PAPEA – A Modular Pipeline for the Automation of Protest Event Analysis

Sebastian Haunss, Priska Daphi, Jan Matti Dollbaum, Lidiya Hristova, Pál Susánszky, Elias Steinhilper

Appendix A: Event selection and codebook

The ProLoc data set was compiled in a four-step process, involving (1) selection of the cities, (2) a keyword search, (3) a fully automated relevance selection procedure, and (4) manual annotation of the selected articles. Each selected city – Bremen, Dresden, Leipzig and Stuttgart – is a regional center, with populations ranging from 500,000 to 600,000. At the same time, we chose cities based on political and structural differences to induce systematic variation with the aim to cover different structural environments for protest, including geographic location (two in the East and two in the West) and political background (two with predominantly center-left and two with predominantly center-right governments over the past two decades).

In the data collection phase for each of the chosen cities spanning the years 2000 to 2020, we relied on major local newspapers: Leipziger Volkszeitung, Sächsische Zeitung (for Dresden), Weser-Kurier (for Bremen), and Stuttgarter Zeitung. For the corpus we initially downloaded articles from Factiva, LexisNexis and Genios containing at least one of the following keywords: "protest, assembly, demonstr*, rally, campaign, social movement, squat, strike, petition, hate crime, unrest, riot, insurrection, boycott, activis*, resistance, mobilis*, citizens' initiative" in all flexions.[1] Through an automated selection procedure utilizing a transformer-based classifier (see Wiedemann et al. 2022), we then filtered out news articles containing protest-related information. Subsequently, aided by a closely supervised team of student assistants, we manually annotated articles adhering to our codebook (see below).

[1] The original German keyword list is: "[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*"

Table A1. Codebook of ProLoc

Variable label	*description*	*label*
FORM	Forms of action	0 No form mentioned 1 Threat of litigation 2 Threat of murder/ manslaughter 3 Occupation 4 Demonstration, assembly, public protest rally 5 Leaflet, resolution, open letter 6 Litigation 7 Non-verbal protest, cultural event 8 Press release, call for action 9 Disruption, obstruction 10 Strike 11 Blockade, sit-in 12 Protest camp 13 Attack with damage to property 14 Petition 15 Scuffle 16 Action resulting in personal injury 17 Manslaughter, murder 18 Attack on persons 19 Threats 20 Broadcasting campaign 21 Boycott 22 Online protest 97 unclear
DATUM	Date on which PE begins	Day.Month-Year
TRAEGER 1, 2, 3...	Individuals or collectives who carry out PE	0 not specified 1 Individual 2 Collective name (workers etc.) 3 Informal group/ citizens' initiative 4 Trade union 5 Association 6 Church 7 Party 8 NGO/association 9 Alliance 10 Local authority 11 Other 12 anonymous 97 unclear
ZAHL	Number of people involved in the PE	Number of participants

Table A1. Codebook of ProLoc

CLAIM1, 2, 3...	Demands of the PE	101 repression 102 rights 103 democracy 109 media 140 foreign_rights 154 solidarity 200 political 400 economy 500 peasants 600 labour 700 social 800 education 900 infrastructure 1000 environment (without nuclear) 1100 nuclear power 1200 gender 1300 migration 1400 peace 1510 anti-far-right 1511 tolerance 1520 far-right 1530 anticapitalist 1600 international 9902 COVID 19
CLAIMADR (für CLAIM1, 2 ,3)	Addressee of the PE	0 not specified 1 State institutions 2 Political parties 3 Trade associations and companies 4 Trade unions 5 Other associations, churches 6 Public officials 7 Private individuals 8 Diffuse, society as a whole 9 Social subgroups 10 Other social movement 11 Other 97 unclear
CLAIMEB	Problem level of the articulated claim	0 Not specified 1 municipal 2 regional 3 nationwide 4 nationwide 5 Europe 6 International 97 unclear
REAKDEMO	Does the PE trigger a counterpro-test	0 Not specified 1 Counter-protest reported 2 No counter-protest reported 3 PE is counter-protest 97 unclear
GEWDEMO	Reports of violence by protesters at the PE	0 Not specified 1 Violence by protesters 2 Explicit: No violence 97 unclear

