should be uniformly distributed among classes to offer variety (measured via entropy):

$$V(\mathcal{B}) = \frac{|C_{\mathcal{B}}|}{|C_{\mathcal{B}} \cup C_{\mathcal{U}}|} \cdot \sum_{c \in C_{\mathcal{B}}} - \left(\frac{\frac{n_{\mathcal{B}c}}{n_{\mathcal{B}}} \cdot \log_2(\frac{n_{\mathcal{B}c}}{n_{\mathcal{B}}})}{\log_2(|C_{\mathcal{B}}|)} \right) \quad (3)$$

where $C_{\mathcal{B}}$ is the set of classes included in the batch and $C_{\mathcal{U}}$ is the set of classes in the unlabeled pool. $V(\mathcal{B}) \in [0, 1]$, with larger values indicating a more varied set of examples with reference to the classes.

4 Evaluation Design

We provide an overview of the study design next by going into detail about the dataset selection, the chosen classification model, the selection of query strategies, and the experimental setup.

4.1 Datasets

We aim at a broad comparison across different datasets to empirically demonstrate the strengths and weaknesses of different query strategies with respect to the introduced user-centric measures. In doing so, we consider six datasets for different multi-class tasks and from diverse domains. An overview is given in Table 1.

DBPedia (Zhang et al., 2015) is a large-scale ontology dataset of Wikipedia articles (title and

Dataset	Task	Domain	$\|C\|$	Train	Val	Test
DBPedia	T	Wikipedia	14	15,000	2,000	4,000
20NG	T	News	20	2,507	354	721
ATIS	I	Flight reservations	17	3,802	537	1,093
TREC-50	Q	Diverse	46	4,163	589	1,196
BILLS	T	Congressional bills	20	15,000	2,000	4,000
CDB	T	Public participation	29	1,372	194	395

Table 1: Details of the six datasets. The task types are topic (T), intent (I), and question (Q) classification. $|C|$ denotes the number of classes.

abstract) and their topics. 20 Newsgroups[2] (20NG) contains messages collected from diverse newsgroups. Airline Travel Information Systems (ATIS; Siddhant and Lipton, 2018) is a dataset of transcribed audio recordings for classifying the intent of costumer utterances. TREC (Li and Roth, 2002) provides answer types for a collection of English-language questions.

These four English-language datasets regularly serve for benchmarking AL. While previous work has mostly relied on TREC-6, which organizes the questions into six main categories, we use the finer answer types of TREC-50 to give more weight to the multi-class setting that motivates this work.

The remaining two datasets come from real-world applications of topic classification in the computational social sciences. The Congressional Bills Corpus (BILLS; Purpura et al., 2008) provides information on bills introduced in the U.S. Congress between 1947 and 2008. One of its purposes is to examine what attention the congress has paid to various issues by thematically analyzing the bill's titles. The Cycling Dialogues Bonn (CDB; Romberg and Escher, 2022) is a German dataset of citizen contributions to a public participation process on cycling infrastructure.

While ATIS, TREC-50, BILLS, and CDB reflect the common class imbalance of real-world data, DBPedia and 20NG have been artificially counterbalanced at creation. To simulate a plausible scenario, we adjust the distribution of the two datasets through sub-sampling. Since we lack knowledge about the original data sources' actual distributions, we assume a distribution according to Zipf's law: the most frequent class should occur about twice as often as the second most frequent class, three times as often as the third most frequent class, and so on.

We follow Ein-Dor et al. (2020) by limiting the size of large datasets to $21K$ (DBPedia and BILLS) and apply a $70\%/10\%/20\%$ split for training, validation and testing. There were predefined splits available for some of the datasets (train/test splits for TREC-50 and 20NG; a train/val/test split for DBPedia), which we rejected for the following reasons: For TREC these are neither consistent in their distribution (Lowell et al., 2019), nor does the test split for TREC-50 contain all of the original 47 classes. For 20NG and DBPedia, we modified the structure of the datasets to a greater extent by adapting them to Zipf's distribution. We therefore decided to define new splits selected according to a stratified random sample. Classes with less than 5 examples were removed.

Table 3 in Appendix A provides detailed insights into the resulting dataset splits. The splits and code for the experiments are available at https://github.com/juliaromberg/ranlp-2023.

4.2 Classification Model

Several studies have shown the potential of AL coupled with pre-trained language models (PTMs) (e.g., Ein-Dor et al. 2020; Yuan et al. 2020; Longpre et al. 2022; Zhang et al. 2022). We adhere to these findings and apply the BERT base model (Devlin et al., 2019), as has been done in much of the related work. For English datasets, we use uncased BERT[3] (pre-trained on English data), and for the German dataset, we rely on cased GBERT[4].

4.3 Query Strategies

We compare a variety of strategies that have stood out in previous work for their strong results and cost-effectiveness when used with PTMs in imbalanced settings. As a baseline, we use *Random Sampling* (Random).

Traditional uncertainty-based acquisition functions select examples according to the confidence of model prediction. They are efficient and have proven to keep up with more advanced AL strategies when used with PTMs (Zhang and Zhang, 2019; Margatina et al., 2021, 2022). We consider *Least Confidence* (LC; Lewis and Gale, 1994), which has proven effective for imbalanced datasets (Ein-Dor et al., 2020; Schröder et al., 2022), and *Breaking Ties* (BT; Luo et al., 2005), which was recommended as a baseline for uncertainty sampling with transformers by Schröder et al. (2022). LC selects those examples for annotation where the model's probability output is lowest for the most

[2] http://qwone.com/ jason/20Newsgroups/

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

[4] https://huggingface.co/deepset/gbert-base

likely class, i.e., cases in which the model is least confident. BT aims to improve classification confidence by selecting examples where the difference in probability outputs between the two most likely classes is the smallest.

Diversity-based query strategies aim to select examples that best represent the full dataset. We include *Core-Sets* (Sener and Savarese, 2018), which have been found to select batches of high diversity and representativeness in addition to a promising boost of model performance in imbalanced settings (Ein-Dor et al., 2020). Core-sets are subsets of examples that represent the dataset in a learned feature space (for PTMs: CLS) in the sense that a model trained on a Core-set is competitive to a model trained on the entire dataset. We rely on the lightweight and fast algorithm for building the Core-sets by Bachem et al. (2018).

As a proxy for functions with a hybrid objective, we choose *Contrastive Active Learning* (CAL; Margatina et al., 2021) which has the potential to outperform alternatives such as BADGE (Ash et al., 2020) and ALPS (Yuan et al., 2020) in terms of computational efficiency and accuracy (Margatina et al., 2021). CAL combines the characteristics of uncertainty- and diversity-based strategies by seeking so-called contrastive examples. These are examples that, despite high similarity in the feature space (i.e., among the k nearest neighbors), exhibit maximum mean Kullback-Leibler divergence between their predictive likelihoods.

4.4 Experimental Setup

In each AL iteration, training runs for 30 epochs on a batch size of 12 and the best model, in terms of validation loss, is retained. To avoid overfitting to the data from previous iterations, BERT is fine-tuned from scratch at each iteration (Hu et al., 2019). We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $2e-5$, beta coefficients of 0.9 and 0.999, and an epsilon of $1e-8$, and set the maximum sequence length to 100 for all datasets.

For each of the six datasets, the unlabeled pool $\mathcal{U}$ is formed by the respective training splits and 50 examples are randomly sampled from the pool to build the set of initially labeled data $\mathcal{L}$. Then, 20 iterations of AL are performed, in each of which a new batch of 50 unlabeled examples is selected from $\mathcal{U}$ according to the respective query strategy. The model performance is evaluated at the end of each iteration using a hold-out test set.

We run the AL simulation five times with different sets of initially labeled data for each combination (datasets × query strategies). To allow for a fair comparison, these seeds remain the same for each dataset across the different query strategies.

In accordance with our experimental setup, $3,156$ experiments (6 datasets $\times$ (5 query strategies $\times$ 5 initial seeds $\times$ (1 initial model $+$ 20 iterations) $+$ 1 full supervision model)) were conducted. The experiments were run on a single Nvidia Tesla P100-PCIE-16GB GPU and with 2.2 GHz Intel Xeon CPU processor.

We refer the reader to Appendix B for further details on hyperparameter selection, reproducibility of the experiments and computational costs.

5 Results

In this section, we report the experimental results. We start by shedding light on the performance of the different query strategies as is common in the literature via a standard measure for classification tasks, in our case the macro averaged F_1 score. Using the newly introduced user-centric measures, we then shift our focus to analyzing additional indicators that can help select an appropriate query strategy for practical use.

5.1 Macro F_1 Performance

Figure 1 illustrates how the macro F_1 score evolves on average over the iterations of AL in the experiments. It can be seen that full supervision performance can be achieved on all datasets within the chosen annotation budget of 20 iterations, except for BILLS.

Our analysis across all datasets shows a clear pattern of superior performance for uncertainty-based sampling compared to the other strategies. In particular, BT performs consistently strong. While hybrid CAL is in the middle of the rankings, it is evident that the diversity-based strategy mostly underperforms.

Based on these findings, from a ML-perspective that is commonly shared among many studies in the field, it seems an obvious conclusion to recommend BT as the strategy for practical application in imbalanced multi-class settings. In the following, we will examine whether this assumption can be supported from a user-centric perspective.

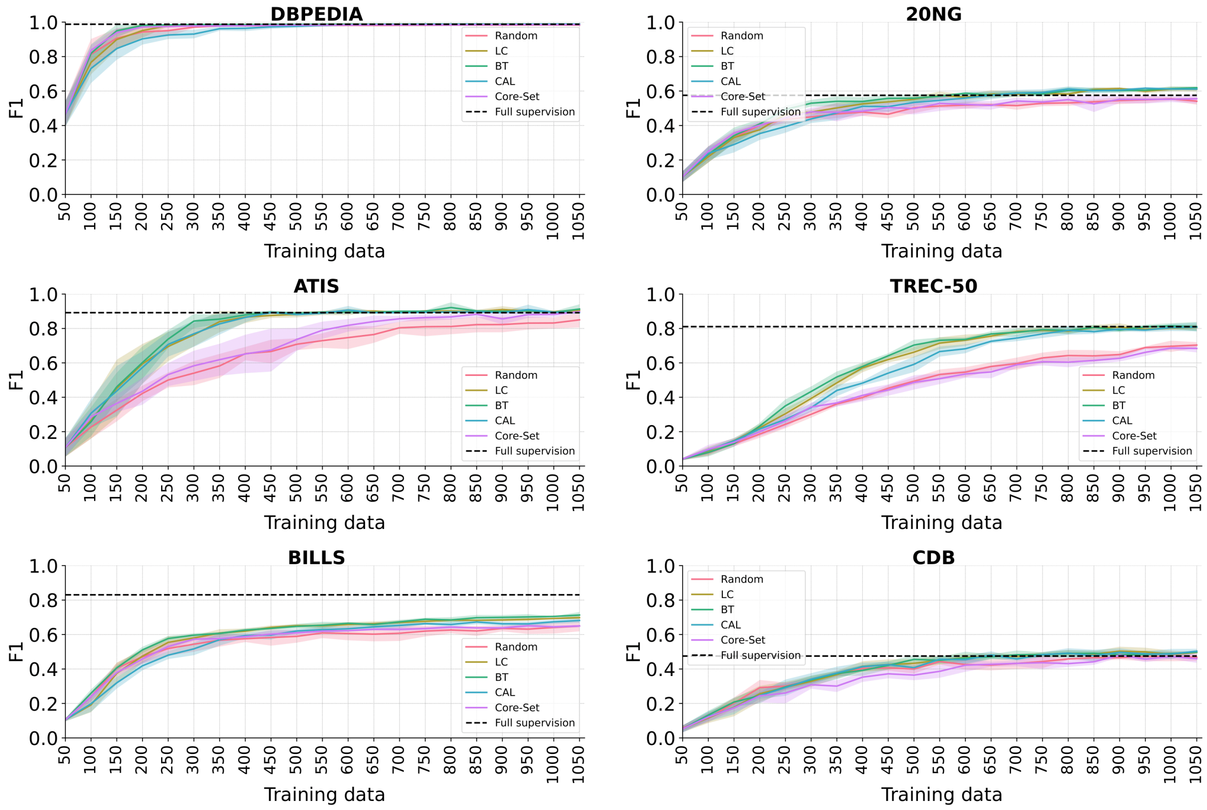


Figure 1: Macro F_1 scores, averaged over the five seeds and with the shaded area illustrating the standard deviation. As a reference for the maximum achievable macro F_1 score for each dataset, the performance of the BERT models trained on the complete training data is indicated (full supervision).

5.2 User-Centric Measures

Table 2 lists the results of the four user-centric measures for the datasets and query strategies, averaged over the iterations of AL for a better overview.

Which strategies favor minority classes? First, we evaluate whether, among the strategies considered, there are such that promote a higher representation of rare classes in the batches. We apply the minority-aware batch distribution measure $M(\mathcal{B})$ for this purpose.

All advanced strategies are found to consider rare classes more than random sampling. In particular, uncertainty-based strategies promote a higher minority representation on average. A detailed look shows that this trend is consistent among datasets, but there are major differences in how pivotal the choice of query strategy is. For BILLS and CDB, this makes a negligible difference. In contrast, the effect is much more dramatic on ATIS, where the scores range from 0.44 to 0.84.

Which strategies favor class coverage? Next, we examine whether there are any query strategies that prioritize quick and extensive class coverage by applying the class coverage measure $K(\mathcal{L})$.

The results show that uncertainty-based and hybrid query strategies stand out positively. BT achieves the highest average class coverage and turns out to be a good choice for a rapid growth in the coverage curve (as a detailed look at progress between iterations confirms).

Are the strategies capable of finding all classes? As argued in Section 3.2, a realistic requirement of the practice may be that all classes that a dataset comprises are found in the AL process. We measure the full coverage with I_K.

Contrary to our expectation, three strategies failed to find all classes within the budget of 20 annotation cycles on the datasets ATIS and TREC-50. In addition to random sampling and Core-Sets, in TREC-50 this surprisingly also affects the previously excelling strategy BT. The failure is systematic in each case, as we can observe it for several random seeds.

