Political Science Research and Methods (2025), 1–18
doi:10.1017/psrm.2025.10013

ORIGINAL ARTICLE

PAPEA: A modular pipeline for the automation of protest event analysis

Sebastian Haunss[1], Priska Daphi[2], Jan Matti Dollbaum[3,4], Lidiya Hristova[1], Pál Susánszky[1] and Elias Steinhilper[5]

[1]Research Institute Social Cohesion, University of Bremen, Bremen, Germany; [2]Institute of Sociology, Justus-Liebig University, Giessen, Germany; [3]Department of European Studies and Slavic Studies, Université de Fribourg, Switzerland; [4]Geschwister-Scholl-Institute for Political Science, Ludwig Maximilian University of Munich, Munich, Germany and [5]Consensus and Conflict Department, German Centre for Integration and Migration Research (DeZIM), Berlin, Germany
Corresponding author: Sebastian Haunss; Email: sebastian.haunss@uni-bremen.de

(Received 3 April 2024; revised 3 October 2024; accepted 12 January 2025)

Abstract

Protest event analysis (PEA) is the core method to understand spatial patterns and temporal dynamics of protest. We show how Large Language Models (LLM) can be used to automate the classification of protest events and of political event data more broadly with levels of accuracy comparable to humans, while reducing necessary annotation time by several orders of magnitude. We propose a modular pipeline for the automation of PEA (PAPEA) based on fine-tuned LLMs and provide publicly available models and tools which can be easily adapted and extended. PAPEA enables getting from newspaper articles to PEA datasets with high levels of precision without human intervention. A use case based on a large German news-corpus illustrates the potential of PAPEA.

Keywords: computational social sciences; large language models; natural language processing; protest event analysis; protest forms; social movements

1. Introduction

Protest is considered a key indicator of political conflict and constitutes a widely used expression of unconventional political participation (Della Porta and Diani, 2006). Given its importance to understand social and political dynamics, a "science to study protest" (Fisher *et al.*, 2019) has emerged in the last decades. While the field of social movement and protest research is characterized by methodological pluralism (della Porta, 2014), the analysis of protest events constitutes the core method to understand the temporal and spatial development of protests. It has been widely used for empirical analysis and theory building on social movements and contentious politics (Tilly, 1978; McAdam *et al.*, 2009; Kriesi *et al.*, 2012) including in recent episodes of contention (Shuman *et al.*, 2022).

Protest event data are usually compiled from newspaper reports or other textual sources, and so far relied on the time- and resource-intensive manual annotation of these sources according to a predefined codebook. As a result, only few larger protest event data sets exist, limiting the method's potential to be applied to the multiple and dynamically unfolding crises of our time.[1]

[1]These include the CCC (https://countingcrowds.org/), PolDem (https://poldem.eui.eu/the-observatory/), and GLOCON (https://glocon.ku.edu.tr/).

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

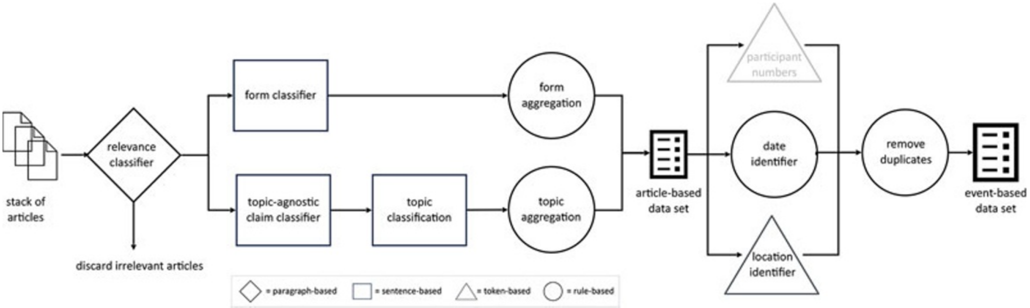

Figure 1. The complete pipeline.

With growing computing power and the development of advanced natural language processing (NLP) methods, several attempts have been made to automate the creation of protest event data (Hanna, 2017; Fisher *et al.*, 2019; Lorenzini *et al.*, 2022). Currently, the automated *identification* of protest events and political event data more broadly reaches levels of accuracy comparable to humans. But the automated *coding of event characteristics* (such as protest claims and forms of action) still remains a research frontier. Scholars have pointed to the weaknesses of existing fully automated event databases such as Global Data on Events Language and Tone (GDELT) (Hoffmann *et al.*, 2022) and concluded that the automated "collection of political event data has not accomplished a high degree of reliability or advanced beyond English language sources" (Fisher *et al.*, 2019, p. 3).

In this article, we present major advances regarding these shortcomings. We combine selected classifiers based on fine-tuned large language models (LLMs) and present a first fully automatic pipeline for the creation of protest event data sets that turns a set of German newspaper articles containing protest-related keywords into a reliable data set of protest events. We call this pipeline *PAPEA, Pipeline for the Automation of Protest Event Analysis*. All steps in the pipeline use publicly available models and tools. PAPEA thus can be easily adapted and extended. With PAPEA, we join ongoing efforts to automate PEA data collection and classification (Durușan *et al.*, 2022; Lorenzini *et al.*, 2022; Caren *et al.*, 2023; Oliver *et al.*, 2023) and provide a first fully documented and adaptable pipeline for this task that can reduce annotation time by several orders of magnitude.

In the following, we will first briefly present the development of PEA automation attempts and discuss the contributions and shortcomings of existing approaches. We will then present PAPEA's overall architecture and discuss each step in the pipeline in detail. Subsequently, we illustrate its potential on a large data set of local newspaper articles. We wrap up this article with a discussion of remaining challenges and necessary steps to extend PAPEA to automatically extract an even larger set of PEA variables and some thoughts on how to increase precision for infrequent categories.

The pipeline that we present fulfills a specific objective—the analysis of protest events. Yet, the individual steps of the pipeline depicted in Figure 1 correspond to much more general problems that are relevant for a broader set of text analysis tasks concerned with action forms and positions of political actors. An application outside protest event analysis would require different fine-tuned models, but the overall architecture of the pipeline could well be adapted. We will discuss this in more detail in Section 5.

2. From manual annotation to automation: Advances and current frontiers

Due to its prominence in the canon of social movement methodology, the genealogy of protest event research is well documented (Hutter, 2014; Hanna, 2017; Fisher *et al.*, 2019). Pioneering work was done by Charles Tilly, who engaged in standardized event coding from the 1960s (Fisher *et al.*, 2019).

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

Since then, countless projects have applied the method contributing significantly to theory building in social movement studies and political sociology more broadly. Prominent examples include the Dynamics of Collective Action (DCA) project (McAdam *et al.,* 2009) in the United States and the New Social Movements project, extending protest event analysis (PEA) to cross-country comparative analysis in Europe (Kriesi *et al.,* 1992). As a key indicator of political conflict, information on protest has also been collected as part of larger conflict event datasets, most prominently the Armed Conflict Location and Event Data Project (ACLED) (Raleigh, 2010). Yet, this dataset is limited to political violence and one specific form of protest—demonstrations—ignoring broader repertoires of protest (e.g. petitions, vigils, occupations, and civil disobedience). Furthermore, the data set has shortcomings with respect to its initial focus on Africa and concerns regarding data quality (Eck, 2012) as well as with respect to potential biases in regional coverage due to its source selection. This means ACLED cannot fully replace the more fine-grained protest event databases that currently are still manually annotated (e.g. PolDem; Kriesi *et al.,* 2020). Olsen and her collaborators provide an overview over many of the currently existing PEA and conflict datasets and their methodological underpinnings (Olsen *et al.,* 2024).

Until today, the bulk of research based on protest event data relies on the extraction of protest events from media reports in national newspapers, usually one per country. The focus on few (national) sources introduces a well-documented bias (Earl *et al.,* 2004), yet it was long indispensable given the labor-intensive data collection process. Fundamentally, protest event analysis combines "a discovery and coding problem" (Fisher *et al.,* 2019, p. 3), involving a "haystack task" (Hanna, 2017, p. 7) in which a small percentage of relevant articles must be selected from a large universe of reports, and a subsequent coding of key event variables. Illustrating the resources needed to exploit the full potential of PEA in long time series, Oliver et al. refer to the DCA project based on the manual reading and coding of protest events based on *New York Times* microfilm archives, which "involved four principal investigators and dozens of graduate students funded over a decade by a series of NSF grants at three institutions" (Oliver *et al.,* 2023, p. 2).