Table A1. Codebook of ProLoc

GEWPOL	Reports of police violence at the PE	0 Not specified 1 Violence against protesters 2 Explicit: No violence 97 Unclear
AUFPOL	Termination of the PE by the police	0 Not specified 1 Termination 97 unclear
TRAGSOZ 1, 2 , 3..	Social groups who carry out the protest	0 not specified 1 Employed 2 Unemployed 3 Asylum seekers 4 Farmers 5 Women 6 Young people/students 8 Pensioners 10 Students 13 Religious groups 18 Ethnic groups 19 heterogeneous 99 other
TPERSON	Individuals highlighted by name individuals	
BEMERK	Other interesting aspects	

Appendix B: Testing a simplified version of the form codebook

Table B1 below documents the process of reducing the granularity of our form codebook to six macro-categories.

Table B1: *Documentation of protest form category reduction*

Old category	*New category*
4 Demonstration, assembly, public protest rally	Symbolic physical
7 Non-verbal protest, cultural event	
15 Scuffle	
5 Leaflet, resolution, open letter	Symbolic non-physical
8 Press release, call for action	
14 Petition	
20 Broadcasting campaign (“Sendeaktion”)	
21 Boycott	
22 Online protest	
3 Occupation	Disruptive non-violent
9 Disruption, obstruction	
11 Blockade, sit-in	
12 Protest camp	
2 Threat of murder/ manslaughter	Violence
13 Attack with damage to property	
16 Action resulting in personal injury	
17 Manslaughter, murder	
18 Attack on persons	
19 Threats	
10 Strike	Strike
1 Threat of litigation	Legal action
6 Litigation	
97 unclear	other
0 No form mentioned	no form

We run two tests with the simplified form codebook. Table B2 and B3 display the results. First, we retrained the model on the simpler version displayed in Table B1 above and applied it to the sentence level. Performance is somewhat better than the results shown in Table 2 (left-hand side) in the manuscript.

***Table B2**. Performance of re-trained form prediction model (sentence level)*

	precision	recall	f1-score	n
no form	0.95	0.94	0.95	1900
symbolic physical	0.82	0.85	0.83	477
symbolic non-physical	0.73	0.82	0.78	118
disruptive non-violent	0.71	0.53	0.61	66
violence	0.66	0.70	0.68	61
strike	0.82	0.93	0.87	84
legal action	0.62	0.31	0.42	16
accuracy			0.90	2722
macro avg	0.76	0.73	0.73	2722
weighted avg	0.90	0.90	0.90	2722

Table B3 uses the original model with the 20-category form codebook, applies it to the article level and simplifies the form categories only after application. This outperforms the article-level version reported in Table 2 of the manuscript (right-hand side). In particular, the unweighted macro F1 increases drastically, which is the result of previously underperforming categories being aggregated into larger ones.

***Table B3**. Performance of original form prediction model aggregated to 6-category scheme after prediction (article level)*

	precision	recall	f1-score	n
symbolic physical	0.97	0.99	0.98	2134
symbolic non-physical	0.98	0.95	0.97	573
disruptive non-violent	0.89	0.77	0.83	150
violence	0.99	0.9	0.94	149
strike	0.97	1.00	0.99	397
legal action	0.82	0.82	0.82	17
macro avg	0.94	0.78	0.92	3425
weighted avg	0.97	0.97	0.97	3425

Appendix C: Direct comparison of human and machine performance against gold standard