To gain better insight into the extent of the failure, we ran additional experiments beyond the AL budget of 20 iterations until full class coverage was achieved for the affected cases. On TREC-50,

	Random	LC	BT	CAL	Core-Set
$M(\mathcal{B})$					
DBPedia	0.852 ± 0.003	$\mathbf{0.918} \pm 0.001$	0.916 ± 0.002	0.916 ± 0.005	0.870 ± 0.002
20NG	0.874 ± 0.002	$\mathbf{0.930} \pm 0.001$	0.928 ± 0.003	0.924 ± 0.001	0.888 ± 0.001
ATIS	0.440 ± 0.006	$\mathbf{0.840} \pm 0.012$	$\mathbf{0.840} \pm 0.007$	0.735 ± 0.009	0.586 ± 0.010
TREC-50	0.925 ± 0.002	$\mathbf{0.947} \pm 0.001$	0.945 ± 0.001	$\mathbf{0.947} \pm 0.001$	0.928 ± 0.001
BILLS	0.918 ± 0.001	$\mathbf{0.931} \pm 0.001$	$\mathbf{0.931} \pm 0.001$	0.928 ± 0.000	0.924 ± 0.001
CDB	0.933 ± 0.001	$\mathbf{0.937} \pm 0.001$	0.936 ± 0.001	0.934 ± 0.000	0.933 ± 0.001
AVG	0.824 ± 0.003	$\mathbf{0.917} \pm 0.003$	0.916 ± 0.002	0.897 ± 0.003	0.855 ± 0.003
$K(\mathcal{L})$					
DBPedia	0.995 ± 0.023	0.995 ± 0.024	0.995 ± 0.023	0.995 ± 0.026	$\mathbf{0.996} \pm 0.022$
20NG	0.971 ± 0.076	0.979 ± 0.071	$\mathbf{0.982} \pm 0.067$	0.977 ± 0.072	0.977 ± 0.072
ATIS	0.864 ± 0.143	0.915 ± 0.162	$\mathbf{0.926} \pm 0.149$	0.924 ± 0.157	0.867 ± 0.137
TREC-50	0.847 ± 0.138	0.869 ± 0.159	$\mathbf{0.889} \pm 0.151$	0.881 ± 0.161	0.822 ± 0.136
BILLS	0.979 ± 0.051	0.981 ± 0.051	$\mathbf{0.984} \pm 0.048$	0.978 ± 0.056	0.983 ± 0.049
CDB	0.958 ± 0.085	$\mathbf{0.968} \pm 0.077$	0.962 ± 0.080	0.964 ± 0.082	0.962 ± 0.083
AVG	0.936 ± 0.086	0.951 ± 0.091	$\mathbf{0.956} \pm 0.086$	0.953 ± 0.092	0.934 ± 0.083
I_K					
DBPedia	1.0 ± 1.2	1.2 ± 1.3	1.0 ± 1.2	1.0 ± 1.0	$\mathbf{0.8} \pm 0.8$
20NG	4.2 ± 0.8	2.6 ± 0.9	$\mathbf{2.0} \pm 1.2$	2.6 ± 0.9	2.8 ± 1.3
ATIS	$26.6 \pm 16.4^*$	8.0 ± 2.4	8.8 ± 2.1	$\mathbf{7.6} \pm 1.3$	$22.8 \pm 6.8^*$
TREC-50	$35.2 \pm 8.1^*$	16.2 ± 2.9	$28.0 \pm 23.8^*$	$\mathbf{15.8} \pm 2.7$	$27.8 \pm 5.9^*$
BILLS	4.4 ± 0.9	3.2 ± 0.5	$\mathbf{3.0} \pm 1.2$	3.8 ± 1.1	3.4 ± 2.5
CDB	7.6 ± 2.4	5.8 ± 1.6	6.6 ± 1.1	$\mathbf{5.0} \pm 0.0$	7.0 ± 2.6
AVG	13.2 ± 5.0	6.2 ± 1.6	8.2 ± 5.1	$\mathbf{6.0} \pm 1.2$	10.8 ± 3.3
$V(\mathcal{B})$					
DBPedia	0.736 ± 0.017	0.516 ± 0.037	0.600 ± 0.018	0.474 ± 0.060	$\mathbf{0.785} \pm 0.007$
20NG	0.636 ± 0.018	0.761 ± 0.008	$\mathbf{0.791} \pm 0.009$	0.737 ± 0.030	0.688 ± 0.014
ATIS	0.216 ± 0.009	0.381 ± 0.020	0.391 ± 0.026	$\mathbf{0.458} \pm 0.007$	0.376 ± 0.010
TREC-50	0.388 ± 0.011	0.393 ± 0.013	$\mathbf{0.426} \pm 0.012$	0.388 ± 0.014	0.400 ± 0.007
BILLS	0.696 ± 0.009	0.676 ± 0.009	0.738 ± 0.019	0.637 ± 0.015	$\mathbf{0.742} \pm 0.016$
CDB	0.606 ± 0.009	0.605 ± 0.016	$\mathbf{0.617} \pm 0.013$	0.581 ± 0.009	0.607 ± 0.006
AVG	0.493 ± 0.012	0.478 ± 0.020	0.512 ± 0.016	0.477 ± 0.021	$\mathbf{0.539 \pm 0.008}$

Table 2: Average results for $M(\mathcal{B})$, $K(\mathcal{L})$, I_K, and $V(\mathcal{B})$ on the six datasets of evaluation. The scores are averaged over the seeds and iterations of AL, and standard deviation is stated. The best scores are marked in bold. Cases in which a strategy failed to reach full coverage within the given budget are marked with an asterix.

Core-Sets and BT both required up to 28 iterations on average. However, the deviations between the different seeds are much more extreme with BT: In the worst case, BT asked for manual labeling of over three quarters of the pool $\mathcal{U}$, which sums up to 60 iterations of AL.

We further discovered that in case of incomplete class coverage, it was the minority classes that were not found. This is why we repeated the experiments for TREC-50 and ATIS with an increased required minimum class support of 20 to spot check how performance changes. As for Random and Core-Sets, this modification allowed all experiments to achieve full class coverage within the given annotation budget. However, for BT, the undesired effects persisted on TREC-50. Moreover, failure even extended to the other two strategies associated with uncertainty, namely LC and CAL.

Overall, in the average comparison between all strategies, the hybrid CAL stands out, requiring on average only 6 iterations to successfully detect all classes.

How variant are the batches in terms of classes? Last, we apply $V(\mathcal{B})$ in order to account for variance in batches with the goal of reducing monotonous patterns.

Here, it is the diversity-based query strategy Core-Sets that on average produces batches that best fulfill the condition. Individually, though, the results are very mixed for the different acquisition functions and datasets. For example, BT performs best on three of the datasets, rendering this query strategy a strong contender.

6 Discussion

We considered several measures that take into account aspects that may determine the practicality of active learning strategies with respect to specific application scenarios. For the datasets under consideration, it can be seen that the macro F_1 score, the rapidity of class coverage, and the minority-awareness in the batches advocate for the use of uncertainty-based acquisition functions, in particular BT, in practical scenarios with multiple and imbalanced classes. However, Core-Sets offer the opportunity to add more variety to the monotonous task of annotation by filling batches with rather different classes and in a more balanced way. This may potentially help prevent annotation fatigue and thus human annotation errors that negatively impact AL. In addition, such variation could be a plus in terms of usability.

What is more, we found weaknesses in reaching full class coverage for all strategies. For random sampling and Core-Sets, we hypothesize that this is caused by extremely rare classes. However, for uncertainty sampling, the problem became even more apparent when excluding those classes. This is of particular interest since full supervision macro F_1 can be well achieved within the annotation budget (see Figure 1).

Although the macro F_1 score and some user-centric measures recommend BT as a favorite, the lack of reliability in achieving full class coverage, which we have empirically determined, may become a decisive criterion for practical applicability. Not only can it have a significant impact on human trust in AL. This finding affects AL in general, as the reliability of models strongly depends on the quality of the datasets.

7 Conclusion

With our results, we were able to illustrate that different query strategies stand out in different aspects that might be desirable or even necessary from the user's perspective in the practical application of AL. So what implications can be drawn for AL research beyond this study? The main reason why research on AL exists is its development and improvement for real-world use. In this, AL is a collaberative interaction between human and machine. However, this particular feature of AL seems to have gradually faded from the community's awareness, with the main focus being on optimizing the established performance measure for the particular machine learning task, e.g. classification. It is true that these established measures have important informational value about the methods. But there are additional requirements that arise specifically from the human factor inherent in the nature of AL, which likewise impact the practical value of AL. These should therefore be taken into account.

Therefore, we argue that future studies on AL should report a wider range of measures in their experimental evaluation. With this broader foundation, practitioners will be able to make a more informed decision when selecting an AL strategy based on academic findings in order to comply with their specific needs for a given application. For example, in applications where the annotation step is simultaneously used to analyze the dataset at hand, features such as a quick overview of all classes or, in particular, minority classes can be desired, as we have discussed in more detail in Section 3.2. Surely, the measures we have suggested are by no means exhaustive. Therefore, this work should also serve as a motivation to cover other aspects of the human component of AL in future research.

Ultimately, selecting an appropriate AL strategy for some practical use case is a matter of balancing different needs. The suggested measures make an important contribution to this, as they enable more reflective decisions, especially in combination with common performance measures like the macro F_1 score.

To sum up, AL has the potential to support ML in scenarios where the annotation budget is limited. We have argued that in order to assist the transfer of such methods from research to practice, both the machine learner and the human annotator must be taken into account. Considering the frequent use case of multi-class text classification with imbalanced classes, we introduced four measures that evaluate the acquired examples w.r.t. class-related requirements from the user's point of view. These measures are based on scientific literature and practical experience. Our results show that as complete a picture as possible should be considered to avoid failures in practical application.

The next step will be to conduct a user study to validate the usefulness of the metrics presented here. In future work, we will also investigate in more detail which influencing factors prevent a fast finding of all classes. This necessitates a study that investigates, among other aspects, the effect of data distribution on the class coverage of the different strategies in order to draw general conclusions.

Acknowledgments

Thanks to Tobias Escher, Gerhard Klassen and Sergej Korlakov for valuable discussions and feedback on earlier forms of this work. 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

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

Olivier Bachem, Mario Lucic, and Andreas Krause. 2018. Scalable k-means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 1119–1127.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305.

Bart Bonikowski, Yuchen Luo, and Oscar Stuhler. 2022. Politics as usual? Measuring populism, nationalism, and authoritarianism in U.S. presidential campaigns (1952–2020) with neural language models. *Sociological Methods & Research*, 51(4):1721–1787.

Adrian Calma, Jan Marco Leimeister, Paul Lukowicz, Sarah Oeste-Reiß, Tobias Reitmaier, Albrecht Schmidt, Bernhard Sick, Gerd Stumme, and Katharina Anna Zweig. 2016. From active learning to dedicated collaborative interactive learning. In *29th International Conference on Architecture of Computing Systems*, pages 1–8.

Adrian Calma and Bernhard Sick. 2017. Simulation of annotators for active learning: Uncertain oracles. In *Proceedings of the Workshop and Tutorial on Interactive Adaptive Learning*, pages 49–58.

Aditi Chaudhary, Antonios Anastasopoulos, Zaid Sheikh, and Graham Neubig. 2021. Reducing confusion in active learning for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 9:1–16.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145.

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.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for BERT: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7949–7962.

Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 144–151.

Peiyun Hu, Zachary C Lipton, Anima Anandkumar, and Deva Ramanan. 2019. Active learning with partial feedback. In *International Conference on Learning Representations*.

Daniel Kottke, Adrian Calma, Denis Huseljic, G. M. Krempl, and Bernhard Sick. 2017. Challenges of reliable, realistic and comparable active learning evaluation. In *Proceedings of the Workshop and Tutorial on Interactive Adaptive Learning*, pages 2–14.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12.

Shoushan Li, Shengfeng Ju, Guodong Zhou, and Xiaojun Li. 2012. Active learning for imbalanced sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 139–148.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 556–562.

Shayne Longpre, Julia Reisler, Edward Greg Huang, Yi Lu, Andrew Frank, Nikhil Ramesh, and Chris DuBois. 2022. Active learning over multiple domains in natural language tasks. *arXiv preprint*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 21–30.

Tong Luo, Kurt Kramer, Dmitry B. Goldgof, Lawrence O. Hall, Scott Samson, Andrew Remsen, Thomas Hopkins, and David Cohn. 2005. Active learning to recognize multiple types of plankton. *Journal of Machine Learning Research*, 6(4):589–613.

Katerina Margatina, Loic Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663.

Stanislav Peshterliev, John Kearney, Abhyuday Jagannatha, Imre Kiss, and Spyros Matsoukas. 2019. Active learning for new domains in natural language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 90–96.

Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling bias in deep active classification: An empirical study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4058–4068.

Stephen Purpura, John Wilkerson, and Dustin Hillard. 2008. The U.S. policy agenda legislation corpus volume 1 – a language resource from 1947 - 1998. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 403–409.

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, Heinrich Heine University Düsseldorf.

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.

Christopher Schröder, Kim Bürgl, Yves Annanias, Andreas Niekler, Lydia Müller, Daniel Wiegreffe, Christian Bender, Christoph Mengs, Gerik Scheuermann, and Gerhard Heyer. 2021. Supporting land reuse of former open pit mining sites using text classification and active learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4141–4152.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909.

Katrin Tomanek and Katherina Morik. 2011. Inspecting sample reusability for active learning. In *Active Learning and Experimental Design workshop in conjunction with AISTATS 2010*, pages 169–181.

Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2:45–66.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696.

Lukas Wertz, Katsiaryna Mirylenka, Jonas Kuhn, and Jasmina Bogojeska. 2022. Investigating active learning sampling strategies for extreme multi label text classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4597–4605.

Eugene Yang, Sean MacAvaney, David D. Lewis, and Ophir Frieder. 2022. Goldilocks: Just-right tuning of BERT for technology-assisted review. In *Proceedings of the European Conference on Information Retrieval*, page 502–517.

Erelcan Yanık and Tevfik Metin Sezgin. 2015. Active learning for sketch recognition. *Computers & Graphics*, 52:93–105.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7948.

Leihan Zhang and Le Zhang. 2019. An ensemble deep active learning method for intent classification. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, pages 107–111.

Shujian Zhang, Chengyue Gong, Xingchao Liu, Pengcheng He, Weizhu Chen, and Mingyuan Zhou.

2022. ALLSH: Active learning guided by local sensitivity and hardness. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1328–1342.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*.

Fedor Zhdanov. 2019. Diverse mini-batch active learning. *arXiv preprint*.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K. Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 1137–1144.

Appendix

A Datasets

Table 3 provides a detailed overview of the six datasets after preprocessing (described in Section 4.1 in the main body of this paper) with respect to the classes and the different splits (train/val/test).