Against this background, the generation of event databases has become a dynamic context for the application of computational social science, most notably machine learning (Beieler *et al.,* 2016; Hanna, 2017; Zhang and Pan, 2019; Caselli *et al.,* 2021; Lorenzini *et al.,* 2022). While the automation of event data collection has proliferated "with declining costs of computational power" (Wiedemann *et al.,* 2022), the results are still mixed. So far, the few fully automated real-time event coding projects such as GDELT or Crisis Early Warning System (ICEWS), which also cover protests, remain confronted with serious validity and reliability issues (Wang *et al.,* 2016; Hoffmann *et al.,* 2022). However, parts of the workflow, notably the "haystack task" of identifying relevant articles, have been successfully automated. As argued by Wiedemann et al. in a recent overview of advances in the identification task (Wiedemann *et al.,* 2022), this step requires striking a tricky balance between recall and precision. By definition, increasing recall comes at the cost of decreasing precision. Since minimizing false negatives is usually considered the priority, a higher number of false positives need to be accepted and filtered out afterward (Croicu and Weidmann, 2015). For this, Zhang and Pan have applied a second-stage classifier concentrating on the elimination of false positives after an initial round of relevance classification (Zhang and Pan, 2019).

An alternative approach is improving the quality of the classifiers. In this regard, innovations in computational capacity have in recent years allowed the application of transformer-based neural networks. This approach does rely not only on the traditional unordered "bag-of-words" semantics but also on the structure and sequence of natural language (Wiedemann and Fedke, 2021). Depending on the strategy used, the results in the relevance detection vary greatly. Hanna reports a maximum *F*2 score of 0.77 (Hanna, 2017). Croicu and Weidmann's (2015) combination of various bag-of-words approaches yields 0.90 recall and 0.58 precision. Zhang and Pan (2019), who apply a recurrent neural network for the detection of protest events in social media, report highly satisfactory recall and precision scores of 0.96 and 0.95, respectively. Similarly, the BERT model applied by Hürriyetoglu et al.

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

achieves an $F1$ score of 0.9 (2019). Despite these highly satisfactory scores, both Hürriyetoglu et al. (2021) and Zhang and Pan (2019) document that the accuracy of their trained models significantly drops when applied out-of-sample. Despite these ongoing challenges, innovations in computational capacity and NLP modeling have led to a situation, in which an automation of the identification task reaches levels of accuracy comparable to humans, thereby reducing the need for human resources drastically.

Beyond the identification of relevant articles, however, various challenges remain in the automated coding of event characteristics. Hanna reports $F1$ scores ranging from 0.11 to 0.85 for protest forms, with only three forms surpassing a 0.5 threshold (Hanna, 2017, p. 23). Likewise, without detailing exact scores, the Count Love project notes that automated models "do not yet perform with sufficient recall and precision for fully automatic data entry" (Leung and Perkins, 2021). Due to unsatisfactory results in this more demanding task, various projects such as the Machine-learning Protest Event Data System (MPEDS; Hanna, 2017), POLCON (Lorenzini *et al.*, 2022), and the Crowd Counting Consortium (CCC; Fisher *et al.*, 2019) continue to rely on a semi-automated process with event annotation primarily performed by human coders. At the same time, several teams of researchers concentrate on the improvement of automated annotation of key variables. A key initiative in this regard is the CASE (Challenges and Application of Automatic Extraction of Socio-political Events from News) series of workshops (Hürriyetoğlu *et al.*, 2021, 2022, 2024), which aims at the collaborative tackling of key shared tasks. The organizers of CASE are currently working on GLOCON, a fully automated data set of "contentious politics events" in Turkey, India, China, Argentina, and Brazil (Duruşan *et al.*, 2022), of which technical details are not yet published. Beyond protest event data sets, BERT-based approaches include, for instance, the attempt to automatically identify 56 thematic categories in party manifestos, currently with $F1$ scores between 0.4 and 0.6 (Koh *et al.*, 2021).

Building upon this state of the art, with PAPEA, we are contributing to the current research frontier of automated classification of protest events. In the remainder of this article, we detail the architecture and performance of our fully automated pipeline and apply it to a use case of German-language newspaper articles.

3. A modular pipeline for the automation of protest event analysis

PAPEA largely follows a workflow similar to creating a manually annotated protest event data set. The overall design of the pipeline is modular. Each module corresponds to one step in the pipeline and has a well-defined input and output. Technically, a module can contain anything from a few lines of code to a sequence of tasks. If alternative and potentially better methods for individual steps become available, they can be "plugged in" into the pipeline according to the respective research project's needs, allowing for a high level of flexibility and adaptability.

Our current pipeline is able to identify and classify four core event variables: *protest form, protest topic, place*, and *date* of the protest. In addition, we have a module to distinguish between actual reports about past protests and articles announcing future protests. In Section 5 we report ideas on how to improve performance on *place* and *date*, as well as how to include modules on *actors* and *numbers* to the pipeline. Figure 1 illustrates the complete text processing pipeline. We start with a set of newspaper articles containing keywords related to protests. This keyword selection usually produces a high number of false positives because some indispensable keywords are polysemic (e.g. to demonstrate) or describe activities that sometimes but not always are a protest (e.g. march). The first step of reducing the original data set to only relevant articles is accomplished with a pre-trained sequence classifier. After reducing our stack of articles to only relevant ones, we identify and classify protest forms and topics in two parallel processing steps. For both tasks, we use sentence transformer models. After aggregating the results from both classification steps at the article level, we then run a token-based classifier to identify protest locations and a rule-based procedure for protest dates. After completion, we remove duplicate events and produce an event-based PEA dataset as the final output.

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

In the following section, we first briefly introduce the data set that was used to train and evaluate the various classifiers used in PAPEA, before discussing each step of the pipeline in detail. All Python and R scripts to run the pipeline as well as an example dataset are provided in the Online Appendix. The fine-tuned models are made available and documented at Huggingface.[2]

3.1. The gold-standard data set for model training and evaluation

The classifiers were trained on a hand-annotated gold standard dataset of local protests in four German cities—Bremen, Dresden, Leipzig, and Stuttgart—between 2000 and 2020. The dataset draws on German-language local newspapers—one outlet for each city (Leipziger Volkszeitung, Sächsische Zeitung [for Dresden], Weser-Kurier [for Bremen], and Stuttgarter Zeitung). We used the first 2500 manually annotated articles to train our classifier for relevance detection described below, and, after successful implementation, used it for selecting relevant articles that were then manually annotated. A team of student assistants manually extracted date, protest form, issue, number of participants, and several other variables. Quality and consistency of annotation were ensured by (a) a closely supervised team of student assistants including a continuous discussion of unclear cases and (b) detailed annotation guidelines that were tested and improved after the first six city years (Bremen 2015–16, Dresden 2014, Leipzig 2015–16, and 2018) were coded. For this purpose, the project's two PIs and two postdocs conducted a double annotation of 607 articles, checked for alignment, and adjusted annotation guidelines if necessary. Newspaper reports were annotated with the open-source web-based annotation software *Inception* (Klie *et al.,* 2018).

3.2. Task 1: Selecting relevant documents

The first module selects articles that actually report about a protest. Following the long-term German protest data set *Prodat* (Rucht *et al.,* 1992, p. 4), we define protest as every collective, public action of non-state actors that expresses a societal or political demand. We thus include not only demonstrations but also strikes, petitions, performances, or collective acts of political violence. For our German language newspaper corpus, we initially select articles that contain at least one of the following keywords[3]: "protest, assembly, demonstr*, rally, campaign, social movement, squat, strike, petition, hate crime, unrest, riot, insurrection, boycott, activis*, resistance, mobilis*, citizens' initiative" in all flexions.[4] This typically returns relevant and irrelevant articles at a ratio of about 1:10. To select relevant articles, we use the best-performing multilingual model described in Wiedemann et al. (2022)—a sequence classifier based on XLM-RoBERTa and fine-tuned with the gold standard dataset outlinedearlier. For the classification, it reduces the article text to only those sentences that contain one of the keywords plus the sentence before and after the respective keyword sentence (kwic + 1). Under optimal conditions, this classifier reaches an $F1$ score of 0.92, and even under the most adverse conditions, it yields an $F1$ score of 0.75 (Wiedemann *et al.,* 2022, p. 3389).