To test the performance of human coders versus the model, we retrospectively created a gold standard of 100 articles that had been both hand and machine annotated.[2] The member of the authors team who was most familiar with the codebook (due to leading the training of the annotators during the duration of the whole 3-year project) identified and manually registered the two most often reported topics and the two most often reported forms in each of the 100 articles. We then identified the most often annotated topics and forms per article in the manually annotated data and aggregated the machine-annotated sentences for topic and form to the article level. In Table C1 below we document macro F1 and weighted macro F1 scores for (1) the comparison of the gold standard and the human annotators and (2) the comparison of the gold standard and the machine annotations. In both cases, we compute F1 scores for the agreement between gold standard on the one hand and humans or the machine on the other in two scenarios: (A) whether humans or the machine identify exactly the most often found claim or form in the gold standard or (B) whether humans or the machine identify either the most often identified *or the second most often identified* topic or form in the gold standard.

Table C1*. Direct comparison of performance of human and machine annotations on 100 articles (upper value: macro F1, lower value: weighted macro F1)*

forms			**topics**		
Comparison of gold standard with...	(1) human coders	(2) machine annotations	Comparison of gold standard with...	(1) human coders	(2) machine annotations
(A) exactly most often identified form per article	.63 .81	.67 .81	(A) exactly most often identified claim per article	.76 .75	.59 .52
(B) most often or second most often identified form per article	.96 .99	.92 .94	(B) most often or second most often identified claim per article	.96 .92	.95 .77

As can be seen from the table, in this test against the retrospective gold standard on the article (not the sentence) level, humans and machines perform very similarly on identifying forms. On topic, humans outperform the model on the identification of most often found one, but when the task is to identify the most or second most often found topic in the article, the machine achieves good to excellent results, too. Figures C1 and C2 below display the confusion matrices for scenario A for both comparisons. It visually confirms the impressions from Table C1. In addition, it shows that the lower performance of the machine compared to human annotators for topics in scenario A appears partly located *within* plausible broader issue categories, mitigating the underperformance: the machine sometimes conflates labour and economy and it sometime conflates far-right, anti-far-right and migration protests. Since far-right protests often invoke migration-related demands and are often countered by anti-far-right demonstrations on the spot, which tends to be reported in the same article, the latter case again demonstrates the yet unsolved

[2] The original gold standard data set reported on in section 3.1 of the manuscript had been used tot rain annotators who had amended their annotations according to the gold standard, so that the original annotations were unavailable for comparing human and machine annotation performance.

problem with articles that report about multiple events (most often: protests and counterprotests) that we have identified in the article. If the topic categories were aggregated after classification to broader issue categories like social, cultural, political issues (see, e.g. Daphi et al. 2025)[3], it is to be expected that the machine performance would be even better.