Class	Train	Val	Test	Share
DBPedia				
Building	4614	615	1231	30.76%
Village	2307	308	615	15.38%
Album	1538	205	410	10.25%
NaturalPlace	1153	154	307	7.69%
MeanOfTransportation	923	123	246	6.15%
Animal	769	102	205	5.12%
WrittenWork	659	88	176	4.40%
EducationalInstitution	577	77	153	3.84%
Film	512	68	137	3.41%
Artist	461	61	123	3.07%
Company	419	56	112	2.80%
Athlete	384	52	102	2.56%
Plant	355	47	95	2.37%
OfficeHolder	329	44	88	2.20%
20 NEWSGROUPS (20NG)				
rec.sport.hockey	699	99	201	27.89%
soc.religion.christian	349	49	101	13.93%
rec.motorcycles	233	33	67	9.30%
rec.sport.baseball	174	25	50	6.95%
sci.crypt	139	20	40	5.56%
rec.autos	116	16	34	4.63%
sci.med	99	14	29	3.96%
comp.windows.x	87	12	25	3.46%
sci.space	77	11	22	3.07%
comp.os.ms-windows.misc	69	10	20	2.76%
sci.electronics	63	9	18	2.51%
comp.sys.ibm.pc.hardware	58	8	17	2.32%
misc.forsale	53	7	16	2.12%
comp.graphics	50	7	14	1.98%
comp.sys.mac.hardware	46	7	13	1.84%
talk.politics.mideast	44	6	12	1.73%
talk.politics.guns	41	6	11	1.62%
alt.atheism	39	5	11	1.54%
talk.politics.misc	37	5	10	1.45%
talk.religion.misc	34	5	10	1.37%
ATIS				
flight	2814	397	809	74.01%
airfare	316	44	91	8.30%
ground_service	185	26	53	4.86%
airline	121	17	35	3.18%
abbreviation	89	13	25	2.34%
aircraft	60	9	17	1.58%
flight_time	36	5	11	0.96%
quantity	36	5	11	0.96%
capacity	26	4	7	0.68%
distance	21	3	6	0.55%
airport	21	3	6	0.55%
flight#airfare	19	3	5	0.50%
ground_fare	17	2	5	0.44%
city	16	2	5	0.42%
flight_no	14	2	4	0.37%
meal	8	1	2	0.20%
restriction	3	1	1	0.09%
TREC-50				
ind	712	101	204	17.10%
other	565	80	162	13.57%
def	381	54	109	9.15%
count	260	37	75	6.25%
desc	232	33	66	5.56%
manner	194	28	56	4.67%
date	185	26	54	4.46%
cremat	145	20	42	3.48%
reason	138	19	40	3.31%
gr	136	19	40	3.28%
country	111	15	32	2.66%
city	103	14	30	2.47%
animal	90	12	26	2.15%
food	75	10	22	1.80%
dismed	73	10	22	1.77%
termeq	70	10	20	1.68%
period	58	8	17	1.40%
exp	55	8	15	1.31%
money	52	7	15	1.24%
state	51	7	15	1.23%
sport	44	6	13	1.06%
event	41	6	11	0.98%
substance	39	6	11	0.94%
dist	35	5	10	0.84%
color	35	5	10	0.84%
product	32	5	9	0.77%
techmeth	27	4	8	0.66%
veh	22	3	6	0.52%
perc	21	3	6	0.50%
title	18	3	5	0.44%
word	18	3	5	0.44%
mount	17	2	5	0.40%
plant	13	2	3	0.30%
lang	13	2	3	0.30%
body	13	2	3	0.30%
abb	12	2	3	0.29%
speed	10	2	3	0.25%
weight	10	2	3	0.25%
temp	9	1	3	0.22%
volsize	9	1	3	0.22%
symbol	8	1	2	0.18%
instru	8	1	2	0.18%
currency	7	1	2	0.17%
letter	6	1	2	0.15%
code	6	1	2	0.15%
ord	4	1	1	0.10%
Congressional Bills Corpus (BILLS)				
Health	2157	288	575	14.38%
Domestic Commerce	1765	235	472	11.77%
Government Operations	1739	231	464	11.59%
Defense	1288	172	343	8.59%
Public Lands	1167	155	311	7.78%
Law and Crime	1144	153	305	7.63%
Education	927	123	248	6.18%
Macroeconomics	540	72	144	3.60%
Energy	540	72	144	3.60%
Environment	526	70	141	3.51%
International Affairs	517	69	137	3.44%
Transportation	457	61	121	3.04%
Labor	455	61	121	3.03%
Immigration	344	46	91	2.29%
Social Welfare	339	45	90	2.26%
Civil Rights	266	36	71	1.78%
Technology	265	36	71	1.77%
Agriculture	240	32	64	1.60%
Housing	201	27	54	1.34%
Foreign Trade	123	16	33	0.82%
Cycling Dialogues Bonn (CDB)				
new_cycle_path	222	32	64	16.22%
unevenness_flaws_or_cracks	115	16	34	8.41%
cycle_path_permanently_parked	113	16	33	8.26%
unclear_traffic_routing	108	15	31	7.85%
too_narrow_width	98	14	28	7.14%
unfavourable_switching	75	10	22	5.46%
safe_road_crossing_is_missing	66	9	19	4.79%

Class	Train	Val	Test	Share
no_or_too_few_parking_facilities	56	8	16	4.08%
obstruction_due_to_stationary_objects	55	7	16	3.98%
marking_of_cycle_path_missing_poorly_vis..	50	7	14	3.62%
ruleadverse_behaviour	45	6	13	3.26%
set_up_road_with_priority_for_cycling	41	6	12	3.01%
signage_of_cycle_path_missing_poorly_vis...	37	5	11	2.70%
new_traffic_light_addition	35	5	10	2.55%
lack_of_visibility	32	5	9	2.35%
cycle_path_often_blocked	28	4	8	2.04%
passages_with_excessive_height_differences	24	4	7	1.78%
cycle_path_use_in_both_directions	24	4	7	1.78%
lighting_is_missing	22	3	6	1.58%
open_oneway_street_for_cycling	22	3	7	1.63%
check_mandatory_use_of_cycle_path	20	3	5	1.43%
repeatedly_dirt_or_water_on_cycle_path	18	3	5	1.33%
missing_local_reference	15	2	4	1.07%
unsuitable_parking_facilities	12	2	3	0.87%
other_notices	10	1	3	0.71%
access_to_cycle_path_only_with_detour	10	1	3	0.71%
speed_limit	7	1	2	0.51%
remove_traffic_light	6	1	1	0.41%
deficiency_report	6	1	2	0.46%

Table 3: Detailed dataset statistics in absolute and percentage terms.

B Implementation Details

Hyperparameters The choice of batch size, number of training epochs, and maximum sequence length is a tradeoff between model performance, runtime, and GPU restrictions. We empirically determined that setting the batch size to 12 yielded good results. As for the number of 30 training epochs, we found that model prediction benefits from this increased number especially when there are only a few labeled examples, but also as the AL process progresses. Future work may consider whether the number of epochs can be curtailed as $\mathcal{L}$ grows larger. In consideration with the runtime due to the chosen number of epochs and the total number of experiments, as well as with regard to GPU constraints, we decided on an overall maximum sequence length of 100. For TREC-50 and ATIS, the longest encountered sequence comprises only 41 respectively 52 tokens, so we set the maximum sequence length correspondingly lower in these cases.

Reproducibility Experiments were performed with the same five random seeds, randomly selected from the range $[1, 9999]$, to make them reproducible.

Computational Costs Table 4 provides the average duration of each AL experiment. The decisive factor for the runtime is model fine-tuning.

Full Supervision Models These (c.f. Figure 1 in the main body) were fit on the full training data of the respective dataset with AdamW, $lr = 2e - 5$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e - 8$. We trained for five epochs in case of large datasets (DBPedia, BILLS) and for 30 epochs in case of small datasets (20NG, ATIS, TREC-50, CDB), and selected the best model by validation loss. To obtain reliable results, we repeated each experiment five times with different random seeds.

	Random	LC	BT	CAL	Core-Set
DBPEDIA	613	672	670	682	675
20NG	466	474	475	475	473
ATIS	422	442	435	447	436
TREC-50	387	422	405	412	411
BILLS	611	712	710	678	665
CDB	545	561	536	560	547

Table 4: Average runtime (seconds) including model training, inference, batch acquisition, and hold-out test set prediction.

5

Argumentation Mining

Argumentation mining was first characterized by Palau and Moens (2009). The authors locate the research area at the intersection of natural language processing, argumentation theory, and information retrieval with the goal of automatically recognizing arguments in text documents and classifying their function in the local (i.e., interactions between different parts within an argument) and global structure of argumentation (i.e., interactions between arguments). Broadly, argumentation mining can be systematized into four sub-tasks (Peldszus and Stede, 2013). These include segmenting text into argumentative discourse units, classifying these segments based on their function in the argumentation, identifying relations between argumentative discourse units, and filling in missing parts of the argument if they are not explicitly stated.

Since its beginnings, argumentation mining has evolved greatly. Researchers have looked at a variety of domains, such as legal texts (Moens et al., 2007; Mochales and Moens, 2011), persuasive essays (Stab and Gurevych, 2014, 2017), broadcast debates (Budzynska et al., 2014), news articles (Ein-Dor et al., 2020), and microblogging services (Schaefer and Stede, 2021), as well as across different domains (Stab et al., 2018). Applications in the realm of argumentation mining have included analyzing support types of user propositions in online comments (Park and Cardie, 2014), predicting the usefulness of product reviews based on their argumentative content (Passon et al., 2018), and determining rhetorical moves in scientific writing (Alliheedi et al., 2019). Research furthermore focused on retrieving arguments in favor of or against a controversial topic (Levy et al., 2014; Rinott et al., 2015; Stab et al., 2018; Ein-Dor et al., 2020). A particular interest was also on the generation of argumentative text, like the synthesis of conclusions (Alshomary et al., 2020), premises (Rajendran et al., 2016) and counter-arguments (Hua and Wang, 2018; Alshomary and Wachsmuth, 2023). For a more detailed overview of research in argumentation mining, we refer the reader to the surveys of Stede and Schneider (2018) and Lawrence and Reed (2019).

As remarkable progress has been made in argumentation mining - not least demonstrated by IBM's Project Debater (Slonim et al., 2021) - the focus of the research area has broadened. More and more, the quality of argumentation has become the focus of attention because of its potential impact on any application task. According to Stede and Schneider (2018), the question of argument quality might be the one that is ultimately decisive for argumentation mining.

There are different observations about what makes a good argument, such as certain logical, rhetorical, and dialectical features (Blair, 2012). Computational approaches to assessing logical quality include the evaluability of arguments (Park and Cardie, 2018) and their cogency (Saveleva et al., 2021). As for rhetorical quality, research has looked, for example, at the convincingness (Habernal and Gurevych, 2016) and effectiveness of arguments (Zhang et al., 2016). Studies of dialectical quality cover persuasion effects (El Baff et al., 2020) and the myside bias, which describes the tendency to ignore opposing arguments (Stab and Gurevych, 2016). Taking a more pragmatic approach, it has also been proposed to compare arguments based on a measure of overall quality (Toledo et al., 2019; Gretz et al., 2020). A comprehensive taxonomy of argument quality dimensions and additional related work can be found in Wachsmuth et al. (2017a).

When it comes to public participation, the domain of our interest, the body of literature has focused on the first three sub-tasks of argumentation mining, i.e., discourse segmentation, segment classification, and relation identification (Kwon et al., 2007, 2006; Park and Cardie, 2014; Park et al., 2015; Konat et al., 2016; Liebeck et al., 2016; Fierro et al., 2017; Lawrence et al., 2017; Niculae et al., 2017; Morio and Fujita, 2018b; Eidelman and Grom, 2019).

One purpose of argumentation mining is to facilitate decision-making processes in situations where the sheer amount of data makes manual evaluation challenging or even impractical. This is exactly in line with our second research goal of analyzing the public's reasoning about the issues under discussion in order to inform policy-making. In this chapter, we therefore look at two relevant aspects of argumentation mining in support of public participation processes. In Section 5.1, we address basic structures of argumentation, while focusing on the robustness of models in terms of their utility for new datasets. In Section 5.2, we then go a step further and address the quality of citizen arguments by introducing concreteness as an overall measure. To be fair to the subjective nature of the task, we furthermore present a first approach in the field of argumentation mining that integrates multiple perspectives into the input representation of machine learning.

5.1 Robust Argument Component Identification and Classification

---

Paper: Julia Romberg and Stefan Conrad. 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. Association for Computational Linguistics, 2021.

Personal Contribution: The research was conducted entirely by Julia Romberg. The manuscript was written by Julia Romberg under the supervision of Stefan Conrad.

Status: published

In many of the works on the domain of public participation, noteworthy contributions have been made to the progress of argumentation mining. However, the important aspect of model robustness has so far lacked attention for our use case. Cocarascu et al. (2020) were the only researchers to conduct a cross-dataset evaluation, with a specific focus on relation prediction. More generally, machine learning models have been evaluated on the same dataset on which they are trained. This renders the estimate of how the performance of models differs on datasets that were not present during training difficult. To best support the evaluation of public participation through argumentation mining, we therefore conclude that one goal should be the development of models that can perform on a variety of processes without significant loss of predictive accuracy.

The main objective of this paper is thus to develop a generic model that can successfully detect argument structures across different datasets of public participation. We tackle two fundamental tasks in argumentation mining: identifying argumentative discourse units and classifying them based on their role within the argument. In doing so, we employ a scheme that distinguishes between major positions (proposed courses of action and policy options) and premises (attacking or supporting reasons), derived from Liebeck et al. (2016). We select promising approaches to computational argumentation for public participation data from the prior work outlined above. Using the CIMT Argument Components sub-corpus that includes a variety of mobility-related urban planning processes, we compare them with current pre-trained language models.

Our results indicate that BERT surpasses previous argumentation mining approaches on German public participation data in both tasks, reaching an average macro F_1 score of 0.77 for the identification of argumentative discourse units and an average macro F_1 score of 0.90 for their classification. In a cross-dataset evaluation, we show that BERT models that were trained on a particular dataset can recognize argument structures in other public participation processes (which were not part of the training) with comparable goodness of fit. By empirically demonstrating such model robustness across different datasets, we take one further step towards the practical applicability of argumentation mining in the public sector.

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

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*
	arg	0.93 (0.00)	0.94 (0.00)	**0.95** (0.00)
CD_B	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)
	arg	0.93 (0.01)	0.87 (0.02)*	**0.95** (0.01)
CD_C	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)
	arg	0.90 (0.01)	0.92 (0.01)	**0.94** (0.01)
CD_M	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)
	arg	0.86 (0.01)	0.89 (0.01)	**0.91** (0.01)
MC_K	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)
	arg	0.94 (0.02)	0.85 (0.02)*	**0.96** (0.01)
CQ_B	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)
	arg	0.91 (0.01)	0.79 (0.03)*	**0.92** (0.01)
THF	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.

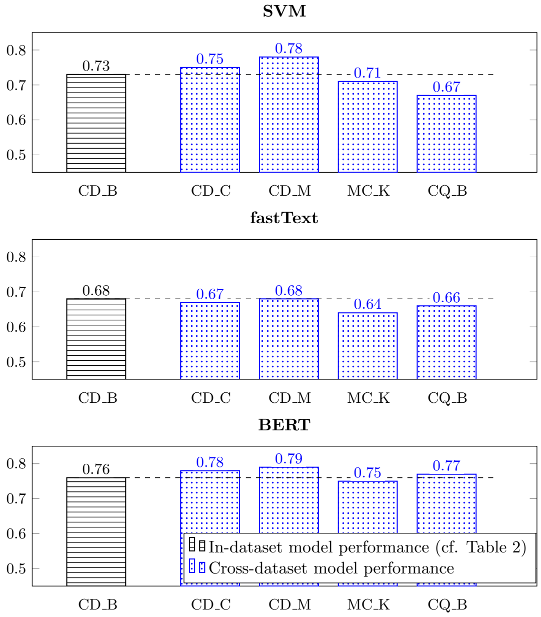


Figure 1: Cross-dataset evaluation for Task A. Results are averaged macro F$_1$ values of the five models trained on CD_B.

