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

Julia Romberg(✉) and Tobias Escher

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

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

Keywords: Topic classification · Public participation · Active learning · Natural language processing

1 Introduction

Democratic authorities are regularly using public participation to consult and involve citizens in order to inform political decisions and increase public support

The original version of this chapter was revised: this chapter was previously published non-open access. The correction to this chapter is available at https://doi.org/10.1007/978-3-031-15086-9_37

M. Janssen et al. (Eds.): EGOV 2022, LNCS 13391, pp. 369–385, 2022.
https://doi.org/10.1007/978-3-031-15086-9_24

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

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

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

how time-efficient the different categorisation approaches are to assess whether these could be used in realistic scenarios (RQ3).

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

2 Approaches to Thematically Structure Contributions

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

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

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

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

consider more appropriate to support categorising contributions thematically [1,4,6,13,14].

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

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

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

Since the study of Purpura et al., the research on NLP and on active learning has evolved. Our goal is to apply state-of-the-art methods to citizen contributions

and to evaluate to what extent the advanced methods can further reduce the amount of training data needed. In addition, we also assess the runtime of these models as another potential barrier for practical application.

3 The Cycling Dialogues Datasets

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

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

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

The coded dataset comprises a total of 3,139 contributions. *Cycling Dialogue Bonn* has received the most contributions with 2,314, whereas *Cycling Dialogue Ehrenfeld* and *Cycling Dialogue Moers* account for 366 and 459 unique contributions respectively. The contributions contain an average of 4.83 (Bonn), 4.66 (Ehrenfeld) and 4.78 (Moers) sentences. Table 1 gives insights into the thematic priorities within the eight categories. Cycling traffic management and cycle path

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

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

quality attracted the most interest in all datasets, followed by either obstacles or traffic lights. The (larger) differences in the amount of contributions as well as the (smaller) difference in the distribution of categories can be attributed to both contextual factors such as city size or local infrastructure, and individual-level factors such as the participating stakeholders.

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

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

4 Methodology

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

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

4.1 Active Learning

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

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

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

4.2 Query Strategies

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

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

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

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

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

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

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

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

4.3 Classifier

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

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

5 Evaluation

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

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

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

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

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

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

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

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

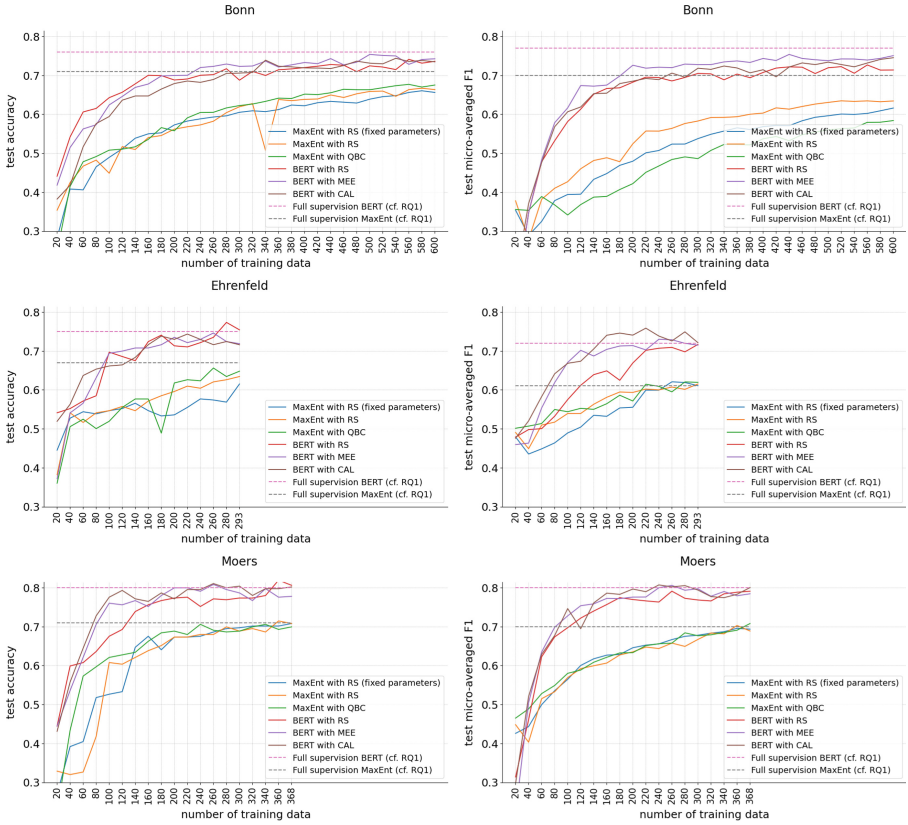
cations. This suggests that predicting all associated labels of a contribution is by no means more difficult than the recognition of the primary topic.