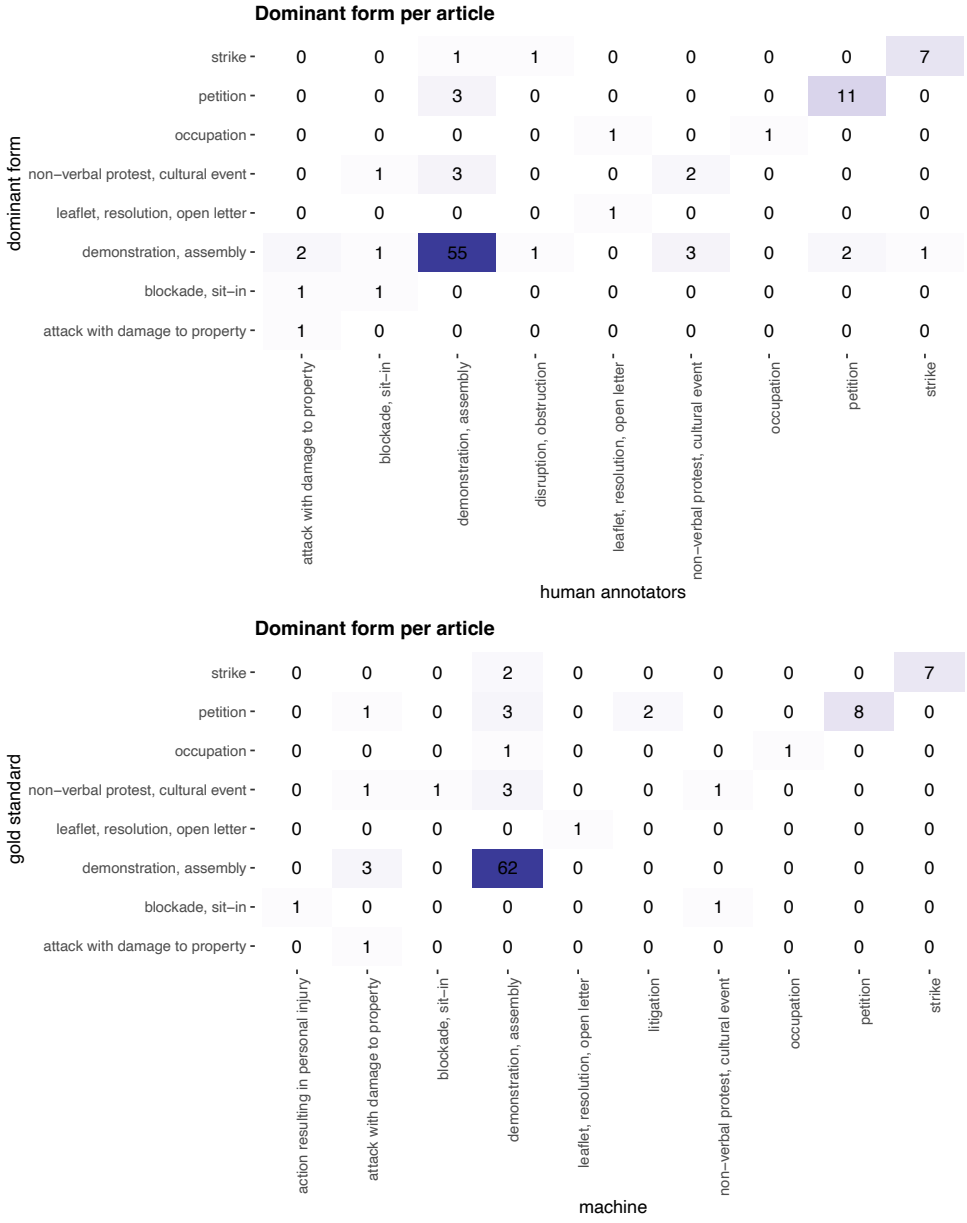

***Figure C1**. Confusion matrix of human and machine performance against the gold standard on form identification (aggregated to claim level, scenario A, see Table C1).*

[3] Daphi, Priska, Jan Matti Dollbaum, Sebastian Haunss, and Larissa Meier. 2025. "Local Protest Event Analysis: Providing a More Comprehensive Picture?" *West European Politics* 48 (2): 449–63.