[2] Step 1 (relevance classification): shaunss/xlmroberta-pea-relevance-de; step 2 (form classification): shaunss/protestforms_mpnet-base-v2; step 3 (claims classification): shaunss/protestclaims_gbert.

[3] It can be argued that working with keywords introduced selection bias as it "might include irrelevant reports or fail to include relevant ones" (Althaus *et al.,* 2022, p. 608). We counter this critique with two points. First, we deliberately chose a broad range of expressions to maximize recall over precision in the selection process, i.e. to find as many true positives as possible, while not bothering about the number of false negatives. Second, under the assumption that using a large number of keywords guards against missing whole sections of the protest-related discourse, we argue that using a keyword-based approach in the phase of training the model is unproblematic because the model should understand the more general protest context and will be able to apply it to text that do not contain the keywords.

[4] The original German keyword list is as follows: "[Pp]rotest* OR Versammlung* OR [Dd]emonstr* OR Kundgebung* OR Kampagne* OR [s]oziale Bewegung* OR Hausbesetzung* OR Streik* OR Unterschriftensammlung* OR Hasskriminalität* OR Unruhen* OR Aufruhr* OR Aufstand* OR Boykott* OR Riot* OR Aktivis* OR Widerstand* OR Mobilisierung* OR Bürgerinitiative* OR Bürgerbegehren*."

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

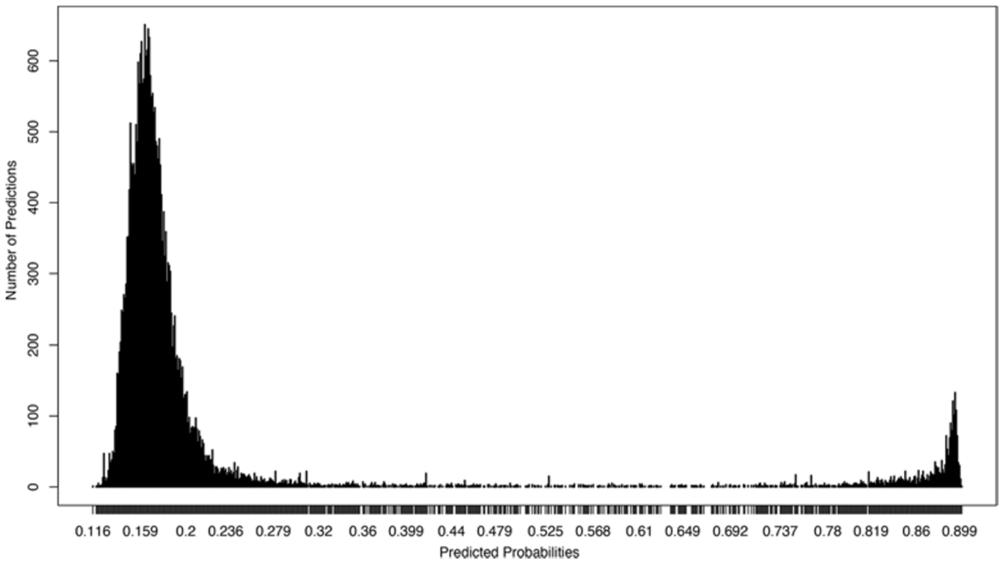

Figure 2. Relevance classification, distribution of predicted probabilities.

Figure 2 shows the distribution of the classifier's predicted probabilities on an example dataset not used for training, documenting a bimodal probability distribution typical for such a binary classification task: at the left local maximum, the model is confident that the documents do not contain a report about protests and, while at the right local maximum, it is confident that the documents do report about a protest. Between the two local maxima, the frequency of true positives decreases and the frequency of false positives increases.

The specific character of the distribution allows us to maximize for recall or for precision by selecting a lower or higher cutoff point for including documents in the final selection.

3.3. Task 2: Classifying protest forms

The protest form identification is realized as a sentence classifier. Information about the form—a demonstration, a petition, a strike, a performance, or another of the overall 22 protest forms in our codebook (Table 1)—can usually be identified at the sentence level, without having to account for a broader context. Therefore, sentence transformer models are ideal for this task.

Table 1 lists the relative frequency of the different forms of protest in our gold standard data set documenting that only 11 forms (highlighted in bold) have a relative frequency of over 1%. Demonstrations make up more than 50% of the protests. Together with strikes, petitions, attacks with property damage, leaflets/open letters, nonverbal/performative protests, blockades, and obstructions, these eight forms make up more than 90% of all protests.

For the form classification task, we fine-tuned a multilingual sentence BERT model (paraphrase-multilingual-mpnet-base-v2; Reimers and Gurevych, 2019) with labeled sentences containing protest form information from the ProLoc data set (Wiedemann *et al.*, 2022; Daphi *et al.*, 2024). In the training data, we included sentences with and without form information at a ratio of 1:2. In real-world data, the proportion of sentences containing form information is much smaller, but oversampling these sentences in the training data improves model performance. For fine-tuning, the dataset was split into a training and test set at the ratio 7:3, where half of the test set (15% of the labeled data) was used to improve model fit during training and the other half was used to evaluate the best fitting model. Sentences of the category "97 unclear" were deleted from the training data.

Table 1. Frequency of protest forms in the ProLoc data set

Protest form	%	Protest form	%	Protest form	%
1 Threat of litigation	0.36	**9 Disruption, obstruction**	**2.39**	17 Manslaughter, murder	0.08
2 Threat of murder/manslaughter	0.10	**10 Strike**	**9.99**	18 Attack on persons	0.92
3 Occupation	**1.44**	**11 Blockade, sit-in**	**3.25**	19 Threats	0.37
4 Demonstration, assembly, rally	**50.94**	12 Protest camp	0.35	20 Broadcasting campaign	0.31
5 Leaflet, resolution, open letter	**6.33**	**13 Attack with damage to property**	**4.54**	21 Boycott	0.18
6 Litigation	0.73	**14 Petition**	**8.17**	22 Online protest	0.08
7 Nonverbal protest, cultural event	**5.87**	**15 Scuffle**	**1.60**	97 Unclear	0.21
8 Press release, call for action	0.24	**16 Action resulting in personal injury**	**1.51**		

Table 2. Evaluation of form prediction

	Model evaluation during training				Evaluation at the article level			
Form	*P*	*R*	*F*1	*n*	*P*	*R*	*F*1	*n*
No form	0.94	0.93	**0.94**	1870				
Threat of litigation	0.50	0.33	0.40	3	0.20	1.00	0.33	1
Threat of murder/manslaughter	0.00	0.00	0.00	1	0.00	0.00	0.00	1
Occupation	0.75	0.32	0.44	19	0.75	0.91	0.82	66
Demonstration, assembly	0.78	0.88	**0.83**	504	0.96	1.00	**0.98**	1966
Leaflet, resolution, open letter	0.44	0.39	0.41	49	0.90	0.91	**0.90**	169
Litigation	0.60	0.43	0.50	7	1.00	0.75	**0.86**	16
Nonverbal protest, cultural event	0.38	0.31	0.34	49	0.91	0.82	**0.86**	163
Press release, call for action	0.00	0.00	0.00	1	1.00	0.14	0.25	7
Disruption, obstruction	0.10	0.10	0.10	21	0.90	0.31	0.46	29
Strike	0.82	0.93	**0.87**	87	0.97	1.00	**0.99**	397
Blockade, sit-in	0.57	0.43	0.49	37	0.95	0.84	**0.89**	43
Protest camp	0.00	0.00	0.00	4	1.00	0.25	0.40	12
Attack with damage to property	0.55	0.68	0.61	53	0.98	0.98	**0.98**	119
Petition	0.89	0.89	**0.89**	70	0.99	0.99	**0.99**	386
Scuffle	0.43	0.20	0.27	15	1.00	0.57	0.73	7
Action resulting in personal injury	0.25	0.12	0.17	16	1.00	0.83	**0.91**	18
Manslaughter, murder	0.00	0.00	0.00	1	0.00	0.00	0.00	11
Attack on persons	0.50	0.08	0.13	13	0.00	0.00	0.00	6
Threats	1.00	0.17	0.29	6	0.00	0.00	0.00	2
Boycott	0.00	0.00	0.00	1	0.00	0.00	0.00	3
Macro avg.	0.45	0.34	0.37	2827	0.90	0.51	0.76	3428
Weighted avg.	0.85	0.86	**0.85**	2827	0.95	0.96	**0.95**	3428

Note: Bold numbers indicate *F*1 scores where the model reaches or exceeds the average score of trained human annotators across all categories (see Table C1 in the Online Appendix).