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

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

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

In our experiments, the active learning process (implemented using the small-text library [21]) is initialised with 20 randomly drawn samples (i.e. contribu-



(a) Single-Label Classification (b) Multi-Label Classification

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

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

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

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

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

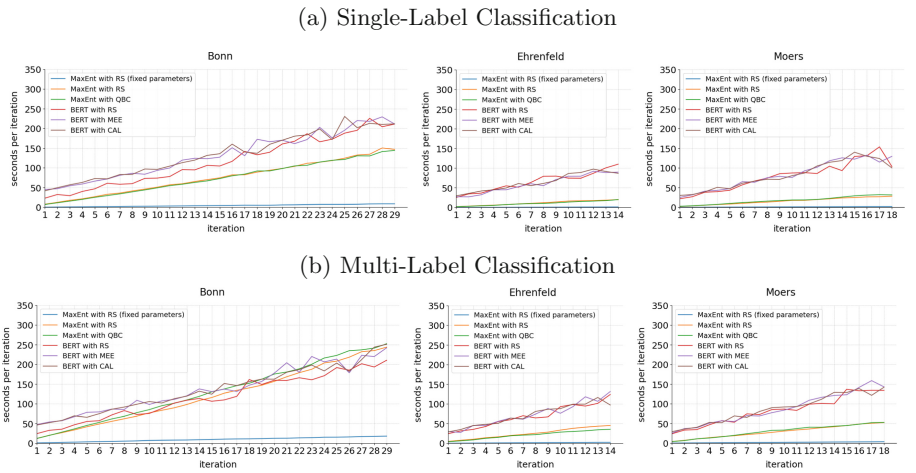


Fig. 2. Time duration of active learning iterations in seconds.

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

Encouragingly in terms of applicability, no iteration in the observation interval lasts longer than five minutes. Taking into account the findings from RQ2, to achieve these results on average a human analyst would have to wait less than three minutes (Bonn) or even less than one minute (Ehrenfeld, Moers) between

the coding sessions. At the same time, however, we can observe that BERT is more computationally intensive than MaxEnt, even though we severely limited hyperparameter tuning in our experiments.

6 Discussion

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

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

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

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

However, this would only offer a useful support for practice if these models can be realistically computed in common administrative settings. Therefore we also investigated the time-efficiency of the different categorisation approaches (RQ3). As it turns out, all of these require only a few minutes per iteration to compute. However, it should be noted that these time benchmarks depend on specific hardware (e.g. GPU and processor). The implications for practical use will need to be investigated in future work.