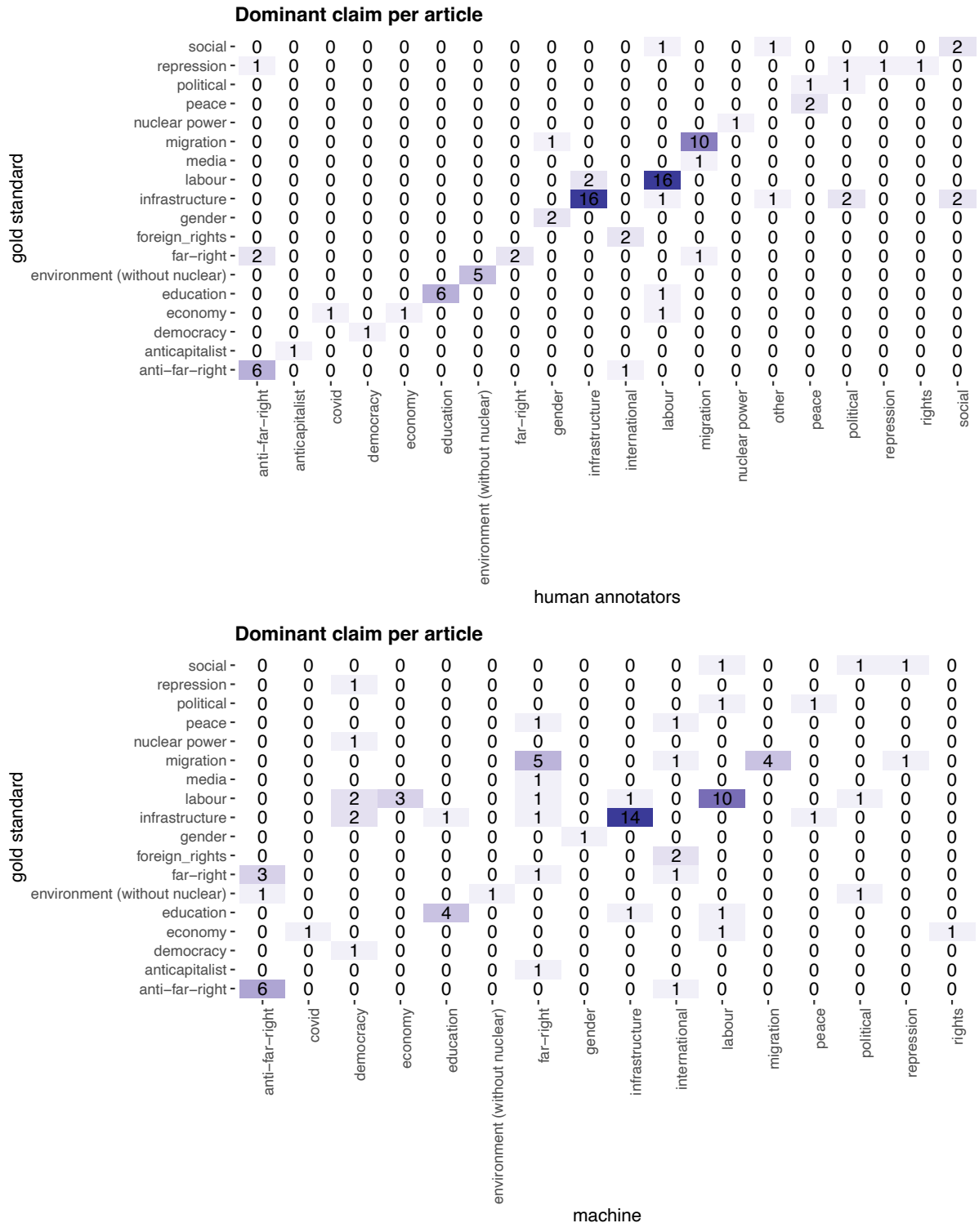

***Figure C2**. Confusion matrix of human and machine performance against the gold standard on claim identification (aggregated to article level, scenario A, see Table C1).*

Appendix D. Test of form classification on English-language data

As described in the main text, we tested our model on English-language newswire data from 30 countries contained in the PolDem dataset sourced through 10 English-language news agencies (Kriesi et al. 2020). Specifically, we passed 4711 PolDem events through the module of our pipeline that classifies protest forms.[4] Table D1 shows the harmonization of PolDem and ProLoc form categories.

Table D1. *Harmonization of form codes across PolDem and ProLoc*

PolDem	**ProLoc**
strikes	10 Strike
demonstrations	4 Demonstration, assembly, public protest rally
	12 Protest camp
confrontations, blockades	2 Threat of murder/ manslaughter
	3 Occupation
	9 Disruption, obstruction
	11 Blockade, sit-in
	19 Threats
violent protest	13 Attack wich damage to property
	15 Scuffle
	16 Action resulting in personal injury
	17 Manslaughter, murder
	18 Attack on persons
petitions, symbolic actions	5 Leaflet, resolution, open letter
	7 Non-verbal protest, cultural event
	8 Press release, call for action
	14 Petition
	20 Broadcasting campaign
	21 Boycot
other protest	1 Threat of litigation
	6 Litigation
	97 other
no form	0 no form