Table 2 documents the performance of the form classifier. The left side (Model evaluation) displays the performance of the model on the evaluation dataset during training. Generally, the model performs much better in the high-frequency classes where it reaches $F1$ scores up to 0.89 (and 0.94 for the no-form class). For the lower frequency classes, $F1$ scores vary between 0.1 and 0.61. The evaluation is performed at the sentence level, i.e. the numbers reflect the accuracy of predictions for each sentence in the articles in the train/test set. This most likely inflates prediction errors from false positives because it de facto measures how good the model replicates the manual annotation, and not whether the model may have found the correct protest form in another sentence in which it was not manually annotated.

On the right side of Table 2, we therefore evaluate the performance of the classifier at the article level based on an extended version of our ProLoc data set containing about 3400 annotated articles that were not part of the training data. In this case, we no longer measure whether every sentence is predicted correctly, but only whether the most often predicted protest form in each article actually corresponds to a protest form that was also annotated manually for this article. We thus account for the information redundancy in newspaper articles and also add a primitive form of error correction by keeping only the most frequently found form for each article.

As a result, for most categories, we receive highly satisfactory results, with a weighted average $F1$ score of 0.95. For the most frequent categories (demonstration, strike, and petition), we even get $F1$ scores of up to 0.99. Lower $F1$ scores are mostly driven by low recall values, while precision is usually very high, even for the infrequent categories. The model thus overlooks some protests with less frequent protest forms, but when it classifies a protest as one of the less frequent forms, this classification is in most cases correct. At the article level, the model performs very well for all categories where the training data contained more than 200 examples (i.e. more than 40 cases in the evaluation dataset reported as n in the left table of Table 2). Performance in lower frequency classes possibly depends on the semantic distinctiveness of the category.

There are some very low-frequency classes that the model never predicts. We assume that this may be the result of the very low prevalence of these protest forms in the training data, and will discuss options to improve prediction of low-frequency categories in Section 5.

3.4. Task 3: Classifying protest topics

Classification of protest topics is realized as a two-step process. The goal is to classify the protests into 24 topics according to the ProLoc codebook (see Online Appendix A). The categories represent an updated and compatible version of the topic categories in the Prodat codebook. After initial tests, we realized that classification on the full text of the articles either with transformer or with topic models does not perform sufficiently well. We therefore decided to try sentence classifiers that have proven to perform very well on the task of classifying claims (i.e. demands) in newspaper texts (Haunss *et al.*, 2020; Dayanik *et al.*, 2022). An inherent challenge associated with sentence classifiers lies in the initial necessity of identifying relevant sentences. Regrettably, those sentences containing protest forms, which the module described in Section 3.3 aptly recognizes, frequently lack the information concerning the protest claims. Our strategy therefore is to first use a binary sentence classifier to identify possible claim sentences and then use a second sentence classifier to determine the categories.

For the first subtask, we use the manually annotated DEbateNet dataset (Lapesa *et al.*, 2020) to train a multi-layer perceptron as a claim identifier. Even though the model is trained on data from political debates on migration in Germany only, it performs surprisingly well also on other topics, with an $F1$ score of 0.78 (precision: 0.77, recall: 0.79) (Ceron *et al.*, 2024). While an $F1$ score of 0.78 may look underwhelming given the results for relevance classification and form detection, the information redundancy in newspaper articles guarantees that usually multiple sentences containing information about the protesters' claims exist. It should be noted that the topic-agnostic claim classifier does not only find claims made by protesters but also claims made, for instance, by addressees of the protest. Like in step 1, we address this issue by limiting the initial set of sentences to only those sentences containing one of our protest keywords plus the respective preceding and following sentences. The fact that in this context some claims come from addressees of the protests and not the protesters is not problematic because these claims still focus on the same topic.

For the second subtask, we fine-tune a sentence classifier (gbert-large-paraphrase-cosine) trained on the manually annotated protest events from the extended ProLoc data set described in Section 3.3. This classifier reaches a weighted macro average $F1$ score of 0.72 on the unseen test data during training (Table 3). One problem here is the strongly skewed distribution of topics, leading to seemingly unsatisfactory model performance in low-frequency categories. In high-frequency categories

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

Table 3. Evaluation of topic prediction

	Model evaluation during training				Evaluation at the article level			
Topic	*P*	*R*	*F*1	*n*	*P*	*R*	*F*1	*n*
Repression	0.30	0.30	0.30	10	0.48	0.68	0.56	40
Rights	0.26	0.36	0.30	14	0.31	0.79	0.45	49
Democracy	0.23	0.38	0.29	8	0.23	0.45	0.30	23
Media	0.33	0.50	0.40	2	0.50	0.25	0.33	3
Foreign_rights	0.50	0.33	0.40	3	1.00	0.83	**0.91**	5
Solidarity	0.00	0.00	0.00	1	0.50	0.25	0.33	3
Political	0.53	0.50	0.51	20	0.56	0.76	0.65	39
Economy	0.23	0.30	0.26	10	0.35	0.38	0.36	17
Peasants	0.33	0.50	0.40	2	1.00	1.00	**1.00**	6
Labur	0.92	0.88	**0.90**	131	0.98	0.92	**0.95**	239
Social	0.50	0.42	0.46	19	0.69	0.76	0.72	47
Education	0.64	0.70	0.67	30	0.91	0.90	**0.91**	69
Infrastructure	0.82	0.77	**0.79**	149	0.94	0.92	**0.93**	223
Environment (without nuclear)	0.76	0.77	**0.76**	48	0.97	0.74	**0.84**	73
Nuclear power	0.88	0.78	**0.82**	9	0.91	0.83	**0.87**	11
Gender	0.93	0.87	**0.90**	15	0.89	0.89	**0.89**	35
Migration	0.79	0.88	**0.83**	113	0.93	0.93	**0.93**	176
Peace	0.91	0.77	**0.83**	26	0.97	0.90	**0.94**	37
Anti-far-right	0.70	0.45	0.55	58	0.93	0.76	**0.84**	59
Tolerance	0.00	0.00	0.00	3	0.50	1.00	0.67	4
Far-right	0.29	0.67	0.40	9	0.50	0.61	0.55	22
Anticapitalist	0.50	0.50	0.50	2	1.00	1.00	**1.00**	2
International	0.59	0.59	0.59	17	0.86	0.78	**0.82**	29
COVID-19	0.38	0.43	0.40	7	0.80	0.63	0.71	15
Macro avg.	0.51	0.53	0.51	706	0.74	0.75	0.73	1226
Weighted avg.	0.74	0.72	**0.72**	706	0.83	0.82	0.82	1226

Note: Bold numbers indicate *F*1 scores where the model reaches or exceeds the average score of trained human annotators across all categories (see Table C1 in the Online Appendix).

like environment, infrastructure, labor, or migration the model reaches above average *F*1 scores between 0.79 and 0.90. The model also performs very well for topics like gender and nuclear power that have a more distinct semantic field.