(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-

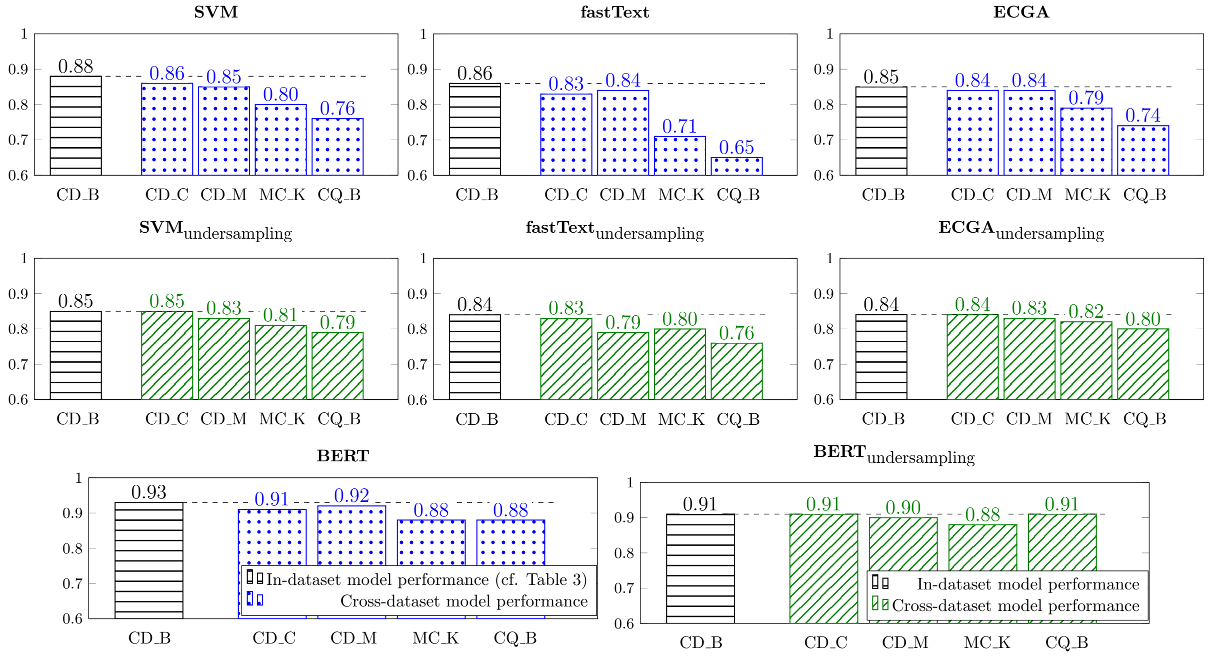


Figure 2: Cross-dataset evaluation for Task B. Results are averaged macro F_1 values of the five models trained on CD_B. (Note that in-dataset performance of CD_B with undersampling has not been reported in Table 3),

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.

5.2 Assessing Argument Concreteness

Paper: Julia Romberg. Is Your Perspective Also My Perspective? Enriching Prediction with Subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125. International Conference on Computational Linguistics, 2022.

Personal Contribution: Julia Romberg solely designed and implemented this research, and prepared the manuscript.

Status: published

To account for the quality of citizens' arguments, Park et al. (2015) focused on evaluability. Specifically, the authors introduced a model of argumentation that classifies propositions according to whether these are backed by a fact (objective evidence), a value (preference, interpretation, or judgment), a policy (course of action), a statement (personal state or fact of experience), or a reference (source of objective evidence).

In this paper, we propose an alternative to measuring the quality of arguments in public participation. Considering the level of concreteness with which propositions are uttered as a global indicator of quality, we define an overall measure that can help human analysts accelerate evaluations by prioritizing detailed input and deferring imprecise ideas until capacity is available.

In doing so, we acknowledge that evaluating the quality of an argument, i.e., how good it is, is subjective to a certain extent. The reason for such conception of reasoning that goes beyond mere rationality is rooted in a number of social psychological phenomena. Among other things, people perceive argumentation differently based on their values (Kiesel et al., 2022; Esau, 2018), their morality (Alshomary et al., 2022), or how they react to particular mechanisms of argumentation such as storytelling (Falk and Lapesa, 2022a, 2023b). While the underlying causes have been studied, no approach was taken to directly model the diversity of perspectives in the knowledge representation step of argumentation mining, as called for by Cabitza et al. (2023).

Motivated by this background, we introduce the first "multi-perspectivist" approach to the field of computational argumentation[1]: *PerspectifyMe* adds subjectivity information to conventional text classification workflows that as a rule learn from a single aggregated ground truth. It translates a given task into two sub-tasks, one of which refers to the original task of predicting aggregated labels, and one of which refers to an artificial task for predicting the subjectivity of the input using a subjectivity score.

We train a variety of models on the CIMT Argument Concreteness sub-corpus, relying on a two-level and a four-level classification of subjectivity. Our results demonstrate that machine learning of argument concreteness is feasible, with a best accuracy of 0.79 and a macro F_1 score of 0.67. Regarding the subjective perception of argument concreteness, the models correctly distinguish objective from subjective examples in three out of four cases. Despite the need for further improvement, our findings hold relevance for practitioners and the onset of multi-perspectivism in argumentation mining.

[1] We use the term "multi-perspectivist" for referring to a machine learning strategy that integrates different valid perspectives during the knowledge representation step (i.e., in setting the ground truth).

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 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 policymaking 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.)

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.

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.

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	**0.52** ± 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	**0.67** ± 0.05	**0.79** ± 0.02	**0.42** ± 0.04	**0.52** ± 0.03	**0.72** ± 0.02	**0.74** ± 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	**0.62** ± 0.07	**0.76** ± 0.04	**0.37** ± 0.06	**0.47** ± 0.05	**0.68** ± 0.06	**0.71** ± 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	**0.56** ± 0.04	**0.73** ± 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)	**0.68** ± 0.07	0.81 ± 0.03	0.38 ± 0.07	0.53 ± 0.07	0.71 ± 0.04	0.74 ± 0.04
	BERT	**0.68** ± 0.06	**0.82** ± 0.03	**0.42** ± 0.05	**0.56** ± 0.04	**0.73** ± 0.04	**0.76** ± 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 subtasks. 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.

6 Conclusion

Within the scope of this dissertation, we investigated how the evaluation of public participation processes can be supported through text classification algorithms. In this concluding chapter, we summarize the different aspects we have researched as part of the main research question. We discuss the contributions to each substantive focus and review our findings in an overarching light. In doing so, we highlight aspects where the presented work could be further improved and outline promising directions of future research.

6.1 State of Research and Data Foundation

We started with a systematic literature review about machine learning approaches to support the evaluation of textual contributions from public participation processes (R5). Taking into account the interdisciplinary nature of the field, which includes research from computational linguistics as well as from digital government, we have provided an overview of its current state by reviewing the strengths and weaknesses of existing approaches to offer guidance for further research. In our endeavor, we sharply focused by only accepting studies that referred to top-down consultations, directly initiated by public authorities to seek public input, and by limiting the selection to those studies that specifically examined the step of evaluation. While some promising approaches exist, such as for grouping data thematically and analysing arguments and opinions, there are still important hurdles to overcome before these can provide reliable support in practice. Major challenges that remain include the need to improve the quality of results, the suitability of methods for non-English language datasets, and the provision of ready-to-use software for practitioners.

Complementary to the studies our focused literature review found, future research can certainly learn from selected studies outside the search scope. For example, self-initiated citizen engagement through so-called bottom-up participation, conducted via petitions or online discussions in social media, among others, and its evaluation has attracted some interest (Belkahla Driss et al., 2019; Simonofski et al., 2021). Despite conceptual and methodological differences (e.g., induced by particular characteristics of social media content), the consideration of research on such additional forms of participation could potentially contribute to the development of methods for evaluating top-down consultations. In addition, it may be worthwhile to look at studies that center on the deliberative processes that are often used to generate the public input. There are a number of similarities here with the downstream evaluation of a collection of previously gathered contributions, especially with regard to the analysis of argumentation (Falk and Lapesa, 2022b, 2023a). Moreover, research on machine learning for discussion and discourse data in general can provide helpful hints on methods to explore (Lei and Huang, 2022; Dutta et al., 2022; Hessel and Lee, 2019).

We then presented the publicly available CIMT PartEval Corpus (R6), which was motivated by the lack of German-language resources for developing text classification models in our application domain. Comprising several thousand citizen contributions from six mobility-related planning processes with different goals and formats, the corpus forms the basis for four evaluation tasks. For these tasks previously either no annotated data was available at all or existing datasets were subject to limitations that we considered essential: In contrast to previous work on the recognition of argumentative text units and their classification for German, we covered a multitude of public participation processes, which allows researchers to evaluate how robustly machine learning models generalize to new datasets. With the annotation of argument concreteness, we have introduced a new scheme for assessing the quality of arguments that is particularly suitable for public participation processes with a spatial focus. The created sub-corpus for text-based document geo-location has established a new application domain. Lastly, we developed a generic schema of mobility for thematic categorization of contributions. The annotated contributions serve as a basis to build models that can universally aid in the topic classification of mobility-related planning processes.

A distinctive feature of our corpus is that the processes collected all deal with mobility-related goals. This is due to the fact that the dissertation at hand was written as part of a research project that examines citizen participation specifically with regard to the transformation of transport. Consequently, models may develop a specific preference for vocabulary associated with mobility during training. As the potential of supervised machine learning to support the evaluation of citizen input extends beyond mobility issues to other sectors such as environment, climate, housing, and education, it can therefore be desirable that trained models not only fulfill their primary application role, but may also be used more generally. To this end, we suggest that additional test datasets with more thematic variance should be added to validate the usefulness of models across sectors. What is more, we have specifically focused on informal participation. These processes are not mandated by law, so their design usually deviates from legally prescribed procedures, which can affect the structure of public input, such as the length and detail of comments. The corpus could thus benefit from expansion in this respect as well. In terms of process formats included, we already covered two prominent sources of written input, namely questionnaires (filled in digitally or handwritten) and comments from online participation platforms. For the sake of completeness, future work may consider other means of textual participation such as paper mail and e-mail (Shulman, 2009), although we suspect that these will play a lesser role in the future.

Reflecting on the size of the respective sub-corpora, the current focus is on argumentation mining. In particular, the amount of data available for the task of thematic categorization, aimed at developing an approach to classify mobility-related contributions into transportation categories, is sparse in the released version of the corpus. For example, the classes “public transport (long-distance)” and “inter- and multimodality” each contain only one contribution, turning their classification into an extreme few-shot problem. Notably, these classes need to be populated with examples so that common patterns can be derived from them. Therefore, we continued our work on the subcorpus after publication, resulting in a current expansion that more than doubled its size. It is planned to make this follow-up version publicly available in the near future.

A further aspect worth thinking of when it comes to the practical usefulness of models developed on the presented corpus lies in the annotation process itself. To perform the annotation tasks, we employed non-expert annotators who were trained using guidelines, some of which were developed in consultation with practitioners. While this corresponds to a common approach to annotating data in natural language processing, the question arises to what extent non-experts can cover the realities of public authorities when assessing given input, especially during the more advanced stages of evaluation. However, using experts with practical experience for downstream annotation seems more than unrealistic in terms of cost and availability. Decisions made as part of actual policy-making processes would provide the best data basis, but are usually not tracked or explicitly noted in the manner required for research purposes. It would therefore be of great interest for advancing the field to negotiate an agreement with public authorities and contracted service providers in order to be able to learn directly from the human evaluation efforts and, in conclusion, to be able to offer solutions that more closely resemble the working reality.

Finally, it should be noted that the annotation schemes still leave room for improvement. In this respect, the annotation of argument components shows the greatest potential. As of now, they follow the tried and tested classic premise-conclusion scheme. However, the dialogue with practitioners has since revealed that their requirements go beyond this basic concept. Rather, a supplementary classification according to factual evidence and personal opinion seems desirable. For German-language datasets from the field of public participation, no such annotation exists yet. However, Park et al. (2015) introduced a corresponding scheme in the context of U.S. eRulemaking, which can be referred to. What is more, we decided to postpone for the time being the relationships between individual argument components. In the long run, these links must be considered, as they are essential for a full understanding of the argument.

6.2 Empowering Topic Classification with Active Learning

With regard to the first focus of this work, the support of thematic pre-structuring by text classification, we specifically addressed the need for models that can be individually tailored to the processes in question. A prerequisite for such process-specific individualization of models to become practically useful is to keep manual effort low. We therefore emphasized the concept of active learning, which we consider a key methodology for a beneficial application of topic classification. In a case study, we demonstrated the potential of recent approaches for the domain of public participation (R4). Our

BERT-based solution yielded higher accuracy and a significant reduction in annotation overhead compared to previous work, while keeping model runtime affordable.

In detail, we were able to demonstrate the methodological potential for categorizing a collection by main topics (eight in our use case), resulting in an average accuracy of 0.77 and a reduction in annotation effort of up to 80% compared to a purely manual evaluation. As demonstrated in the study through an illustrative calculation, the savings are substantial even when misclassifications are taken into account. These very encouraging results signal a clear relief in pre-structuring according to the relevant fields of discussion. In addition to further improving performance, an upcoming direction for research would be to classify the contributions within each identified discussion area into more detailed subtopics. For example, the main categories of the three public participation processes that we considered in the case study can be subdivided into 30 subcategories. Initial research on the question whether active learning-based topic classification can also support a finer categorization of public participation input yields a mixed picture (Purpura et al., 2008; Thome, 2022). The increased number of classes and the simultaneously low occurrence of many classes in the training data pose difficulties for both traditional models and advanced language models. More in-depth research is needed to explore the potential of recent methods to handle such scenarios of data scarcity and imbalance. Here, the work of Dayanik et al. (2022) offers a promising approach that exploits hierarchical relations between categories to overcome the problems arising from the rare occurrence of fine-grained classes. Likewise, the SetFit classifier and its extensions (Tunstall et al., 2022; Bates and Gurevych, 2023), which are geared towards few-shot learning, could prove useful.

While we have successfully showcased the methodological possibilities, the feasibility of implementing active learning with large-scale language models in municipal settings still raises concerns as it has been found that the application of advanced technologies in public administrations is regularly hampered by a lack of technological expertise and capacity (Giest, 2017; Poel et al., 2018). Our study was conducted under the assumption of a well-equipped computing infrastructure including a graphics processing unit. Due to the model size (i.e., 110 million parameters) and the iterative approach of active learning, where the model is repeatedly fit to the pool of labeled data, fast computations are a necessity. Thus, it remains to be clarified which computing infrastructure can be used in municipalities or, in the case of outsourcing to service providers, which additional costs are indicated and whether these are bearable. This is where another difficulty comes into play: There are data protection regulations that must be complied with (e.g., the European Union's GDPR[1]). For example, the transmission of data to third parties, such as external servers running software solutions, can only take place if there is a legal basis for the transfer, such as the consent of the affected individual. Overall, current laws remain inadequately clarified in many aspects for dealing with automated decision-making (Cobbe, 2019; Busche, 2023), placing considerable pressure on the responsible public officers.

In the further course of the thesis, we turned more specifically to the evaluation of active learning and the demands of real-world applicants on query strategies, emphasizing practical utility as the primary goal of development (R2). For this purpose, we

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

have taken the input from practitioners and translated it into four measures that can be helpful in selecting a suitable strategy for specific use cases. We revealed that the common measure for imbalanced text classification scenarios, the macro F_1 score, cannot sufficiently account for class-related requirements in data selection that users may place on active learning in practice. As a consequence, we insisted that the measures currently in use for evaluating research setups of active learning need to be amended to more accurately simulate the needs of practitioners.

Our metrics draw their rationale from interviews with public authorities and service providers as well as from the literature. As a first step, our study focused on empirical evidence of the additional gain of information from the proposed measures in contrast to standard measures of evaluation. In the end, the pertinent question is how the choice of an active learning strategy supported by the user-centric measures affects the user experience and, thus, the evaluation process. A user study examining these aspects is therefore planned as a second step to prove the benefits of the additional measures.