Table D2 displays the results. Despite the fact that the PolDem data come from a different kind of source (newswires) and a different language (English) compared to what the model was trained on we receive very good (.79) to excellent (.96) F1 scores for the identification of the substantive PolDem categories. As can also be seen from the confusion matrix (Figure D1), the only underperformance is the precision for "no protest", meaning that our model identifies several protest forms where the PolDem data does

[4] Unfortunately, due to strong differences in the codebooks between the two projects, a similar test on topics was impossible, but the two teams are currently working on harmonizing the issue categories to allow for such tests.

not show any. Nonetheless, the performance on the substantive categories strongly suggests that the model is applicable to different data in other languages with a similar degree of accuracy on actual forms, while somewhat overpredicting protest, particularly violent forms of protest and demonstrations.

Table D2. *Results of form prediction on PolDem data*

	precision	**recall**	**f1**	**n**
confrontations, blockades	1.00	0.78	**0.88**	223
demonstrations	1.00	0.92	**0.96**	2087
none	0.14	0.99	0.25	679
other protest	1.00	0.06	0.11	3
petitions, symbolic actions	1.00	0.75	**0.86**	197
strikes	1.00	0.85	**0.92**	273
violent protest	1.00	0.66	0.79	331
macro	0.88	0.71	0.68	3793
macro average	0.85	0.89	**0.80**	

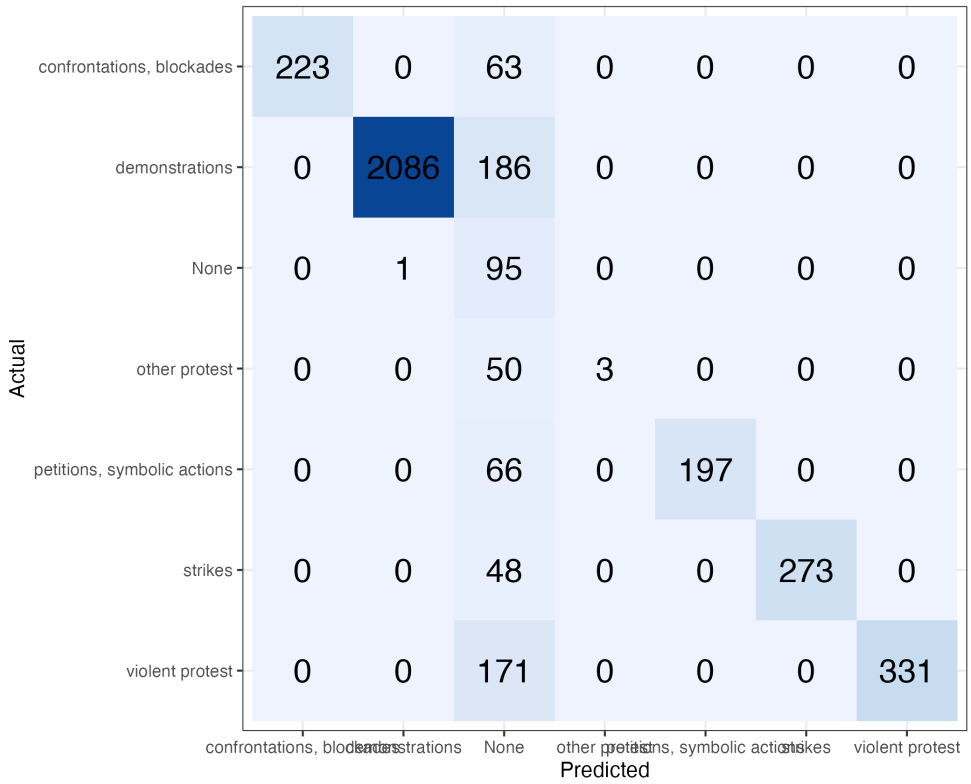

Figure D1. *Confusion matrix, prediction of PolDem data*