Again, like in the case of the protest forms, performance at the article level is much better. When we only take the most often predicted topic per article, the weighted macro average *F*1 score rises to 0.82 and almost all frequent categories—and even some of the rare ones—are predicted with very high levels of accuracy (precision scores usually above 0.9 and F1 scores between 0.84 and 0.95). We can thus, again, profit from the redundancy in newspaper reporting. The model, however, overpredicts the rather abstract topic categories of repression, rights, and democracy.[5]

3.5. Task 4: Identifying protest location

Location, date, and number of participants are essential features of a PEA data set. In contrast to the procedures for classifying relevance, protest form, and protest issue that we discussed earlier, these features are not identified through paragraph- or sentence-based classifiers. The location identifier instead employs Named Entity Recognition (NER) at the token level because this currently still outperforms transformer-based approaches (Tanev and De Longueville, 2023) and does not require labeled training data. We use spaCy (Honnibal *et al.*, 2020) to identify location information—more specifically the pipeline optimized for German language processing, "de_core_news_md"—and then extract German cities based on a publicly available list of cities and ZIP codes. To be able to use the list

[5] In addition to comparing machine annotations to the human-generated test data, in Online Appendix C we report direct comparisons of human and machine performance on identifying forms and topics to a gold standard generated by the author team on a small subset of 100 articles.

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

of German cities for enhancing the results, a custom “EntityRuler” was incorporated into the pipeline following the pre-trained NER component, while the tok2vec, tagger, morphologizer, parser, lemmatizer, and sentence detector were disabled in order to reduce the processing time. The EntityRuler enables the specification of patterns for identification within the text—in this case the city names—and could be used with additional conditions, for instance, using information from the preceding step of the pipeline. If no city is found, we use the local newspaper’s place of publication.

Because our gold standard dataset only contains four cities, we manually annotated the actual protest location in 500 local newspaper articles in a separate dataset drawn from our experiment described in Section 4 to test the accuracy of the location prediction. Among all 469 true positives within these 500 articles (a precision of 0.94), the correct location was among the automatically identified locations in 78% of articles. However, the NER often identifies more than one location, which necessitates a selection rule. For this, we use the first identified location. This selection rule is crude but plausible, given that newspaper articles often begin by stating the place of the action they report on. With this rule, we can fully automatically identify the correct location in over two-thirds of true positives (68%).

We used the Mordecai 3 neural geoparser in Python[6] as an alternative. After running the Mordecai 3 model on our subsample, we selected the locations with the highest predicted scores. The model correctly identified the location in 64% of cases, which is slightly less precise than the NER in spaCy. Nonetheless, Mordecai 3 can serve as a reasonable alternative tool in other cases.

Since these values are sufficient for a first orientation, but far from the performance we achieve for relevance, forms, and topics, we discuss some ideas for improvement in Section 5.

3.6. Task 4: Identifying protest dates

To identify the protest dates, we apply a three-step process on the same protest keyword-containing subset of sentences as described for task 1. First, we search for time-indicating keywords such as weekdays, “weekend,” or relative time expressions like “today,” “tomorrow,” or “yesterday.” We consider the article publication date as the reference. For instance, if the text mentions “Monday” and the article is published on a “Wednesday,” we code the previous Monday as the protest event date. With this procedure, we cover 62% of all the cases. In the second step, we search the remaining articles for various other date formats found in the texts (e.g. “15 March 2007,” “8 April,” “October 2007,” and “in August”). In cases where the exact day is missing, we set the date as the 15th of the given month.

Finally, in those cases where the text does not contain any date expressions, we approximate the protest date by setting it as the day before the publication of the article. Applying this multi-step process, we identify the correct exact protest date in 52.5% of the total cases. In 80.1% of the cases, the predicted date is within 1 week of the actual date.

It is worth noting that the date detection process performs especially poorly when the temporal gap between the article publication and protest date is wider than one week. In 77.9% of the inaccurately predicted cases, the reported protest event occurred more than a week before the article’s publication.

4. Applying PAPEA to create a data set on local protests in Germany

In this section, we describe a first application of the PAPEA pipeline to a large collection of newspaper articles. This constitutes a real-world application case in which the pipeline is used

[6] See https://andrewhalterman.com/post/mordecai3/.

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

to extract information on protest events from unlabeled newspaper data. Given that the test data are unlabeled, we cannot compute precise performance measures. However, this test case does illustrate the profound advantages of the approach compared to time-intensive manual annotation.

4.1. Protest reporting in local newspapers

For the use case, we rely on the GermanLocalNews data set compiled at DeZIM. It contains articles from 146 German local newspapers published in a 10-year period (2010–2019). For our application of PAPEA, we limited the data to articles published in 2019 and ran them through our protest keyword list described earlier, arriving at 713,616 articles. Note that the manual annotation of about 500,000 articles that cover the 20 years of local newspaper reporting on protests in four cities in ProLoc took our five research assistants, each working 7–8 hours per week, about 3 years, and contrast this with the 17 hours compute time that it took to send the GermanLocalNews articles through the PAPEA pipeline.[7] This comparison alone underlines the immense advantages of finding well-performing automation procedures for the generation of PEA data.

From the 713,616 articles, we first subtract those that are identical duplicates, arriving at 536,995 articles from 89 newspapers. After relevance classification, we are left with 56,443 articles, which roughly corresponds to the 1:10 relation that we already noted in ProLoc. This suggests that the classifier trained on ProLoc works similarly on articles coming from different regions and newspapers than the ones it was trained on. Further removing articles in which our sentence-based classifiers did not find topics or forms reduces the data to 55,605 articles. We finally remove articles that are identical on publication date, topic, form, and identified location, thus removing articles that duplicate the protest event information, arriving at a data set containing 45,328 events.[8] Note that GermanLocalNews contains only 146 of the 498 local newspapers issued in Germany (Wikipedia, 2024) and thus contains only a fraction of the whole number of reported events (even though that fraction is likely larger than the fraction of covered newspapers). But still, this number is two orders of magnitude larger than, for instance, the 415 events that the PolDem-30 data report for Germany in 2019 based on national-level sources (Kriesi *et al.*, 2020). The difference in magnitude remains even if the number of 45,328 events would, for example, be halved by a more precise process of duplicate deletion.

4.2. Some protest trends in Germany 2019

We now illustrate some protest trends for Germany in 2019. Figure 3 shows a kernel density plot of the distribution of events. Two observations are immediately apparent. First, the protest wave of the Fridays for Future is clearly visible with its climate strikes in March, September, and November. Second, in contrast to PEA data sets that focus on national reporting and thus tend to mirror attention biases more than local-sourced PEA data (Hocke, 1996), we see much protest activity beyond these peaks as well.

Next, Figure 4 (left panel) displays the distribution of topics in the automatically generated data set. Environmental protests are strongly represented as the second largest category, suggesting again that the pipeline correctly identifies the FFF protest wave of 2019. However, the results also point to topics that were not present on the national news agenda to a similar extent: the largest and third-largest categories are infrastructure and labor-related protest. This matches the distribution for 2019 in the manually annotated ProLoc data set (Figure 4, right panel). As expected, the categories of

[7] Predictions were run on one NVIDIA V100 GPU node on a dedicated GPU server.

[8] Note that this procedure is dependent on correct identification of location and date, two areas where—despite some good first results—we still see room for improvement.

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

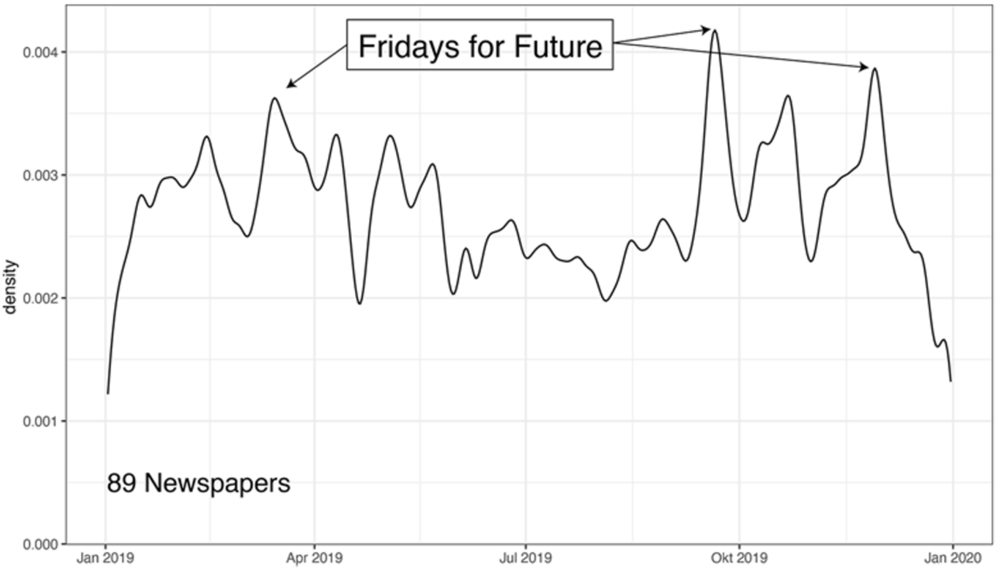

Figure 3. Distribution of automatically annotated articles containing protest events, Germany 2019.
Note: Kernel bandwidth chosen to maximize information while smoothing over seasonality (drop in articles every Sunday).