With respect to our two publications on topic classification, it is worth noting that modifications in active learning experiments are particularly expensive due to the inherent repeated training of models over multiple iterations, especially in the era of large language models. As a result, the versatility of the experimental design is significantly curtailed given a natural limitation in computational capacity and available time. Thus, a meaningful simulation of active learning poses immense challenges to the entire field. Margatina and Aletras (2023) give a comprehensive overview of the many factors that count in. These include considerations of the active learning setup such as the choice of seed dataset, the number of iterations (i.e., the assumed acquisition budget), the choice of query strategy, and the choice of machine learning model including training specifics. Due to the intertwining of many different aspects, which mostly influence each other, experiments often provide only a limited understanding of the anticipated performance in real-world scenarios. The question arises as to the significance for practical application when, for example, basic settings are changed. This is why, in our experiments, we were careful to create a realistic environment by selecting small annotation batches (20 to 50 documents) and accounting for the waiting time between annotation cycles. In our work, hyperparameter tuning could be performed to a very limited extent only, but once the classifier and active learning strategy are determined, more effort can be devoted to this aspect in practical deployment. We therefore believe that the results of our experiments show a lower bound of the performance possible, leaving space for enhancement in practical applications.

Another aspect that can impact the success of active learning based methods is the human component. There has been little focus on this in computational linguistics research to date, as experiments are usually simulations using existing annotated datasets. However, this approach makes several assumptions that are generally not the rule in the real-world, such as the gold standard quality of human annotation, the inevitability of making annotation decisions, and the equal cost associated with each annotation decision (Donmez and Carbonell, 2008). It has been shown that humans in their role as annotators can influence active learning performance differently. For example, strategic sampling (i.e., query strategies that go beyond random sampling) can result in examples that are more difficult and time-consuming to annotate (Hachey et al., 2005). In addition, the effectiveness of active learning strategies may depend heavily on the expertise of the annotators (Baldridge and Palmer, 2009). To incorporate such factors into active learning simulations, the costs associated with the annotator need to be taken into account (e.g., Donmez and Carbonell 2008; Tomanek and Hahn 2010; Calma and Sick 2017). Our work has focused on another facet, namely the expectations that humans, as practical users, have of the behavior of active learning strategies. In order to comprehensively consider real-world conditions in the simulation of active learning to increase the validity of the findings, it is critical to thoroughly control for all effects arising from the human factor. Despite their significance, these aspects are still largely overlooked. Therefore, future work on active learning must urgently address this issue to make meaningful progress – both in general and specifically in the topic classification of citizen input.

On a general note, assigning textual contributions to topics can be approached in different ways. In our work we have emphasized supervised machine learning with a reduced training data expense through active learning. We chose this approach because we believe it resembles the natural workflow in public administration, where there is usually prior knowledge of the topics discussed. The literature, however, also stresses unsupervised solutions such as k-means and k-medoids clustering (Yang and Callan, 2009; Simonofski et al., 2021), non-negative matrix factorization (Arana-Catania et al., 2021a), associative networks (Teufl et al., 2009), and latent Dirichlet allocation (Hagen et al., 2015; Hagen, 2018; Arana-Catania et al., 2021a). While the primary challenge in using supervised machine learning is the necessary amount of training data, the fundamental problem with unsupervised machine learning methods is that the discovered topic clusters often do not align with the user's requirements. In order to use the latter in a useful way, it was suggested to provide human supervision for the adaptation of the clustering to a user's needs via interactive topic modeling (Cai et al., 2018).

To make automation attractive to practitioners, this means that either the required supervision in supervised approaches must be reduced or that supervision must be introduced in unsupervised approaches. Despite the drastic reduction of manual effort that we could demonstrate in our studies, it would be fair to shed light on the potentials of the other side as well. Unsupervised machine learning methods like structural topic models (Roberts et al., 2013), which have shown promising results in a variety of social science applications (e.g. Lucas et al., 2015; Chen et al., 2020; Ferrario and Stantcheva, 2022), should be examined for their performance and the required additional manual effort for adaptations. A direct comparison of the two strands should then be carried out by means of a user study. It is of particular interest here to investigate what impact solutions initiating collaboration between humans and machines can have in public administration – possibly even beyond the mere goodness of topic assignment. For example, it is conceivable that such close interaction makes it easier for users to develop trust in machine predictions. A – at least perceived – direct influence on the automated processes could possibly create a sense of greater controllability, which might increase the acceptance of the new technologies. These qualities can provide an additional incentive to deploy interactive machine learning solutions.

6.3 Robust and Subjective Argumentation Mining

The second major focus of this thesis was on the detailed analysis of input by means of arguments. Here, we had a look at machine learning methods for argument component identification and classification regarding their in-dataset performance, as well as their performance across different public participation datasets that were not part of the training process (R3). We found remarkable and robust performance of pre-trained language models that generalizes to further datasets with little deviation. This constitutes an important leap towards practical applicability. In particular, BERT-based solutions demonstrated a strong performance in the classification of argumentative sentences as either major positions or premises on data from our domain of application, yielding an average macro F_1 value of 0.90.

Conversely, the identification of sentences as argumentative or not argumentative continues to require major attention in order to further improve the current average macro F_1 value of 0.77 that we were able to achieve. One reason why the models have particular difficulty learning to distinguish between non-argumentative and argumentative sentences may be the strong imbalance between these two classes (15% versus 85%). For this reason, methods that take greater account of the challenges associated with imbalanced datasets appear promising for improving our approach in future research, such as augmenting the minority class with additional data or making use of specific loss functions. Henning et al. (2023) provide an extensive survey of such methods aimed at enhancing the adaptation of deep learning to imbalanced datasets, from which one can draw motivation.

Moreover, we proposed future research directions for improving the machine-assisted analysis of fundamental argumentation structures in public input in the discussion of the CIMT PartEval Corpus in Section 6.1. These include improving argument component classification through an argumentation scheme that is even more tailored to practitioners' needs and exploring the relations between different parts of the argument.

In addition to fundamental argument structures, we focused on assessing the overall quality of citizens' arguments in spatial planning by accounting for the degree of concreteness propositions exhibit (R1). Our experiments demonstrated the potential that BERT holds for determining the concreteness of arguments, with a pleasing average accuracy of 0.79. Nevertheless, the correspondingly reached averaged macro F_1 value of 0.67 is somewhat sobering. The similar performance across the different subsets of the data (i.e., joint or split by the different types of argument components) that we found indicates that it is not the small overall size of the dataset that is causing the problem. Instead, the problem seems to rather lie in adapting BERT to the extreme data imbalance between the three classes (63% high, 30% intermediate, and 7% low concreteness). Notably, the smallest class occurs only 82 times in the entire dataset. Once again, as with argumentative sentence detection, attention in follow-up work should therefore be paid to corresponding techniques, such as data augmentation or an exchange of loss functions.

On top of argument concreteness, other quality traits of argumentation may also provide guidance for the human analysts when evaluating public input. Existing work could be exploited here, for example, on the global relevance of arguments (Wachsmuth et al., 2017b). This approach is of particular interest for our scenario because it does not base the relevance assessment on human judgments (i.e., transferred to our use case, how public authorities or service providers perceive argument relevance), but more objectively considers the structural prevalence of arguments in a collection (i.e., transferred to our use case, how strongly these arguments are represented in the citizenry). Thus, such a relevance ranking reflects the arguments to which citizens themselves collectively attach more importance. It might also be helpful to examine the validity and novelty of argumentative conclusions (Heinisch et al., 2022) in order to get clues about the conclusiveness of the argument and the added value of the content. All these indicators, just like the concreteness of arguments, can be used to pre-structure the citizens' reasoning for manual analysis.

What is more, we addressed the subjective nature of argumentation using the perception of concreteness as an example. While related work has explored the backing mechanisms that lead to different perspectives in argumentation (Ajjour et al., 2019; Kobbe et al., 2020; Kiesel et al., 2022; Alshomary et al., 2022; Falk and Lapesa, 2022a, 2023b) or has covered the range of perspectives in the output, e.g., by retrieving a variety of valid stances for a given claim (Chen et al., 2019), we contributed the first approach to the field of argumentation mining that integrates multiple perspectives as knowledge into the machine learning process itself. In doing so, our method PerspectifyMe predicts a value of subjectivity in parallel to an aggregated ground truth in order to inform the user about possible valid variations in label decisions.

In our use case of argument concreteness, we could see that predicting whether an argumentative text unit triggers different perspectives is feasible by machine to a certain extent. In particular, we found that length was a good indicator, but surprisingly, pretrained language knowledge had a very limited effect. It remains an open question which characteristics constitute the perception of examples that could not be correctly classified as subjective or objective so far. At this stage, it is still unclear whether the task of subjectivity prediction can be solved at all based on the textual representation of arguments alone, or whether further information is required to do so. There is an urgent need to fill this knowledge gap in future work. The exploration of correlations between variation in labeling behavior and individual traits like personal experiences or socio-demographic characteristics is also relevant in this context, as these may provide clues to the causes of differing perspectives.

Especially when assessing the quality of argumentation, multiple perspectives are often permissible, the suppression of which cannot be purposeful in the long run. PerspectifyMe is a first attempt to integrate human label variation into machine learning workflows of argumentation mining with the strength of complementing existing machine learning workflows with a second model for predicting subjectivity. So far, we made use of two classification models in this process that were separately trained. However, subsequent work should also look at whether learning both tasks simultaneously through multi-task models can lead to improved performance. The benefits arising from such synergies have been identified as helpful in previous work on argumentation mining (Schulz et al., 2018; Morio et al., 2022). An important next step towards the inclusion of perspectives in argumentation mining will furthermore consist in applying multi-annotator models that explicitly learn from the non-aggregated ground truth, as suggested by Davani et al. (2022).

The framework we presented offers many possibilities. It is not limited to classification tasks, but can also be extended to regression tasks. Likewise, the approach is not restricted to the field of argumentation mining, but is in principle conceivable for any other subjective task, either in the context of supporting the evaluation of public participation or beyond. Consideration may also be given to using this approach with objectives other than pointing out subjective items to an end user. For example, the prediction of potentially subjective label decisions may be used in low-resource dataset creation scenarios to identify items that should be reviewed by multiple annotators.

6.4 Concluding Remarks

Following the comprehensive analysis of individual contributions and suggesting future research avenues for each part of the thesis, we now draw an overall conclusion on the goal of machine-assisted text classification of citizen contributions, and the potential of machine learning for public participation in general.

We have emphasized research questions of practical relevance, such as the conditions under which active learning-embedded text classification can be profitable, the much needed robustness of models in practical use, and the explicit consideration of subjective label decisions inherent in human natural language understanding. As a means to achieve this, we utilized selected machine learning methods that are known for their good performance in text classification tasks (i.e., BERT). To further improve the results obtained, future work might want to consider a broader range of approaches. In this context, student work (Thome, 2022; Padjman, 2022) in continuation to our publications R3 and R4 has taken a look at the parameter-reduced model variant DistilBERT (Sanh et al., 2019). In both tasks, topic classification and argument component classification, DistilBERT was able to keep up with the predictive performance of BERT with only little deviation, while cutting down significantly on the amount of time required for training and inference. This insight is particularly important for the iterative and thus expensive methodology of active learning. Furthermore, it became apparent that GPT-2 (Radford et al., 2019), as an alternative topic classification model, fell off considerably compared to BERT (Thome, 2022). The decisive factor is presumably the model size, which seems to require more training data for a good fine-tuning than is provided by the public participation processes considered here, especially when concurrently aiming at minimizing manual annotation effort. This limitation could likely apply to other models with an excessive number of parameters as well, such as the subsequent GPT versions. Nevertheless, it would be intriguing to investigate what advantage, for example, the large general knowledge of the highly praised ChatGPT[2] can provide for text classification tasks in our application domain. This might include exploring whether a prompt-based approach with human annotation guidelines provided could enhance the identification of argumentative sentences.

To evaluate the performance of the applied text classification methods, we have given priority to statistical measures within the scope of this thesis. This course of action has provided us with valuable insights, especially with regard to prediction accuracy. In the interest of practical orientation, in the long term we plan to evaluate the methods directly in practice by users. In this way, additional insights into the methods' weaknesses and strengths can be gained, which may not have been considered or noticed during development. The positive findings of such a user study can drive progress and, furthermore, provide confirmation of existing concepts. The negative findings can in turn be used to address important issues in improving approaches, for example, by means of practice-oriented measures as we suggested in publication R2. In the end, a variety of aspects matter for methods to be effective in practice and for users to also want to use them. These can only be uncovered in the best possible way using different evaluation angles that complement each other.

Within the manual evaluation cycle, there are generally a number of starting points where machine learning can assist human analysts. As part of this thesis, we have

[2] https://openai.com/blog/chatgpt

specifically concentrated on two key tasks, namely the thematic categorization of contributions and the analysis of arguments. There is also active, albeit less, research on the summarization of public input (Arana-Catania et al., 2021b) and the analysis of citizens' sentiment and emotion in comments (Aitamurto et al., 2016; Jasim et al., 2021). Further tasks along the way to support practice include detecting duplicate contributions (Yang and Callan, 2005, 2006; Yang et al., 2006), identifying stakeholders (Arguello and Callan, 2007), assessing the urgency of urban issues (Masdeval and Veloso, 2015), or detecting the textual description of locations in public participation processes with spatial reference (Padjman, 2021; using the CIMT Geographic Location sub-corpus). What all these works have in common is that they focus on contributions available in textual form. We are not aware of any work that explicitly attempts to support the evaluation of participation that is collected in other formats, such as images, videos, or audio recordings[3]. However, the public can utter its opinion in very different ways. For example, a common means of participation is through workshops, either on-site or online, where ideas are expressed primarily verbally. When it comes to asking citizens for help in identifying road hazards or other infrastructure deficiencies, photographs are a viable tool for getting information quickly. As these examples make clear, it would be highly beneficial not only to focus on the possibilities of machine learning support for (digitized) texts, but to broaden the picture in order to arrive at a more holistic assistance. And even if spoken language may be converted to written text with little loss using transcription software, image and video data do require their own unique solutions.

Future work should also address the interplay of different communication media (i.e., multimodality) within formats of participation instead of developing isolated solutions in such cases. For example, the chat function in web services deployed for online workshops can be used to intervene in the ongoing verbal discussion through textual input, on-site workshop organizers often resort to additional resources such as whiteboards or maps to summarize discussion points, and citizens can back up their textual statements on online participation platforms with images to substantiate them. In these situations, a joint machine learning solution seems logical and can even be a necessity in order to provide the best possible support. Starting points for research on multimodal argumentation mining, for example, are provided by Mestre et al. (2021, 2023)'s work using presidential debates in text and audio to detect arguments and classify argument relations, and Liu et al. (2022)'s work on predicting stance and persuasiveness of a tweet in consideration of accompanying images.

