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 and TOBIAS ESCHER, Department of Social Sciences, Heinrich Heine University Düsseldorf, Germany

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 corpuses 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 modeling.

CCS Concepts: • **Computing methodologies** → **Natural language processing** • **Applied computing** → **Computing in government**;

Additional Key Words and Phrases: Policy analytics, citizen participation, computational linguistics

ACM Reference format:
Julia Romberg and Tobias Escher. 2023. 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. *Digit. Gov. Res. Pract.* 00, JA, Article 00 (June 2023), 30 pages.
https://doi.org/10.1145/3603254

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

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.
Author's address: J. Romberg (corresponding author) and T. Escher, Department of Social Sciences, Heinrich Heine University Düsseldorf, Universitaetsstrasse 1, 40225 Duesseldorf, Germany.

2639-0175/2023/06-ART00
https://doi.org/10.1145/3603254

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 article 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 article 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 article 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 judgment. 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 articles 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 articles 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. Articles 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, articles 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 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 articles in our list of relevant documents. This strategy resulted in 28 additional studies.

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.

[1] https://aclanthology.org/.
[2] See Appendix F for the search terms that were employed.

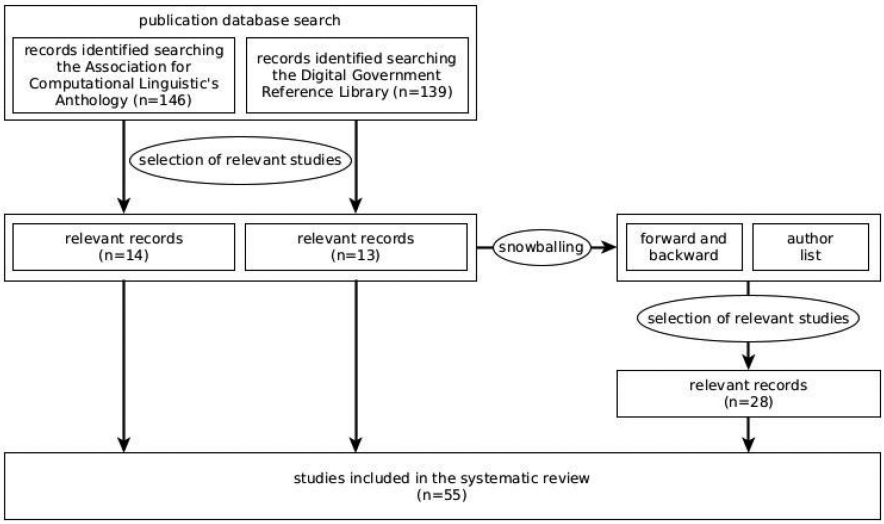

Fig. 1. Study identification and selection process.

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.

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

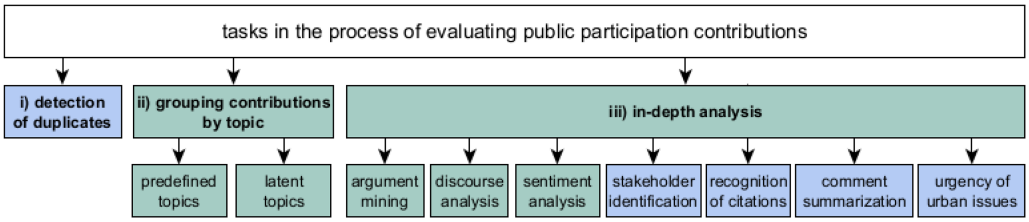

Fig. 2. Overview of tasks in the evaluation of public participation contributions.

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 DUplicate Removal In lArge collectioNs (DURIAN) 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.

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,

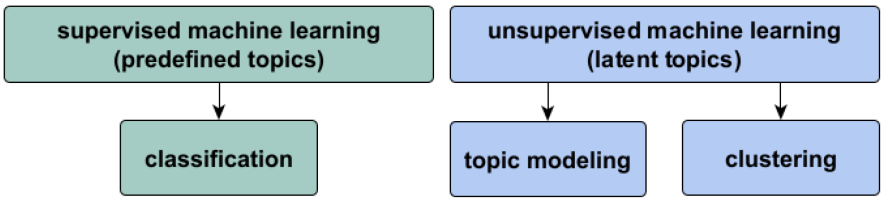

Fig. 3. Approaches to grouping the data collection by topic.

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 article, 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 labeled 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 judgment 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 judgment. (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.

6.1.1 Segmentation. In the segmentation step, citizen contributions are divided into units of argumentative content.[3] All articles 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]

[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.

and parallel constrained pointer architectures (PCPA) [60] lead to the best but not yet sufficient results in classifying argumentative sentences on the respective datasets.

6.1.2 Segment 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 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.

6.1.3 Relation 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 articles 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 articles, 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 judgment 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 articles 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 articles 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

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

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 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 modeling. 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.

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)
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

(Continued)

[6]For the complete list, please refer to the article.
[7]The authors state that NB and CRF were also evaluated. However, the results are not further reported.
[8]We could not find the total number of topics or an overview of all subtopics in the article.
[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.