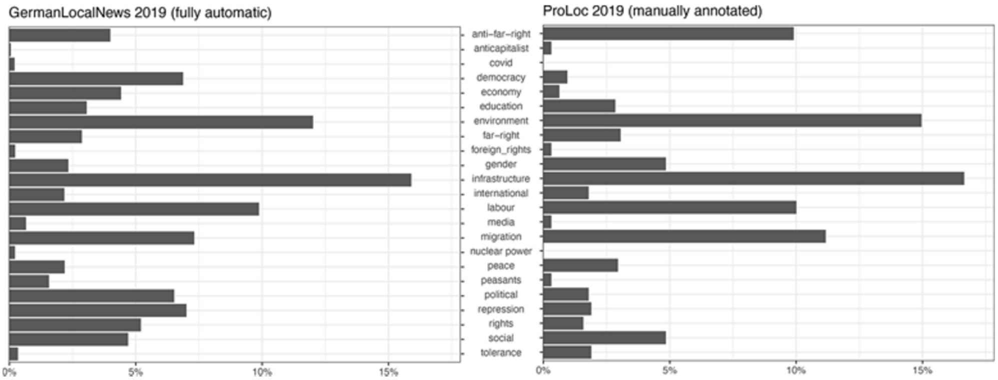

Figure 4. Topics addressed in automatically annotated protest events, Germany 2019.

rights, repression, and democracy are overpredicted, and some idiosyncrasies of our city selection in ProLoc come to the fore (for instance the higher share of far-right and anti-far-right protests owing to the inclusion of Leipzig and Dresden). But the picture is still very similar across the data sets. Environmental and infrastructure protests make up the largest categories with 15% and 17%, and migration, labor, and social protests are all among the most frequent categories in both graphs. This comparison again suggests that the automatically annotated protest event data captures empirically meaningful trends.

This short overview indicates that our pipeline can be put to productive use at this stage in its current, fully automated state, particularly to identify broad protest trends across a large number of sources and, for these trends, will deliver results of similar quality as data sets based on labor-intensive manual coding. At the same time, we identified several areas for improvement to further optimize the quality of the data set.

5. Generalizability and limitations

The pipeline that we have developed fulfills a very specific task: the automation of protest event analysis for German-language newspapers. To what degree can the same pipeline be used for more general tasks? We consider here two scenarios: (1) protest event analysis in other languages and (2) political actions other than protest.

With regard to the first question, we tested PAPEA on English-language newswire reports used in the PolDem project (Kriesi *et al.*, 2020). Our model for protest form classification was fine-tuned with German language training sentences only. But the underlying base model (paraphrase-multilingual-mpnet-base-v2; Reimers and Gurevych, 2019) was trained on texts from 50 languages and therefore our fine-tuned model should in theory also perform well on texts in these other languages. And, indeed, the fine-tuned classifier worked very well on the English language newswires and on protest forms (see Online Appendix D). Adapting our more fine-grained protest form categories to the less fine-grained protest form categories from the PolDem project (as mentioned earlier), we can report *F*1 scores between 0.79 for "violent protests" and 0.96 for "demonstrations." While we could not test the performance on protest topics (due to the incompatibility of the PolDem topic categories with our topic categories), the results for the protest forms show that our pipeline is not limited to German language texts and can be applied without modification to English newswire texts—and likely also to other news sources.

Second, an application of the PAPEA pipeline for other forms of political claims-making beyond protest would be feasible without much modification: it would require replacing the protest-form classification module with a new fine-funed model for other forms, like discursive interventions or lobbying. As long as the topics cover a similar field, the topic classification module could stay the same, since it already classifies protest topics based on only the claim sentences in the newspaper articles.

Nevertheless, PAPEA is not meant to be a general-purpose text classification toolkit, nor can it replace highly resource-intense general event collection and classification efforts like GDELT. Instead, it is a pipeline that can be used by relatively small research teams with access to limited computational and financial resources to automatically create protest event datasets tailored to their specific research interests.

In our current pipeline, we see three areas in which improvement is needed and where we have some suggestions on how to address current shortcomings in updated modules of the pipeline: *infrequent categories, precise location*, and *precise date*. Two further aspects would require adding new modules to the pipeline: *actor identification* and automatic recognition of the *number of protesters*.

5.1. Infrequent categories

The two main blocks of the pipeline, protest topic and form classification, work already very well for the more frequent categories. The lesser performance in the low-frequency categories where we had significantly less than 200 annotated examples is most likely a result of insufficient training data. To improve model performance for these categories, we suggest using the existing pipeline to identify in not yet annotated newspaper articles a larger number of form and topic sentence candidates that most likely cover these categories, then evaluate them manually and retrain the model with this additional data. This should be doable with a reasonable amount of work. Or one could even use LLMs to synthetically generate training data for empirically infrequent categories (Halterman, 2025).

Another option is to reduce the number of categories. Since our codebook for protest forms is relatively fine-grained—much more fine-grained, for instance, than the PolDem project (Kriesi *et al.*, 2020)—we computed two additional scenarios for a simplified form codebook. In the first, we aggregated our 20 forms into 6 macro-categories (symbolic physical, symbolic nonphysical, disruptive nonviolent, violent, strike, and legal action). The process of simplification is reported in Online Appendix B. In the first test, we retrained the model on the set of reduced categories and applied

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

it to the test data. We achieve a weighted macro *F*1 score of 0.9 on the sentence level (see Online Appendix B), which is already somewhat better than the 0.85 reported earlier. The best result, however, is achieved when predicting with the original 20-category model and later aggregating to the 6 macro categories. On the article level, we achieve 0.92 unweighted and 0.97 weighted *F*1 (see Online Appendix Table B3), which outperforms the original fine-grained model. Our results thus show that aggregating the predictions of the original model with more categories delivers better results than directly predicting the fewer aggregated categories. The reason for this is that broader categories are usually also semantically more diverse which sometimes makes them harder to predict than more specific categories.

5.2. Protest location

In addition to our country-specific solution, we also tested more general and more sophisticated geolocation tools like Mordecai 3, but were not able to get better results at the city level. Currently, we think that for projects that require geolocation at the city level, still some manual curation of the cases with lower certainty scores is necessary. For projects that require geographical precision at the state or national level only, existing general geolocation tools already perform very well (Halterman, 2023).

5.3. Precise dates

Experiments with more sophisticated rule-based time extraction models like *Heideltime* (Strötgen and Gertz, 2015) did not lead to better results than our rather crude search strategy. Caren et al. (2023) have recently used ChatGPT 3.5 for this task on a set of 500 locally reported protest events of the Black Lives Matter movement with very promising results (over 90% of correctly identified dates). After replicating and adapting their pipeline for our German language data, the results, however, were disappointing. In our test data of 2402 articles, ChatGPT 3.5 identified the correct date only in 47% of cases. One reason for this large discrepancy may be our much broader definition of protest, another reason may be differing performance of ChatGPT on German and English newspaper data in general. For the time being, we would argue that precise location and date identification still requires some human post-processing to deliver reliable data.

5.4. Actor identification

Identifying potential actors with current NER tools is almost trivial. But the problem of distinguishing between the protest actors, the addressees of the protest and other actors mentioned in the text, and canonization of actor names is still not solved. For canonization, some approaches exist. Actor recommenders like RePAIR (Solaimani *et al.,* 2017) work for actors prominent enough to be mentioned in Wikipedia or other knowledge bases. More recently, Cereon et al. (2024) have shown that combining LLMs for actor identification and pre-trained XLM models for canonization potentially can tackle the problem of distinguishing between actors with different roles.

5.5. Protester numbers

Again, using NER tools, it is easy to extract verbal and numeric representations of numbers from texts. Unfortunately, many of our texts contain multiple numbers, and determining which of these actually represent the protesters is far from trivial.