Further, of course, it is not only the machine support of evaluating the final collection of input gathered through public participation processes that is rewarding. The support at other stages of the process is equally essential. On the one hand, this includes the preparation and organization of the participation. For example, machine learning could be used to advertise public participation to specific audiences, potentially improving the representation of different socio-demographic groups in the process, which constitutes a frequent problem in participation efforts (Verba et al., 1995; Marien et al., 2010). Another application scenario is the design of public participation processes, in which the extensive knowledge of models such as ChatGPT can be helpful in selecting the most appropriate participation formats for a given context. On the other

[3] We hereby refer to audio recordings in their raw form and not as transcriptions, which we consider to be in textual format.

hand, the phase of active participation, e.g. through deliberative events, offers great potential for machine support. This includes improving the experience of citizens on online platforms (Arana-Catania et al., 2021a) and assistance with the moderation of discussions (Falk et al., 2021). Related work reveals the many methodological overlaps with the step of subsequent evaluation: Techniques for structuring debates thematically find favor as do argumentation mining methods, which have shown potential in improving online citizen debates (Ito et al., 2022; Anastasiou and De Liddo, 2021). The approaches are often quite similar, although they are applied in distinct contexts (supporting the ongoing participation or the evaluation of completed participation) and with different goals (supporting the analyst, the moderator, or the citizen). This is why future research could benefit from exploiting these synergies to a greater extent.

Apart from all these opportunities to further advance research on automated support for public administration in the evaluation of public participation processes and beyond, however, the most critical issue remains the transfer into practice[4]. As we discovered during our literature review, in many research projects, the final step towards the explicit implementation of the often promising research results is not carried out at all or does not succeed. This applies in particular to the desirable open source solutions that prioritize transparency and free use as licensed, making them accessible to a wide range of users. Research initiatives usually seem to fail in getting to a finished product that meets the needs of end users because of the level of effort and cost required. The reasons can be manifold, such as a lack of expertise in ready-to-use software deployment, the usually short periods of project funding or the prioritization of research (often due to the employment of doctoral students that are under pressure to perform within their own discipline). However, if public participation is to be sustainably strengthened in the coming decades through the use of machine learning, it is crucial to meticulously contemplate long-term funding for initiatives and a strategic emphasis on roles like proficient software engineers responsible for effective implementation. A potential approach to realize this goal might involve establishing binding partnerships between research institutions, public authorities and service providers. In this collaborative framework, the tasks involved in developing the final software could be distributed, allowing the different parties to leverage each other's specialized knowledge and skills.

In conclusion, the mechanism of public participation holds a significant role within the democratic framework. By involving citizens in political decision-making, it promotes a well-informed process and provides an opportunity to increase public acceptance. Empirical research has demonstrated that involving the public can positively impact decision-making processes (Dietz and Stern, 2008; Hudson, 2018; Chen and Aitamurto, 2019; Jager et al., 2019), but the manual evaluation of collected input is regularly challenged by the amount of contributions and temporal restrictions. At the same time, it is important to maintain the high standards of democratic procedures. As we were able to confirm, machine learning offers great potential to escape this dilemma by fulfilling an assisting role in the various sub-tasks of evaluation. Specifically, the pre-structuring of citizen input greatly benefits from text classification algorithms, as demonstrated by our different studies. However, as soon as it comes to more far-reaching support, such as actual decision-making, the possibilities that ma-

[4] The methods developed as part of this thesis will also eventually be made available as open source software.

chine learning offers are strongly curtailed by the fundamental democratic principles to which public participation is subject. A concern that arises in this context is certainly whether the use of such technologies for governmental procedures with important implications for the democratic process is ethically justifiable. Decisive factors in this regard include the transparency with which algorithms arrive at their results (Daniell et al., 2016), the question of accountability for machine decisions (König and Wenzelburger, 2020), and the (perceived) fairness of these (Starke et al., 2022). From the public authorities' point of view, it is therefore essential that each contribution is reviewed at least once by a human analyst and that the ultimate step of decision-making remains under human control in order to ensure a transparent treatment in which all opinions are given equal consideration (Dahl, 1989). Only if it is ensured that machine solutions meet these high requirements, if they can arrive at fair decisions and justify them to citizens and authorities, is an expansion of support conceivable.

On the one hand, there thus remains a significant need for further research, encompassing efforts to enhance the transparency and fairness of the already powerful language models. On the other hand, society must also take a step toward the "artificial intelligence". Establishing more trust in machine solutions is of utmost importance (e.g., through interaction and collaboration as discussed in Section 6.2). Society needs to open up to technological progress, especially if these methods are meant only as a support and not as a substitute for human labor. Ultimately, the power of artificial intelligence, and consequently machine learning, for society lies in the individual strengths of humans and machines that can complement each other. The use case highlighted in this thesis, namely the machine-assisted text classification of public participation contributions, serves as an excellent example of this symbiotic potential.

Publications of the Author

Publications Forming Part of the Dissertation

[R1] J. Romberg. Is your perspective also my perspective? Enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125. International Conference on Computational Linguistics, 2022. URL `https://aclanthology.org/2022.argmining-1.11`.

[R2] J. Romberg. Mind the user! Measures to more accurately evaluate the practical value of active learning strategies. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 996–1006. INCOMA Ltd., 2023. URL `https://aclanthology.org/2023.ranlp-1.107/`.

[R3] J. Romberg and S. Conrad. 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. Association for Computational Linguistics, 2021. URL `https://aclanthology.org/2021.argmining-1.9`.

[R4] J. Romberg and T. Escher. Automated topic categorisation of citizens' contributions: Reducing manual labelling efforts through active learning. In *Electronic Government*, pages 369–385. Springer, 2022. URL `https://link.springer.com/chapter/10.1007/978-3-031-15086-9_24`.

[R5] J. Romberg and T. Escher. Making sense of citizens' input through artificial intelligence: A review of methods for computational text analysis to support the evaluation of contributions in public participation. *Digital Government: Research and Practice*, 2023. URL `https://dl.acm.org/doi/10.1145/3603254`. Just Accepted.

[R6] J. Romberg, L. Mark, and T. Escher. A corpus of German citizen contributions in mobility planning: Supporting evaluation through multidimensional classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2874–2883. European Language Resources Association, 2022. URL `https://aclanthology.org/2022.lrec-1.308`.

Further Publications

[F1] T. Cabanski, J. Romberg, and S. Conrad. HHU at SemEval-2017 Task 5: Fine-grained sentiment analysis on financial data using machine learning methods. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 832–836. Association for Computational Linguistics, 2017. URL `https://aclanthology.org/S17-2141`.

[F2] A. Oberstrass, J. Romberg, A. Stoll, and S. Conrad. HHU at SemEval-2019 Task 6: Context does matter - Tackling offensive language identification and categorization with ELMo. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 628–634. Association for Computational Linguistics, 2019. URL `https://aclanthology.org/S19-2112`.

[F3] J. Romberg. Actor identification and relevance filtering in movie reviews. In *Proceedings of the 28th GI-Workshop on Foundations of Databases*, pages 92–97. CEUR Workshop Proceedings, 2016. URL `https://ceur-ws.org/Vol-1594/paper17.pdf`.

[F4] J. Romberg. Comparing relevance feedback techniques on German news articles. In *Datenbanksysteme für Business, Technologie und Web (BTW 2017) - Workshopband*, pages 301–310. Gesellschaft für Informatik e.V., 2017. URL `https://dl.gi.de/handle/20.500.12116/926`.

[F5] J. Romberg. GDWDS: First insights from a student-based key phrase annotation process of medical information needs on a novel German diabetes web data set. In *Proceedings of the 30th GI-Workshop on Foundations of Databases*, pages 89–94. CEUR Workshop Proceedings, 2018. URL `https://ceur-ws.org/Vol-2126/paper14.pdf`.

[F6] J. Romberg, J. Dyczmons, S. O. Borgmann, J. Sommer, M. Vomhof, C. Brunoni, I. Bruck-Ramisch, L. Enders, A. Icks, and S. Conrad. Annotating patient information needs in online diabetes forums. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 19–26. Association for Computational Linguistics, 2020. URL `https://aclanthology.org/2020.smm4h-1.3`.

Bibliography

T. Aitamurto, K. Chen, A. Cherif, J. S. Galli, and L. Santana. Civic CrowdAnalytics: Making sense of crowdsourced civic input with big data tools. In *Proceedings of the 20th International Academic Mindtrek Conference*, pages 86–94. Association for Computing Machinery, 2016. URL https://dl.acm.org/doi/abs/10.1145/2994310.2994366.

Y. Ajjour, M. Alshomary, H. Wachsmuth, and B. Stein. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2922–2932. Association for Computational Linguistics, 2019. URL https://aclanthology.org/D19-1290.

M. Alliheedi, R. E. Mercer, and R. Cohen. Annotation of rhetorical moves in biochemistry articles. In *Proceedings of the 6th Workshop on Argument Mining*, pages 113–123. Association for Computational Linguistics, 2019. URL https://aclanthology.org/W19-4514.

M. Alshomary and H. Wachsmuth. Conclusion-based counter-argument generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 957–967. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.eacl-main.67.

M. Alshomary, S. Syed, M. Potthast, and H. Wachsmuth. Target inference in argument conclusion generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.acl-main.399.

M. Alshomary, R. El Baff, T. Gurcke, and H. Wachsmuth. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.acl-long.601.

L. Anastasiou and A. De Liddo. Making sense of online discussions: Can automated reports help? In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7. Association for Computing Machinery, 2021. URL https://doi.org/10.1145/3411763.3451815.

M. Arana-Catania, F.-A. V. Lier, R. Procter, N. Tkachenko, Y. He, A. Zubiaga, and M. Liakata. Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice*, 2(3):1–22, 2021a. URL `https://dl.acm.org/doi/abs/10.1145/3452118`.

M. Arana-Catania, R. Procter, Y. He, and M. Liakata. Evaluation of abstractive summarisation models with machine translation in deliberative processes. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 57–64. Association for Computational Linguistics, 2021b. URL `https://aclanthology.org/2021.newsum-1.7`.

J. Arguello and J. Callan. A bootstrapping approach for identifying stakeholders in public-comment corpora. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, pages 92–101. Digital Government Society of North America, 2007. URL `https://dl.acm.org/doi/abs/10.5555/1248460.1248475`.

Á. Þ. Árnason and C. Dupré. *Icelandic Constitutional Reform: People, Processes, Politics*. Routledge, 2020.

J. Baldridge and A. Palmer. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305. Association for Computational Linguistics, 2009. URL `https://aclanthology.org/D09-1031`.

D. Balta, P. Kuhn, M. Sellami, D. Kulus, C. Lieven, and H. Krcmar. How to streamline AI application in government? A case study on citizen participation in Germany. In *Electronic Government*, pages 233–247. Springer, 2019. URL `https://link.springer.com/chapter/10.1007/978-3-030-27325-5_18`.

L. Bates and I. Gurevych. Like a good nearest neighbor: Practical content moderation with sentence transformers. *arXiv preprint arXiv:2302.08957*, 2023. URL `https://arxiv.org/abs/2302.08957`.

O. Belkahla Driss, S. Mellouli, and Z. Trabelsi. From citizens to government policy-makers: Social media data analysis. *Government Information Quarterly*, 36(3): 560–570, 2019. URL `https://www.sciencedirect.com/science/article/pii/S0740624X18302983`.

P. Besnard and A. Hunter. *Elements of Argumentation*. MIT Press, 2008.

J. A. Blair. *Groundwork in the Theory of Argumentation*. Springer, 2012.

L. Bobbio. Designing effective public participation. *Policy and Society*, 38(1):41–57, 2018. URL `https://academic.oup.com/policyandsociety/article/38/1/41/6403983`.

S. Bock and B. Reimann. Mit dem Los zu mehr Vielfalt in der Bürgerbeteiligung? Chancen und Grenzen der Zufallsauswahl. *Kursbuch Bürgerbeteiligung# 4*, pages 184–199, 2021.

R. J. Branham. *Debate and Critical Analysis: The Harmony of Conflict*. Routledge, 2013.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

K. Budzynska, M. Janier, J. Kang, C. Reed, P. Saint-Dizier, M. Stede, and O. Yaskorska. Towards argument mining from dialogue. In *Proceedings of the Fifth International Conference on Computational Models of Argument*, pages 185–196. IOS Press, 2014. URL `https://ebooks.iospress.nl/doi/10.3233/978-1-61499-436-7-185`.

D. Busche. Einführung in die Rechtsfragen der künstlichen Intelligenz. *JA*, 6:441–446, 2023.

F. Cabitza, A. Campagner, and V. Basile. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868, 2023. URL `https://ojs.aaai.org/index.php/AAAI/article/view/25840`.

G. Cai, F. Sun, and Y. Sha. Interactive visualization for topic model curation. In *Proceedings of the Workshop on Exploratory Search and Interactive Data Analytics*. CEUR Workshop Proceedings, 2018. URL `https://ceur-ws.org/Vol-2068/esida5.pdf`.

A. Calma and B. Sick. Simulation of annotators for active learning: Uncertain oracles. In *Proceedings of the Workshop and Tutorial on Interactive Adaptive Learning*, pages 49–58. CEUR Workshop Proceedings, 2017. URL `https://ceur-ws.org/Vol-1924/ialatecml_paper4.pdf`.

C. Cardie, C. R. Farina, and T. R. Bruce. Using natural language processing to improve eRulemaking [project highlight]. *Cornell e-Rulemaking Initiative Publications*, Paper 9, 2006. URL `https://scholarship.law.cornell.edu/ceri/9`.

C. Cardie, C. Farina, A. Aijaz, M. Rawding, and S. Purpura. A study in rule-specific issue categorization for e-Rulemaking. In *Proceedings of the 9th Annual International Conference on Digital Government Research*, pages 244–253. Digital Government Research Center, 2008a. URL `https://scholarship.law.cornell.edu/ceri/4/`.

C. Cardie, C. Farina, M. Rawding, and A. Aijaz. An eRulemaking corpus: Identifying substantive issues in public comments. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 2757–2763. European Language Resources Association, 2008b. URL `http://www.lrec-conf.org/proceedings/lrec2008/pdf/699_paper.pdf`.

K. Chen and T. Aitamurto. Barriers for crowd's impact in crowdsourced policymaking: Civic data overload and filter hierarchy. *International Public Management*

Journal, 22(1):99–126, 2019. URL https://www.tandfonline.com/doi/full/10.1080/10967494.2018.1488780.

K. Chen, L. Bao, A. Shao, P. Ho, S. Yang, C. D. Wirz, D. Brossard, M. Brauer, and L. DiPrete Brown. How public perceptions of social distancing evolved over a critical time period: Communication lessons learnt from the American state of Wisconsin. *Journal of Science Communication*, 19(05), 2020. URL https://doi.org/10.22323/2.19050211.

S. Chen, D. Khashabi, W. Yin, C. Callison-Burch, and D. Roth. Seeing things from a different angle: Discovering diverse perspectives about claims. 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 542–557, 2019. URL https://aclanthology.org/N19-1053.

J. Cobbe. Administrative law and the machines of government: Judicial review of automated public-sector decision-making. *Legal Studies*, 39(4):636–655, 2019. URL https://doi.org/10.1017/lst.2019.9.

O. Cocarascu, E. Cabrio, S. Villata, and F. Toni. Dataset independent baselines for relation prediction in argument mining. *Frontiers in Artificial Intelligence and Applications*, 326:45–52, 2020. URL https://ebooks.iospress.nl/publication/55356.

D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996. URL https://www.jair.org/index.php/jair/article/view/10158.

R. A. Dahl. *Democracy and Its Critics*. Yale University Press, 1989.

S. Damer and C. Hague. Public participation in planning: A review. *The Town Planning Review*, 42(3):217–232, 1971. URL https://www.jstor.org/stable/40102750.