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

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

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

7 Conclusion and Future Work

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

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

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

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

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

Appendix A: Hyperparameter Tuning

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

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

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

References

1. Aitamurto, T., Chen, K., Cherif, A., Galli, J.S., Santana, L.: Civic CrowdAnalytics: making sense of crowdsourced civic input with big data tools. In: Proceedings of the 20th International Academic Mindtrek Conference, AcademicMindtrek 2016, pp. 86–94. Association for Computing Machinery, New York (2016)
2. Arana-Catania, M., et al.: Citizen participation and machine learning for a better democracy. Digit. Gov. Res. Pract. **2**(3), 1–22 (2021)
3. Ash, J.T., Chicheng, Z., Akshay, K., John, L., Alekh, A.: Deep batch active learning by diverse, uncertain gradient lower BoundsDeep batch active learning by diverse, uncertain gradient lower bounds. In: International Conference on Learning Representations 2020 (ICLR 2020) (2020)
4. Balta, D., Kuhn, P., Sellami, M., Kulus, D., Lieven, C., Krcmar, H.: How to streamline AI application in government? A case study on citizen participation in Germany. In: Lindgren, I., et al. (eds.) EGOV 2019. LNCS, vol. 11685, pp. 233–247. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27325-5_18
5. Cai, G., Sun, F., Sha, Y.: Interactive visualization for topic model curation. In: Proceedings of the ACM IUI 2018 Workshop on Exploratory Search and Interactive Data Analytics (2018)
6. Cardie, C., Farina, C., Aijaz, A., Rawding, M., Purpura, S.: A study in rule-specific issue categorization for e-rulemaking. In: Proceedings of the 9th International Conference on Digital Government Research, pp. 244–253 (2008)
7. Chen, K., Aitamurto, T.: Barriers for crowd's impact in crowdsourced policymaking: civic data overload and filter hierarchy. Int. Public Manag. J. **22**(1), 99–126 (2019)
8. Dryzek, J.S., et al.: The crisis of democracy and the science of deliberation. Science **363**(6432), 1144–1146 (2019)
9. Ein-Dor, L., et al.: Active learning for BERT: an empirical study. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7949–7962. Association for Computational Linguistics (2020)
10. Hagen, L.: Content analysis of e-petitions with topic modeling: how to train and evaluate LDA models? Inf. Process. Manag. **54**(6), 1292–1307 (2018)
11. Hagen, L., Uzuner, Ö., Kotfila, C., Harrison, T.M., Lamanna, D.: Understanding citizens' direct policy suggestions to the federal government: a natural language processing and topic modeling approach. In: 2015 48th Hawaii International Conference on System Sciences, pp. 2134–2143 (2015)
12. Holub, A., Perona, P., Burl, M.C.: Entropy-based active learning for object recognition. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8 (2008)
13. Kim, B., Yoo, M., Park, K.C., Lee, K.R., Kim, J.H.: A value of civic voices for smart city: a big data analysis of civic queries posed by Seoul citizens. Cities **108**, 102941 (2021)

14. Kwon, N., Shulman, S.W., Hovy, E.: Multidimensional text analysis for eRulemaking. In: Proceedings of the 2006 International Conference on Digital Government Research, dg.o 2006, pp. 157–166. Digital Government Society of North America (2006)
15. Levy, K.E.C., Franklin, M.: Driving regulation: using topic models to examine political contention in the U.S. trucking industry. Soc. Sci. Comput. Rev. **32**(2), 182–194 (2014)
16. Ma, B., Zhang, N., Liu, G., Li, L., Yuan, H.: Semantic search for public opinions on urban affairs: a probabilistic topic modeling-based approach. Inf. Process. Manag. **52**(3), 430–445 (2016)
17. Margatina, K., Vernikos, G., Barrault, L., Aletras, N.: Active learning by acquiring contrastive examples. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 650–663. Association for Computational Linguistics, Punta Cana (2021)
18. OECD: Promise and Problems of E-Democracy. OECD (2003)
19. Parry, G., Moyser, G.: More participation, more democracy? In: Beetham, D. (ed.) Defining and Measuring Democracy. Sage, London (1994)
20. Purpura, S., Cardie, C., Simons, J.: Active learning for e-rulemaking: public comment categorization. In: Proceedings of the 9th International Conference on Digital Government Research, pp. 234–243 (2008)
21. Schröder, C., Müller, L., Niekler, A., Potthast, M.: Small-text: active learning for text classification in Python (2021)
22. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 287–294 (1992)
23. Simonofski, A., Fink, J., Burnay, C.: Supporting policy-making with social media and e-participation platforms data: a policy analytics framework. Gov. Inf. Q. **38**(3), 101590 (2021)
24. Teufl, P., Payer, U., Parycek, P.: Automated analysis of e-participation data by utilizing associative networks, spreading activation and unsupervised learning. In: Macintosh, A., Tambouris, E. (eds.) ePart 2009. LNCS, vol. 5694, pp. 139–150. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-03781-8_13
25. Yang, H., Callan, J.: OntoCop: constructing ontologies for public comments. IEEE Intell. Syst. **24**(5), 70–75 (2009)