In principle—because both actor and number identification are easy to solve for humans—language models should in the future also be able to solve this, and then the respective modules can be added to the pipeline.

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

5.6. Multiple events per article

So far, our pipeline is able to identify possible cases of multiple events per article, e.g. when several forms and/or topics are identified, but as of yet, we do not have a reliable procedure for separating them. In the local newspaper data, however, multiple events are a relatively rare phenomenon limited to short and intense periods of protest-counterprotest interaction (in our data, this was the case in 2014–15 in Leipzig and Dresden). Given the highly complex way of local newspaper reporting in such cases, currently the most promising way to deal with this problem is to identify possible cases (either by multiple forms/topics or through a classifier trained on detecting protest-counterprotest interaction), and flag them for manual control.

A more general concern is that our pipeline might soon be outdated. This could be the case, for instance, when generative AI applications like ChatGPT become so computationally efficient that they can be used on millions of newspaper or online texts. Given the rapid change in machine learning development, this is clearly a possibility. In our view, here the modular character of the pipeline is again an advantage: single modules can easily be replaced if generative AI or other tools outperform our transformer-based classifiers, while keeping the ones intact that seem sensible to retain. Currently, generative LLMs do not yet outperform fine-tuned transformer-based models for our PEA tasks. In addition, the latter are still faster, are more energy efficient, and require less technical expertise to be run locally.

6. Conclusion

In current times of multiple, intersecting crises, the analysis of protest as a visible indicator of social and political conflict is more pertinent than ever. Protest event analysis has for decades been one of the methodological backbones of the study of protest, social movements, and contentious politics. But, because of its dependency on manual annotation, the applicability of protest event analysis has remained limited, and therefore, various research teams are working on pushing the frontiers of this method by leveraging advances in NLP methods (Fisher *et al.*, 2019; Oliver *et al.*, 2023; Hürriyetoğlu *et al.*, 2024).

This is paralleled with ongoing efforts in the automated identification and coding of political events more broadly. As a successor of ICEWS, Halterman and colleagues have recently come up with the "Political Event Classification, Attributes, and Types" (POLECAT) dataset based on a new coding ontology (Halterman *et al.*, 2023), arguing that "the global news sources available for coding near-real-time event data have completely outpaced any projects attempting to code these news sources with only human coders" (Halterman *et al.*, 2023).

Currently, we see several large-scale efforts to collect and classify political event data on conflicts on a global level. Projects like GDELT, ACLED, or POLECAT are resource-intensive endeavors, financed and supported by transnational internet companies, government agencies, or national security agencies. The benefits of these efforts are evident: automated event data collection would allow for an extension of geographic and temporal coverage, and thus enable answering much broader and more comparative research questions, compared to what is currently possible. Nevertheless, existing fully automated event databases still face severe issues of validity (Wang *et al.*, 2016; Hoffmann *et al.*, 2022) and usually remain limited to few forms of protest action. Moreover, tools and processing pipelines of these projects remain out of the reach of most social scientists with limited financial and computational resources in their academic institutions.

PAPEA is not intended to compete with these global event classification efforts. Instead, the focus lies on two aspects that will enable a broad applicability in many research settings: first, we push the limits of automating protest event coding and thus reduce annotation time by several orders of magnitude while maintaining high standards of accuracy. The combination of fine-tuned transformer models with rule-based approaches leads to high precision and recall along the pipeline. In addition,

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

PAPEA's modularity guarantees that alternative and potentially better methods can be "plugged in" once they become available.

Second, PAPEA is a pipeline that requires only moderate computational resources for its application and even for its training. It is thus a resource that can be used and controlled locally by various social scientists and allows them to build their own datasets tailored to their specific research interests. All models and tools used in PAPEA are publicly available which invites adaptation and improvement in a collective endeavor to improve the automation of PEA.

So far, our pipeline allows for an automated coding of the key variables—protest form and protest issue—at an accuracy level comparable to that of human coders. Other variables, including date and location, still require human post-processing to reach fully satisfactory results. Yet, we are rather optimistic that further technological advancements and adaptations of researchers working on the issue will continuously improve the performance of these steps in the pipeline.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2025.10013. To obtain replication material for this article, https://doi.org/10.7910/DVN/KVP7HA.

Acknowledgements. Research for this project was funded by the Federal Ministry of Education and Research (BMBF) (grant number 01UG2450CY) and the Federal Ministry of Family Affairs, Senior Citizens, Women and Youth (BMFSFJ; DeZIM).

References

Althaus S, Peyton B and Shalmon D (2022) A Total Error Approach for Validating Event Data. *American Behavioral Scientist* **66**(5), 603–624. doi:10.1177/00027642211021635.

Beieler J, Brandt PT, Halterman A, Schrodt PA and Simpson EM (2016) Generating political event data in near real time. In Alvarez RM (ed.), *Computational Social Science*. New York, NY: Cambridge University Press, 98–120.

Caren N, Andrews KT and Ray R (2023) Extracting protest events from newspaper articles with ChatGPT. doi:10.31235/osf.io/dvht7.

Caselli T, Mutlu O, Basile A and Hürriyetoğlu A (2021).PROTEST-ER: Retraining BERT for protest event extraction. *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021*), 12–19. doi:10.18653/v1/2021.case-1.4

Ceron T, Barić A, Blessing A, Haunss S, Kuhn J, Lapesa G, Padó S, Papay S and Zauchner P (2024) Automatic analysis of political debates and manifestos: Successes and challenges. *Proceedings of the 1st International Conference on Recent Advances in Robust Argumentation Machines*.

Croicu M and Weidmann NB (2015) Improving the selection of news reports for event coding using ensemble classification. *Research & Politics* **2**(4), 2053168015615596. doi:10.1177/2053168015615596

Daphi P, Dollbaum JM, Haunss S and Meier L (2024) Local protest event analysis: Providing a more comprehensive picture?. *West European Politics*. doi:10.1080/01402382.2024.2363709 (accessed June 21, 2024).

Dayanik E, Blessing A, Blokker N, Haunss S, Kuhn J, Lapesa G and Pado S (2022). Improving neural political statement classification with class hierarchical information. *Findings of the Association for Computational Linguistics: ACL 2022*, 2367–2382. https://aclanthology.org/2022.findings-acl.186 (accessed May 19, 2022).

Della Porta D (Ed.) (2014) *Methodological Practices in Social Movement Research*. Oxford: Oxford University Press.

Della Porta D and Diani M. (2006). *Social Movements. An Introduction*. 2nd Edn. Malden, MA: Blackwell.

Duruşan F, Hürriyetoğlu A, Yörük E, Mutlu O, Yoltar Ç, Gürel B and Comin A (2022) *Global Contentious Politics Database (GLOCON) Annotation Manuals* https://arxiv.org/abs/2206.10299 (accessed July 2, 2024).

Earl J, Martin A, McCarthy A and Soule SA (2004) The use of newspaper data in the study of collective action. *Annual Review of Sociology* **30**, 65–80.

Eck K (2012) In data we trust? A comparison of UCDP GED and ACLED conflict events datasets. *Cooperation and Conflict* **47**(1), 124–141. doi:10.1177/0010836711434463

Fisher DR, Andrews KT, Caren N, Chenoweth E, Heaney MT, Leung T, Perkins LN and Pressman J (2019) The science of contemporary street protest: New efforts in the United States. *Science Advances* **5**(10). doi:10.1126/sciadv.aaw5461

Halterman A (2023) Mordecai 3: A neural geoparser and event geocoder. arXiv:2303.13675. arXiv. doi:10.48550/arXiv.2303.13675 (accessed June 25, 2024).

Halterman A (2025) Synthetically generated text for supervised text analysis. *Political Analysis* 1–14. doi:10.1017/pan.2024.31

Halterman A, Bagozzi BE, Beger A, Schrodt PA and Scarborough GI (2023) Plover and Polecat: A new political event ontology and dataset. International Studies Association Conference Paper.

Hanna A (2017) MPEDS: automating the generation of protest event data. doi:10.31235/osf.io/xuqmv (accessed January 18, 2024).