K. A. Daniell, A. Morton, and D. Ríos Insua. Policy analysis and policy analytics. *Annals of Operations Research*, 236:1–13, 2016. URL https://doi.org/10.1007/s10479-015-1902-9.

A. M. Davani, M. Díaz, and V. Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. URL https://aclanthology.org/2022.tacl-1.6.

E. Dayanik, A. Blessing, N. Blokker, S. Haunss, J. Kuhn, G. Lapesa, and S. Padó. Improving neural political statement classification with class hierarchical information. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2367–2382. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.findings-acl.186.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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. Association for Computational Linguistics, 2019. URL https://aclanthology.org/N19-1423.

T. Dietz and P. C. Stern. *Public Participation in Environmental Assessment and Decision Making*. The National Academies Press, 2008. URL https://doi.org/10.17226/12434.

J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020. URL https://arxiv.org/abs/2002.06305.

P. Donmez and J. G. Carbonell. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 619–628. Association for Computing Machinery, 2008. URL https://doi.org/10.1145/1458082.1458165.

S. Dutta, J. Juneja, D. Das, and T. Chakraborty. Can unsupervised knowledge transfer from social discussions help argument mining? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7774–7786. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.acl-long.536.

V. Eidelman and B. Grom. Argument identification in public comments from eRulemaking. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 199–203. Association for Computing Machinery, 2019. URL https://dl.acm.org/doi/10.1145/3322640.3326714.

L. Ein-Dor, A. Halfon, A. Gera, E. Shnarch, L. Dankin, L. Choshen, M. Danilevsky, R. Aharonov, Y. Katz, and N. Slonim. Active learning for BERT: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7949–7962. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.emnlp-main.638.

L. Ein-Dor, E. Shnarch, L. Dankin, A. Halfon, B. Sznajder, A. Gera, C. Alzate, M. Gleize, L. Choshen, Y. Hou, Y. Bilu, R. Aharonov, and N. Slonim. Corpus wide argument mining - A working solution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7683–7691, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6270.

R. El Baff, H. Wachsmuth, K. Al Khatib, and B. Stein. Analyzing the persuasive effect of style in news editorial argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.acl-main.287.

P. Esaiasson. Will citizens take no for an answer? What government officials can do to enhance decision acceptance. *European Political Science Review*, 2(3):351–371, 2010. URL https://doi.org/doi:10.1017/S1755773910000238.

K. Esau. Capturing citizens' values: On the role of narratives and emotions in digital participation. *Analyse & Kritik*, 40(1):55–72, 2018. URL https://doi.org/10.1515/auk-2018-0003.

N. Falk and G. Lapesa. Reports of personal experiences and stories in argumentation: Datasets and analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5530–5553. Association for Computational Linguistics, 2022a. URL https://aclanthology.org/2022.acl-long.379.

N. Falk and G. Lapesa. Scaling up discourse quality annotation for political science. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3301–3318. European Language Resources Association, 2022b. URL https://aclanthology.org/2022.lrec-1.353.

N. Falk and G. Lapesa. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488. Association for Computational Linguistics, 2023a. URL https://aclanthology.org/2023.findings-eacl.187.

N. Falk and G. Lapesa. StoryARG: A corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372. Association for Computational Linguistics, 2023b. URL https://aclanthology.org/2023.acl-long.132.

N. Falk, I. Jundi, E. M. Vecchi, and G. Lapesa. Predicting moderation of deliberative arguments: Is argument quality the key? In *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.argmining-1.13.

B. Ferrario and S. Stantcheva. Eliciting people's first-order concerns: Text analysis of open-ended survey questions. *AEA Papers and Proceedings*, 112:163–69, 2022. URL https://www.aeaweb.org/articles?id=10.1257/pandp.20221071.

C. Fierro, C. Fuentes, J. Pérez, and M. Quezada. 200K+ crowdsourced political arguments for a new Chilean constitution. In *Proceedings of the 4th Workshop on Argument Mining*, pages 1–10. Association for Computational Linguistics, 2017. URL https://aclanthology.org/W17-5101.

J. B. Freeman. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. De Gruyter Mouton, 1991. URL https://www.degruyter.com/document/doi/10.1515/9783110875843/html.

General Secretariat, Presidency of Chile. *Quantitative Summary of the 2016 Chilean Constituent Process, Participative Phase (in Spanish)*. Ministry General Secretariat of the Presidency of Chile, 2017. Available at http://archivoweb.bibliotecanacionaldigital.cl/unaconstitucionparachile/2017-03-08/sintesis_de_resultados_etapa_participativa.pdf (Accessed 17 December 2022).

A. Giannakopoulos, M. Coriou, A. Hossmann, M. Baeriswyl, and C. Musat. Resilient combination of complementary CNN and RNN features for text classification through attention and ensembling. In *2019 6th Swiss Conference on Data Science*, pages 57–62. IEEE, 2019. URL https://ieeexplore.ieee.org/abstract/document/8789863.

S. Giest. Big data for policymaking: Fad or fasttrack? *Policy Sci*, 50:367–382, 2017. URL https://link.springer.com/article/10.1007/s11077-017-9293-1.

S. Gretz, R. Friedman, E. Cohen-Karlik, A. Toledo, D. Lahav, R. Aharonov, and N. Slonim. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6285.

I. Habernal and I. Gurevych. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599. Association for Computational Linguistics, 2016. URL https://aclanthology.org/P16-1150.

B. Hachey, B. Alex, and M. Becker. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 144–151. Association for Computational Linguistics, 2005. URL https://aclanthology.org/W05-0619.

L. Hagen. Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management*, 54(6): 1292–1307, 2018. URL https://www.sciencedirect.com/science/article/pii/S0306457317307240.

L. Hagen, Ö. Uzuner, C. Kotfila, T. M. Harrison, and D. Lamanna. Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach. In *2015 48th Hawaii International Conference on System Sciences*, pages 2134–2143. IEEE, 2015. URL https://ieeexplore.ieee.org/abstract/document/7070069.

P. Heinisch, A. Frank, J. Opitz, M. Plenz, and P. Cimiano. Overview of the 2022 validity and novelty prediction shared task. In *Proceedings of the 9th Workshop on Argument Mining*, pages 84–94. International Conference on Computational Linguistics, 2022. URL https://aclanthology.org/2022.argmining-1.7.

S. Henning, W. Beluch, A. Fraser, and A. Friedrich. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, 2023. URL https://aclanthology.org/2023.eacl-main.38.

J. Hessel and L. Lee. Something's brewing! Early prediction of controversy-causing posts from discussion features. 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 1648–1659. Association for Computational Linguistics, 2019. URL https://aclanthology.org/N19-1166.

T. A. Hollihan and K. T. Baaske. *Arguments and Arguing: The Products and Process of Human Decision Making.* Waveland Press, 2022.

A. Holub, P. Perona, and M. C. Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008. URL https://ieeexplore.ieee.org/document/4563068.

J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics, 2018. URL https://aclanthology.org/P18-1031.

X. Hua and L. Wang. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230. Association for Computational Linguistics, 2018. URL https://aclanthology.org/P18-1021.

X. Hua and L. Wang. Efficient argument structure extraction with transfer learning and active learning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 423–437. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.findings-acl.36.

A. Hudson. When does public participation make a difference? Evidence from Iceland's crowdsourced constitution. *Policy & Internet*, 10(2):185–217, 2018. URL https://onlinelibrary.wiley.com/doi/full/10.1002/poi3.167.

J. E. Innes and D. E. Booher. Reframing public participation: Strategies for the 21st century. *Planning Theory & Practice*, 5(4):419–436, 2004. URL https://www.tandfonline.com/doi/abs/10.1080/1464935042000293170.

T. Ito, R. Hadfi, and S. Suzuki. An agent that facilitates crowd discussion. *Group Decision and Negotiation*, 31:621–647, 2022. URL https://link.springer.com/article/10.1007/s10726-021-09765-8.

N. W. Jager, J. Newig, E. Challies, and E. Kochskämper. Pathways to implementation: Evidence on how participation in environmental governance impacts on environmental outcomes. *Journal of Public Administration Research and Theory*, 30(3):383–399, 2019. URL https://doi.org/10.1093/jopart/muz034.

M. Jasim, E. Hoque, A. Sarvghad, and N. Mahyar. Communitypulse: Facilitating community input analysis by surfacing hidden insights, reflections, and priorities. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, pages 846–863. Association for Computing Machinery, 2021. URL https://doi.org/10.1145/3461778.3462132.

M. Kaase. Politische Beteiligung/Politische Partizipation. *Handwörterbuch des politischen Systems der Bundesrepublik Deutschland*, pages 473–478, 2000. URL https://link.springer.com/chapter/10.1007/978-3-322-93232-7_105.

P. Karisani, N. Karisani, and L. Xiong. Multi-view active learning for short text classification in user-generated data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6441–6453. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.findings-emnlp.481.

J. Kiesel, M. Alshomary, N. Handke, X. Cai, H. Wachsmuth, and B. Stein. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.acl-long.306.

B. Kim, M. Yoo, K. C. Park, K. R. Lee, and J. H. Kim. A value of civic voices for smart city: A big data analysis of civic queries posed by Seoul citizens. *Cities*, 108:102941, 2021. URL https://www.sciencedirect.com/science/article/abs/pii/S0264275120312890.

J. Kobbe, I. Rehbein, I. Hulpuș, and H. Stuckenschmidt. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.argmining-1.4.

B. Konat, J. Lawrence, J. Park, K. Budzynska, and C. Reed. A corpus of argument networks: Using graph properties to analyse divisive issues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3899–3906. European Language Resources Association, 2016. URL https://aclanthology.org/L16-1617.

D. Kottke, A. Calma, D. Huseljic, G. M. Krempl, and B. Sick. Challenges of reliable, realistic and comparable active learning evaluation. In *Proceedings of the Workshop and Tutorial on Interactive Adaptive Learning*, pages 2–14. CEUR Workshop Proceedings, 2017. URL https://ceur-ws.org/Vol-1924/ialatecml_paper0.pdf.

N. Kwon, S. W. Shulman, and E. Hovy. Multidimensional text analysis for eRulemaking. In *Proceedings of the 7th Annual International Conference on Digital Government Research*, pages 157–166. Digital Government Research Center, 2006. URL https://dl.acm.org/doi/abs/10.1145/1146598.1146649.

N. Kwon, L. Zhou, E. Hovy, and S. W. Shulman. Identifying and classifying subjective claims. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, pages 76–81. Digital Government Society of North America, 2007. URL https://dl.acm.org/doi/abs/10.5555/1248460.1248473.

P. D. König and G. Wenzelburger. Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly*, 37 (3):101489, 2020. URL https://www.sciencedirect.com/science/article/pii/S0740624X1930245X.

J. Lawrence and C. Reed. Argument mining: A survey. *Computational Linguistics*, 45 (4):765–818, 2019. URL `https://aclanthology.org/J19-4006`.

J. Lawrence, J. Park, K. Budzynska, C. Cardie, B. Konat, and C. Reed. Using argumentative structure to interpret debates in online deliberative democracy and eRulemaking. *ACM Transactions on Internet Technology*, 17(3):1–22, 2017. URL `https://dl.acm.org/doi/abs/10.1145/3032989`.

Y. Lei and R. Huang. Few-shot (dis)agreement identification in online discussions with regularized and augmented meta-learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5581–5593. Association for Computational Linguistics, 2022. URL `https://aclanthology.org/2022.findings-emnlp.409`.

R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim. Context dependent claim detection. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500. Dublin City University and Association for Computational Linguistics, 2014. URL `https://aclanthology.org/C14-1141`.

M. Liebeck. *Automated Discussion Analysis in Online Participation Projects*. PhD thesis, Heinrich Heine University Düsseldorf, 2017. URL `https://docserv.uni-duesseldorf.de/servlets/DerivateServlet/Derivate-51087/liebeck_thesis.pdf`.

M. Liebeck, K. Esau, and S. Conrad. 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*, pages 144–153. Association for Computational Linguistics, 2016. URL `https://aclanthology.org/W16-2817`.

Z. Liu, M. Guo, Y. Dai, and D. Litman. ImageArg: A multi-modal tweet dataset for image persuasiveness mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 1–18. International Conference on Computational Linguistics, 2022. URL `https://aclanthology.org/2022.argmining-1.1`.

M. A. Livermore, V. Eidelman, and B. Grom. Computationally assisted regulatory participation. *Notre Dame L. Rev.*, 93:977, 2017.

S. Longpre, J. Reisler, E. G. Huang, Y. Lu, A. Frank, N. Ramesh, and C. DuBois. Active learning over multiple domains in natural language tasks. *arXiv preprint arXiv:2202.00254*, 2022. URL `https://arxiv.org/abs/2202.00254`.

D. Lowell, Z. C. Lipton, and B. C. Wallace. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 21–30, 2019. URL `https://aclanthology.org/D19-1003`.

C. Lucas, R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley. Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2): 254–277, 2015. URL `https://doi.org/10.1093/pan/mpu019`.

N. Mahyar, D. V. Nguyen, M. Chan, J. Zheng, and S. P. Dow. The civic data deluge: Understanding the challenges of analyzing large-scale community input. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 1171–1181. Association for Computing Machinery, 2019. URL https://dl.acm.org/doi/10.1145/3322276.3322354.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

M. Maragoudakis, E. Loukis, and Y. Charalabidis. A review of opinion mining methods for analyzing citizens' contributions in public policy debate. In *Electronic Participation*, pages 298–313. Springer, 2011. URL https://link.springer.com/chapter/10.1007/978-3-642-23333-3_26.

K. Margatina and N. Aletras. On the limitations of simulating active learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4402–4419. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.findings-acl.269.

K. Margatina, G. Vernikos, L. Barrault, and N. Aletras. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.emnlp-main.51.

K. Margatina, L. Barrault, and N. Aletras. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.acl-short.93.

S. Marien, M. Hooghe, and E. Quintelier. Inequalities in non-institutionalised forms of political participation: A multi-level analysis of 25 countries. *Political Studies*, 58 (1):187–213, 2010. URL https://doi.org/10.1111/j.1467-9248.2009.00801.x.

C. Masdeval and A. Veloso. Mining citizen emotions to estimate the urgency of urban issues. *Information Systems*, 54:147–155, 2015. URL https://www.sciencedirect.com/science/article/pii/S030643791500126X.

R. Mestre, R. Milicin, S. E. Middleton, M. Ryan, J. Zhu, and T. J. Norman. M-Arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, 2021. URL https://aclanthology.org/2021.argmining-1.8.

R. Mestre, S. E. Middleton, M. Ryan, M. Gheasi, T. Norman, and J. Zhu. Augmenting pre-trained language models with audio feature embedding for argumentation mining in political debates. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 274–288, 2023. URL https://aclanthology.org/2023.findings-eacl.21.

S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys*, 54(3), 2021. URL `https://dl.acm.org/doi/abs/10.1145/3439726`.

R. Mochales and M.-F. Moens. Argumentation mining. *Artificial Intelligence and Law*, 19:1–22, 2011. URL `https://doi.org/10.1007/s10506-010-9104-x`.

M.-F. Moens, E. Boiy, R. M. Palau, and C. Reed. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pages 225–230. Association for Computing Machinery, 2007. URL `https://dl.acm.org/doi/10.1145/1276318.1276362`.