Table 1. Continued

Dataset	Language	Classification granularity	Categorization schema	Research article	Input features	Algorithms
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
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
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

(Continued)

[10] [68] systematize argument mining as three consecutive subtasks: (1) segmentation, (2) segment classification, (3) relation identification.

Table 3. Continued

Dataset	Language	Argument mining schema	Research article	subtask	Input features	Algorithms
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 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".

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.

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. *Digital Government: Research and Practice* 2, 3 (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, 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. *Journal of Information Technology & Politics* 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 *Proceedings of the 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. *Government Information Quarterly* 36, 3 (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.* 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. Cornell e-Rulemaking Initiative Publications. Paper 5. http://scholarship.law.cornell.edu/ceri/5.

[10] Guoray Cai, Feng Sun, and Yongzhong Sha. 2018. Interactive visualization for topic model curation. *IUI'2018: International Conference on Intelligent User Interfaces, Workshop Exploratory Search and Interactive Data Analytics (ESIDA)*, Tokyo. http://personal.psu.edu/gxc26/Pubs/2018-topic%20model%20curation.pdf.

[11] Iván Cantador, María E. Cortés-Cediel, and Miriam Fernández. 2020. Exploiting open data to analyze discussion and controversy in online citizen participation. *Information Processing & Management* 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 *Proceedings of the 9th Annual International Conference on Digital Government Research*. 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 6th International Conference on Language Resources and Evaluation*. 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*. 401–406.

[15] Yannis Charalabidis, Euripidis N. Loukis, Aggeliki Androutsopoulou, Vangelis Karkaletsis, and Anna Triantafillou. 2014. Passive crowdsourcing in government using social media. *Transforming Government: People, Process and Policy* 8, 2 (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. *International Public Management Journal* 22, 1 (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 Computing Surveys* 54, 10s (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. *Annals of Operations Research* 236, 1 (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 17th International Conference on Artificial Intelligence and Law*. Association for Computing Machinery, New York, NY, 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, 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, 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, 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 *Proceedings of the 2019 6th Swiss Conference on Data Science*. 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 (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. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (2018), 596–599. 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. *Political Analysis* 21, 3 (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 *Proceedings of the Coling 2016, The 26th International Conference On Computational Linguistics: Technical Papers*. 2740–2751.

[34] Loni Hagen. 2018. Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management* 54, 6 (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. *Government Information Quarterly* 36, 4 (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 *Proceedings*

of the 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 *Proceedings of the Designing Interactive Systems Conference 2021*, ACM, New York, NY, 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, (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)* 7444 LNCS, S. Online verfügbar unter, 132–143. http://link.springer.com/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 10th International Conference on Language Resources and Evaluation*. 3899–3906.

[42] Pascal D. König and Georg Wenzelburger. 2020. Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly* 37, 3 (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*. ACM, New York, New York, 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*. ACM, New York, New York, 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*. 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 Transactions on Internet Technology* 17, 3 (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 (2014), 182–194. DOI: https://doi.org/10.1177/0894439313506847

[50] Matthias Liebeck. 2017. *Automated Discussion Analysis in Online Participation Projects*. (Doctoral dissertation, Heinrich-Heine-Universität Düsseldorf, Med.-Fak.)

[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 3rd Workshop on Argument Mining*. 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. *Information Processing & Management* 52, 3 (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, 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*. 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. *Annals of Operations Research* 236, 1 (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. *Information Systems* 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 *Proceedings of the 2018 IEEE/WIC/ACM International Conference on Web Intelligence*. 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, 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, 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, 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 1st 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 11th International Conference on Language Resources and Evaluation*.

[66] Joonsuk Park and Claire Cardie. 2018. A Corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*. 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 (2013), 1–31. 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 (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 *Proceedings of the 9th Annual International Conference on Digital Government Research*. 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. *Digital Government: Research and Practice* 2, 4 (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. *Electronic Government: 21st IFIP WG 8.5 International Conference, EGOV 2022, Linköping, Sweden, September 6–8, 2022, Proceedings*. 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 13th 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. *Agriculture and 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 (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. *Government Information Quarterly* 38, 3 (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 (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, 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. *International Journal of Electronic Governance* 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 (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. *European Journal of Political Research* 58, 2 (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* 13, 4 (2021), 464–484. 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 *Proceedings of the International Conference on Electronic Participation*. Ann Macintosh and Efthimios Tambouris (Eds.). Springer, Berlin, 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 Annua International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, New York, 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, 1 (2019), 1–5. 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* (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, 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. *International Journal of Public Administration* 42, 7 (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*. ACM, New York, New York, 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*. ACM, New York, New York, 421. DOI: https://doi.org/10.1145/1148170.1148243

[99] Hui Yang and Jamie Callan. 2009. OntoCop: Constructing ontologies for public comments. *IEEE Intelligent Systems* 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 *Proceedings of the Advances in Neural Information Processing*. 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 Intelligent Systems* 30, 6 (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. *Government Information Quarterly* 38, 3 (2021), 101577. DOI: https://doi.org/10.1016/j.giq.2021.101577

Received 9 September 2022; revised 8 March 2023; accepted 3 May 2023