Haunss S, Kuhn J, Padó S, Blessing A, Blokker N, Dayanik E and Lapesa G (2020) integrating manual and automatic annotation for the creation of discourse network data sets. *Politics and Governance* **8**(2), 326–339. doi:10.17645/pag.v8i2.2591

Hocke P (1996). *Determining the selection bias in local and national newspaper reports on protest events* (FS III 96-103). WZB. https://bibliothek.wzb.eu/pdf/1996/iii96-103.pdf (accessed June 16, 2023).

Hoffmann M, Santos FG, Neumayer C and Mercea D (2022) Lifting the veil on the use of big data news repositories: A Documentation and critical discussion of a protest event analysis. *Communication Methods and Measures* **16**(4), 283–302. doi:10.1080/19312458.2022.2128099

Honnibal M, Montani I, Van Landeghem S and Boyd A (2020) spaCy: Industrial-strength natural language processing in Python. doi:10.5281/zenodo.1212303 (accessed March 3, 2024).

Hürriyetoğlu A, Tanev H, Zavarella V and Yörük E (2022). Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE). Association for Computational Linguistics. https://aclanthology.org/2022.case-1.0 (accessed October 3, 2024).

Hürriyetoğlu A, Tanev H, Thapa S and Uludoğan E (Eds.) (2024) Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024). Association for Computational Linguistics. https://aclanthology.org/2024.case-1.0 (accessed March 3, 2024).

Hürriyetoğlu A, Yörük E, Mutlu O, Duruşan F, Yoltar Ç, Yüret D and Gürel B (2021) Cross-context news corpus for protest event-related knowledge base construction. *Data Intelligence* **3**(2), 308–335. doi:10.1162/dint_a_00092

Hürriyetoğlu A, Yörük E, Yüret D, Yoltar Ç, Gürel B, Duruşan F and Mutlu O (2019) A task set proposal for automatic protest information collection across multiple countries. In Yörük E, Yüret D, Yoltar BG, Duruşan F and Mutlu O (eds). *Lecture Notes in Computer Science*. Vol. 11438, 316–323. Cham: Springer. doi:10.1007/978-3-030-15719-7_42.

Hutter S (2014) Protest event analysis and its offspring. In Della Porta D (ed.), *Methodological Practices in Social Movement Research*. Oxford: Oxford University Press, 335–367.

Klie J-C, Bugert M, Boullosa B, Eckart de Castilho R and Gurevych I (2018) The INCEpTION Platform: Machine-assisted and knowledge-oriented interactive annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. https://aclanthology.org/C18-2002 (accessed November 11, 2022).

Koh A, Boey DKS and Béchara H (2021) Predicting policy domains from party manifestos with BERT and convolutional neural networks. doi:10.31235/osf.io/fjh4q (accessed January 16, 2024).

Kriesi H, Grande E, Dolezal M, Heibling M, Höglinger D, Hutter S and Wüest B (2012) Cambridge: Cambridge University Press.

Kriesi H, Koopmans R, Duyvendak JW and Giugni MG (1992) New social movements and political opportunities in Western Europe. *European Journal of Political Research* **22**(2), 219–244. doi:10.1111/j.1475-6765.1992.tb00312.x

Kriesi H, Wüest B, Lorenzini J, Makarov P, Enggist M, Rothenhäusler K, Kurer T, Häusermann S, Wangen P, Altiparmakis A, Borbáth E, Bremer B, Gessler T, Hunger S, Hutter S, Schulte-Cloos J and Wang C (2020). PolDem-protest dataset 30 european countries. accessed 2022-09-21, Version 1. https://poldem.eui.eu/downloads/pea/poldem-protest_30_codebook.pdf (accessed September 21, 2022).

Lapesa G, Blessing A, Blokker N, Dayanik E, Haunss S, Kuhn J and Padó S (2020). DEbateNet-mig15:Tracing the 2015 immigration debate in Germany over time. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 919–927. https://aclanthology.org/2020.lrec-1.115 (accessed November 22, 2022).

Leung T and Perkins LN (2021) Counting protests in news articles: A dataset and semi-automated data collection pipeline. arXiv:2102.00917). arXiv. doi:10.48550/arXiv.2102.00917 (accessed June 16, 2023).

Lorenzini J, Kriesi H, Makarov P and Wüest B (2022) Protest event analysis: Developing a semiautomated NLP approach. *American Behavioral Scientist* **66**(5), 555–577. doi:10.1177/00027642211021650

McAdam D, McCarthy J, Olzak S and Soule S (2009) *Dynamics of Collective Action*. https://web.archive.org/web/20230317173319/https://web.stanford.edu/group/collectiveaction/cgi-bin/drupal/

Oliver P, Hanna A and Lim C (2023) Constructing relational and verifiable protest event data: Four challenges and some solutions. *Mobilization: An International Quarterly* **28**(1), 1–22. doi:10.17813/1086-671X-28-1-1

Olsen H, Simon É, Velldal E and Øvrelid L (2024) Socio-political events of conflict and unrest: A survey of available datasets. In Hürriyetoğlu A, Tanev H, Thapa S and Uludoğan E (Eds.), *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text* (CASE 2024), 40–53. Association for Computational Linguistics. https://aclanthology.org/2024.case-1.5 (accessed October 10, 2024).

Raleigh C, Linke R, Hegre H and Karlsen J (2010) Introducing ACLED: An Armed Conflict Location and Event Dataset. *Journal of Peace Research* **47**(5), 651–660. doi:10.1177/0022343310378914

Reimers N and Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. doi:10.18653/v1/D19-1410

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press

Rucht D, Hocke P and Ohlemacher T (1992) Dokumentation und Analyse von Protestereignissen in der Bundesrepublik Deutschland (Prodat) Codebuch. WZB. https://bibliothek.wzb.eu/pdf/1992/iii92-103.pdf (accessed February 1, 2024).

Shuman E, Hasan-Aslih S, Van Zomeren M, Saguy T and Halperin E (2022) Protest movements involving limited violence can sometimes be effective: Evidence from the 2020 BlackLivesMatter protests. *Proceedings of the National Academy of Sciences* **119**(14), e2118990119. doi:10.1073/pnas.2118990119

Solaimani M, Salam S, Khan L, Brandt PT and D'Orazio V (2017). RePAIR: Recommend political actors inreal-time from news websites. *IEEE International Conference on Big Data*. http://ieeexplore.ieee.org/servlet/opac?punumber=8241556 (accessed August 23, 2024).

Strötgen J and Gertz M (2015) A baseline temporal tagger for all languages. In Màrquez L, Callison-Burch C and Su J (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 541–547. Association for Computational Linguistics. doi:10.18653/v1/D15-1063

Tanev H and De Longueville B (2023) Where "where" matters: Event location identification with a bert language model. *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text*, 11–17. doi:10.26615/978-954-452-089-2_002

Tilly C (1978) *From Mobilization to Revolution*. Reading, Mass: Addison-Wesley.

Wang W, Kennedy R, Lazer D and Ramakrishnan N (2016) Growing pains for global monitoring of societal events. *Science* **353**(6307), 1502–1503.

Wiedemann G, Dollbaum JM, Haunss S, Daphi P and Meier LD (2022) A generalizing approach to protest event detection in German local news. *Proceedings of the 13th Conference on Language Resources and Evaluation*, 3883–3891. http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.413.pdf (accessed August 22, 2022).

Wiedemann G and Fedke C (2021) From frequency counts to contextualized word embeddings: The Saussurean turn in automatic content analysis. In Engel U, et al. (ed.), *Handbook of Computational Social Science*. London: Routledge, Vol. 2, 366–385.

Wikipedia (2024) *Liste deutscher Zeitungen*. Wikipedia. https://de.wikipedia.org/w/index.php?title=Liste_deutscher_Zeitungen&oldid=241476104 (accessed November 26, 2023).

Zhang H and Pan J (2019) CASM: a deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology* **49**(1), 1–57. doi:10.1177/0081175019860244

Cite this article: Haunss S, Daphi P, Dollbaum JM, Hristova L, Susánszky P and Steinhilper E (2025) PAPEA: A modular pipeline for the automation of protest event analysis. *Political Science Research and Methods*, 1–18. https://doi.org/10.1017/psrm.2025.10013

https://doi.org/10.1017/psrm.2025.10013 Published online by Cambridge University Press