G. Morio and K. Fujita. Annotating online civic discussion threads for argument mining. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 546–553. IEEE, 2018a. URL `https://ieeexplore.ieee.org/abstract/document/8609644`.

G. Morio and K. Fujita. 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. Association for Computational Linguistics, 2018b. URL `https://aclanthology.org/W18-5202`.

G. Morio, H. Ozaki, T. Morishita, and K. Yanai. End-to-end argument mining with cross-corpora multi-task learning. *Transactions of the Association for Computational Linguistics*, 10:639–658, 2022. URL `https://aclanthology.org/2022.tacl-1.37`.

V. Niculae, J. Park, and C. Cardie. 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. Association for Computational Linguistics, 2017. URL `https://aclanthology.org/P17-1091`.

Organisation for Economic Co-operation and Development. *Promise and Problems of E-Democracy.* OECD Publishing, 2003. URL `https://www.oecd-ilibrary.org/governance/promise-and-problems-of-e-democracy_9789264019492-en`.

S. Padjman. Unterstützung der Auswertung von verkehrsbezogenen Bürger*innenbeteiligungsverfahren durch die automatisierte Erkennung von Verortungen, 2021. Project report, Heinrich Heine University Düsseldorf.

S. Padjman. Mining argument components in public participation processes. Master's thesis, Heinrich Heine University Düsseldorf, 2022.

R. M. Palau and M.-F. Moens. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107. Association for Computing Machinery, 2009. URL `https://doi.org/10.1145/1568234.1568246`.

J. Park and C. Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38. Association for Computational Linguistics, 2014. URL `https://aclanthology.org/W14-2105`.

J. Park and C. Cardie. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association, 2018. URL https://aclanthology.org/L18-1257.

J. Park, C. Blake, and C. Cardie. Toward machine-assisted participation in eRulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 206–210. Association for Computing Machinery, 2015. URL https://dl.acm.org/doi/abs/10.1145/2746090.2746118.

M. Passon, M. Lippi, G. Serra, and C. Tasso. Predicting the usefulness of Amazon reviews using off-the-shelf argumentation mining. In *Proceedings of the 5th Workshop on Argument Mining*, pages 35–39. Association for Computational Linguistics, 2018. URL https://aclanthology.org/W18-5205.

A. Peldszus and M. Stede. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7:1–31, 2013. URL http://doi.org/10.4018/jcini.2013010101.

M. Poel, E. T. Meyer, and R. Schroeder. Big data for policymaking: Great expectations, but with limited progress? *Policy & Internet*, 10(3):347–367, 2018. URL https://onlinelibrary.wiley.com/doi/10.1002/poi3.176.

S. Prabhu, M. Mohamed, and H. Misra. Multi-class text classification using BERT-based active learning. *arXiv preprint arXiv:2104.14289*, 2021. URL https://arxiv.org/abs/2104.14289.

J. D. Priscoli. What is public participation in water resources management and why is it important? *Water International*, 29(2):221–227, 2004. URL https://www.tandfonline.com/doi/abs/10.1080/02508060408691771.

S. Purpura, C. Cardie, and J. Simons. Active learning for e-Rulemaking: Public comment categorization. In *Proceedings of the 9th Annual International Conference on Digital Government Research*, pages 234–243. Digital Government Research Center, 2008. URL https://scholarship.law.cornell.edu/ceri/7/.

X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, 2020. URL https://link.springer.com/article/10.1007/s11431-020-1647-3.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

P. Rajendran, D. Bollegala, and S. Parsons. Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews. In *Proceedings of the Third Workshop on Argument Mining*, pages 31–39. Association for Computational Linguistics, 2016. URL https://aclanthology.org/W16-2804.

B. Reynante, S. P. Dow, and N. Mahyar. A framework for open civic design: Integrating public participation, crowdsourcing, and design thinking. *Digital Government: Research and Practice*, 2(4), 2021. URL https://dl.acm.org/doi/10.1145/3487607.

R. Rinott, L. Dankin, C. Alzate Perez, M. M. Khapra, E. Aharoni, and N. Slonim. 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. Association for Computational Linguistics, 2015. URL https://aclanthology.org/D15-1050.

M. E. Roberts, B. M. Stewart, D. Tingley, and E. M. Airoldi. The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, pages 1–20. Curran Associates, Inc., 2013. URL https://scholar.harvard.edu/sites/scholar.harvard.edu/files/dtingley/files/stmnips2013.pdf.

J. Romberg and T. Escher. Analyse der Anforderungen an eine Software zur (teil-) automatisierten Unterstützung bei der Auswertung von Beteiligungsverfahren. Working Paper 1, CIMT Research Group, Heinrich Heine University Düsseldorf, 2020. URL https://www.cimt-hhu.de/wp-content/uploads/2020/12/cimt_working_paper1.pdf.

G. Rotman and R. Reichart. Multi-task active learning for pre-trained transformer-based models. *Transactions of the Association for Computational Linguistics*, 10: 1209–1228, 2022. URL https://aclanthology.org/2022.tacl-1.70.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

E. Saveleva, V. Petukhova, M. Mosbach, and D. Klakow. Graph-based argument quality assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 1268–1280. INCOMA Ltd., 2021. URL https://aclanthology.org/2021.ranlp-1.143.

R. Schaefer and M. Stede. Argument mining on twitter: A survey. *it – Information Technology*, 63(1):45–58, 2021. URL https://doi.org/10.1515/itit-2020-0053.

V. A. Schmidt. Democracy and legitimacy in the European Union revisited: Input, output and 'throughput'. *Political Studies*, 61(1):2–22, 2013. URL https://journals.sagepub.com/doi/pdf/10.1111/j.1467-9248.2012.00962.x.

C. Schröder, K. Bürgl, Y. Annanias, A. Niekler, L. Müller, D. Wiegreffe, C. Bender, C. Mengs, G. Scheuermann, and G. Heyer. Supporting land reuse of former open pit mining sites using text classification and active learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4141–4152. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.acl-long.320.

C. Schröder, A. Niekler, and M. Potthast. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.findings-acl.172.

C. Schulz, S. Eger, J. Daxenberger, T. Kahse, and I. Gurevych. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41. Association for Computational Linguistics, 2018. URL https://aclanthology.org/N18-2006.

T. Searle, Z. Kraljevic, R. Bendayan, D. Bean, and R. Dobson. MedCATTrainer: A biomedical free text annotation interface with active learning and research use case specific customisation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 139–144. Association for Computational Linguistics, 2019. URL https://aclanthology.org/D19-3024.

B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, Department of Computer Sciences, 2009. URL https://minds.wisconsin.edu/handle/1793/60660.

B. Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, 2011. URL https://aclanthology.org/D11-1136.

S. Shulman, J. Callan, E. Hovy, and S. Zavestoski. SGER collaborative. *Journal of E-Government*, 1(1):123–127, 2004. URL https://www.tandfonline.com/doi/abs/10.1300/J399v01n01_09.

S. Shulman, E. Hovy, J. Callan, and S. Zavestoski. Language processing technologies for electronic rulemaking: A project highlight. In *Proceedings of the 2005 National Conference on Digital Government Research*, pages 87–88. Digital Government Society of North America, 2005. URL https://dl.acm.org/doi/10.5555/1065226.1065248.

S. W. Shulman. An experiment in digital government at the United States national organic program. *Agriculture and Human Values*, 20(3):253–265, 2003. URL https://link.springer.com/article/10.1023/A:1026104815057.

S. W. Shulman. The case against mass e-mails: Perverse incentives and low quality public participation in US federal rulemaking. *Policy & Internet*, 1(1):23–53, 2009. URL https://onlinelibrary.wiley.com/doi/abs/10.2202/1944-2866.1010.

A. Simonofski, J. Fink, and C. Burnay. Supporting policy-making with social media and e-participation platforms data: A policy analytics framework. *Government Information Quarterly*, 38(3):101590, 2021. URL https://www.sciencedirect.com/science/article/abs/pii/S0740624X21000265.

N. Slonim, Y. Bilu, C. Alzate, R. Bar-Haim, B. Bogin, F. Bonin, L. Choshen, E. Cohen-Karlik, L. Dankin, L. Edelstein, et al. An autonomous debating system. *Nature*, 591 (7850):379–384, 2021. URL `https://doi.org/10.1038/s41586-021-03215-w`.

A. Snider and M. Schnurer. *Many Sides: Debate Across the Curriculum.* International Debate Education Association, 2002.

A. Snijders, D. Kiela, and K. Margatina. Investigating multi-source active learning for natural language inference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2187–2209. Association for Computational Linguistics, 2023. URL `https://aclanthology.org/2023.eacl-main.160`.

K. Soulis, I. Varlamis, A. Giannakoulopoulos, and F. Charatsev. A tool for the visualisation of public opinion. *International Journal of Electronic Governance*, 6(3): 218–231, 2013. URL `https://www.inderscienceonline.com/doi/abs/10.1504/IJEG.2013.058404`.

C. Stab and I. Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 46–56. Association for Computational Linguistics, 2014. URL `https://aclanthology.org/D14-1006`.

C. Stab and I. Gurevych. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining*, pages 113–118. Association for Computational Linguistics, 2016. URL `https://aclanthology.org/W16-2813`.

C. Stab and I. Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, 2017. URL `https://aclanthology.org/J17-3005`.

C. Stab, T. Miller, B. Schiller, P. Rai, and I. Gurevych. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674. Association for Computational Linguistics, 2018. URL `https://aclanthology.org/D18-1402`.

C. Starke, J. Baleis, B. Keller, and F. Marcinkowski. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2), 2022. URL `https://doi.org/10.1177/20539517221115189`.

M. Stede and J. Schneider. *Argumentation Mining.* Synthesis Lectures on Human Language Technologies. Morgan and Claypool, 2018. URL `https://doi.org/10.2200/S00883ED1V01Y201811HLT040`.

M. A. Strebel, D. Kübler, and F. Marcinkowski. The importance of input and output legitimacy in democratic governance: Evidence from a population-based survey experiment in four west european countries. *European Journal of Political Research*, 58 (2):488–513, 2019. URL `https://ejpr.onlinelibrary.wiley.com/doi/10.1111/1475-6765.12293`.

E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650. Association for Computational Linguistics, 2019. URL https://aclanthology.org/P19-1355.

A. Suominen and A. Hajikhani. Research themes in big data analytics for policymaking: Insights from a mixed-methods systematic literature review. *Policy & Internet*, 13 (4):464–484, 2021. URL https://onlinelibrary.wiley.com/doi/full/10.1002/poi3.258.

A. Tamkin, D. Nguyen, S. Deshpande, J. Mu, and N. Goodman. Active learning helps pretrained models learn the intended task. In *Advances in Neural Information Processing Systems*, volume 35, pages 28140–28153. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b43a0e8a35b1c044b18cd843b9771915-Paper-Conference.pdf.

P. Teufl, U. Payer, and P. Parycek. Automated analysis of e-Participation data by utilizing associative networks, spreading activation and unsupervised learning. In *Electronic Participation*, pages 139–150. Springer, 2009. URL https://link.springer.com/chapter/10.1007/978-3-642-03781-8_13.

Y. Theocharis and J. W. Van Deth. The continuous expansion of citizen participation: A new taxonomy. *European Political Science Review*, 10(1):139–163, 2018. URL https://doi.org/10.1017/S1755773916000230.

B. Thome. Topic classification of citizen comments using active learning. Master's thesis, Heinrich Heine University Düsseldorf, 2022.

A. Toledo, S. Gretz, E. Cohen-Karlik, R. Friedman, E. Venezian, D. Lahav, M. Jacovi, R. Aharonov, and N. Slonim. 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*, pages 5625–5635. Association for Computational Linguistics, 2019. URL https://aclanthology.org/D19-1564.

K. Tomanek and U. Hahn. A comparison of models for cost-sensitive active learning. In *Coling 2010: Posters*, pages 1247–1255. Coling 2010 Organizing Committee, 2010. URL https://aclanthology.org/C10-2143.

K. Tomanek and K. Morik. Inspecting sample reusability for active learning. In *Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010*, Proceedings of Machine Learning Research, pages 169–181. PMLR, 2011. URL https://proceedings.mlr.press/v16/tomanek11a.html.

L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*, 2022. URL https://arxiv.org/abs/2209.11055.

S. Verba, K. L. Schlozman, and H. E. Brady. *Voice and Equality: Civic Voluntarism in American Politics*. Harvard University Press, 1995.

H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, and B. Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics, 2017a. URL `https://aclanthology.org/E17-1017`.

H. Wachsmuth, B. Stein, and Y. Ajjour. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127. Association for Computational Linguistics, 2017b. URL `https://aclanthology.org/E17-1105`.

M. Wang and M. Liu. An empirical study on active learning for multi-label text classification. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 94–102. Association for Computational Linguistics, 2023. URL `https://aclanthology.org/2023.insights-1.12`.

L. Wertz, K. Mirylenka, J. Kuhn, and J. Bogojeska. Investigating active learning sampling strategies for extreme multi label text classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4597–4605, 2022. URL `https://aclanthology.org/2022.lrec-1.490`.

B. W. Wirtz, J. C. Weyerer, and C. Geyer. Artificial intelligence and the public sector - applications and challenges. *International Journal of Public Administration*, 42(7):596–615, 2019. URL `https://www.tandfonline.com/doi/abs/10.1080/01900692.2018.1498103`.

H. Yang and J. Callan. Near-duplicate detection for eRulemaking. In *Proceedings of the 2005 National Conference on Digital Government Research*, pages 78–86. Digital Government Research Center, 2005. URL `https://dl.acm.org/doi/10.5555/1065226.1065247`.

H. Yang and J. Callan. Near-duplicate detection by instance-level constrained clustering. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 421–428. Association for Computing Machinery, 2006. URL `https://dl.acm.org/doi/abs/10.1145/1148170.1148243`.

H. Yang and J. Callan. Ontocop: Constructing ontologies for public comments. *IEEE Intelligent Systems*, 24(5):70–75, 2009.

H. Yang, J. Callan, and S. Shulman. Next steps in near-duplicate detection for eRulemaking. In *Proceedings of the 7th Annual International Conference on Digital Government Research*, pages 239–248. Digital Government Research Center, 2006. URL `https://dl.acm.org/doi/abs/10.1145/1146598.1146663`.

M. Yuan, H.-T. Lin, and J. Boyd-Graber. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7948. Association for Computational Linguistics, 2020. URL `https://aclanthology.org/2020.emnlp-main.637`.

J. Zhang, R. Kumar, S. Ravi, and C. Danescu-Niculescu-Mizil. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141. Association for Computational Linguistics, 2016. URL `https://aclanthology.org/N16-1017`.

S. Zhang, C. Gong, X. Liu, P. He, W. Chen, and M. Zhou. ALLSH: Active learning guided by local sensitivity and hardness. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1328–1342. Association for Computational Linguistics, 2022a. URL `https://aclanthology.org/2022.findings-naacl.99`.

Z. Zhang, E. Strubell, and E. Hovy. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190. Association for Computational Linguistics, 2022b. URL `https://aclanthology.org/2022.emnlp-main.414`.

Z.-H. Zhou. *Machine Learning*. Springer, 2021.

A. Zuiderwijk, Y.-C. Chen, and F. Salem. Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, 38(3):101577, 2021. URL `https://www.sciencedirect.com/science/article/pii/S0740624X21000137`.