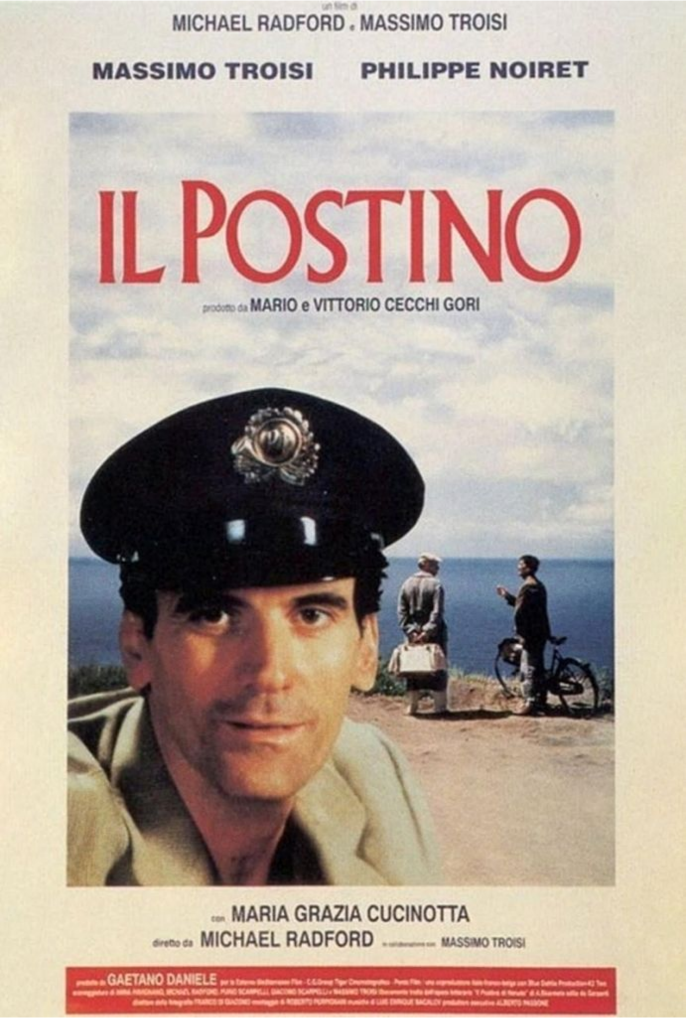
Analysing Cross-Country Protest Dynamics: A Transformer-based Approach to Newspaper Content

Giuliano Formisano (University of Oxford), Caterina Froio (SciencesPo, Paris), Pietro Castelli Gattinara (SciencesPo, Université Libre de Bruxelles)

EPSA Conference 2024



What do newspaper articles about “il postino” and far-right protests have in common?

Motivation

- The evolving nature of collective action and abundant digital text prompt researchers to develop new tools to study its complexities.

- Limiting time, costs and increasing replicability of protest event analysis (Hutter 2014a, 2014b; Lorenzini et al., 2022, Nardulli 2015; Zhang and Pan 2019).

- Previous models present some weaknesses:
  - Protest event selection bias: struggle generalise to unseen texts.
  - Bag-of-words-based approaches ignoring word semantics and order leading to information loss and data sparsity issues.
  - Previous Transformer approaches limited in language and tasks.

Research question & goals

- **Research question:** How LLMs can enhance the efficiency and accuracy of cross-country protest event analysis?

- Two key objectives:

1. To identify articles discussing protest events beyond mentioning keywords.
2. To annotate nuanced characteristics of protest events: issue (i.e., religious and ethnic minorities vs others) and protest form (i.e., demonstrative vs violent).

Data

- **FARPO dataset:** 4,002 manually annotated protest events described in newspaper articles in Austria, Belgium-Wallonia, France, the Netherlands, Germany, Sweden, and Spain (https://farpo.eu/).

- **Time window:** 2008-2018 (focus: economic and cultural impact of the 2008 global financial crisis and 2015 EU migration policy crisis).

- **Sampling:** Actor-centered keyword search on Factiva and Lexis-Nexis (Berkhout et al. 2015).

- Appendices 1 and 2 contain a data summary by task and country.

Annotations

- **Coders:** 6 coders speaking at least one of the languages under analysis.

- **Intercoder reliability tests:** To check consistency and description biases - Cronbach's alphas, finding high levels of reliability, averaging 0.97.

- **Protest identification:** Annotation for relevance, distinguishing articles discussing actual protest events from those merely mentioning keywords.

- **Protest characteristics:** The dataset is annotated for event characteristics: (1) ethnic/religion minorities and (2) violence.

Methods

- We train **3 multilingual supervised machine learning classifiers** on three tasks (protest identification, issue, and forms of action).
- **Models**: XLM-roBERTa (2020) and mBERT (2019).
- **Splits (training, validation, test)**: 60/20/20%, 70/15/15%, and 80/10/10%.
- **Seed:** Randomly selected seeds to split human-coded texts into sets, and in the sequence classification work.
- **Evaluation metrics:** Accuracy, Fscore, AUC.
- **Robustness**: Parameters' freezing and zero-rule baseline.

Results

Table 1: Descriptive statistics of best performing algorithms on the testing sets

Task	Model	Seed	Split	Accuracy	Average F-score	Individual F-scores	AUC
Protest Identification	XLM roBERTa	373	80% training 10% validation 10% testing	80%	80%	78% vs 81%	0.80
Protest Issue		387	70% training 15% validation 15% testing	75%	75%	72% vs 77%	0.75
Protest Action Form		973	70% training 15% validation 15% testing	75%	75%	78% vs 71%	0.77

Note: The table shows the descriptive statistics of the best performing algorithms on the testing sets for each task under analysis. We present the random seed and split used. We also provide four performance metrics: Accuracy, Average F-score, Individual F-scores for each class in the model, and Area-under-the-curve (AUC). Appendices contain descriptive analyses of each classifier using all configurations (model, seed, split, and metrics).

Conclusions

- Integrating machine learning models (LLMs) into protest event analysis enhances efficiency and accuracy while addressing time and cost constraints.

- Training classifiers to identify articles discussing protest events beyond keyword mentions improves dataset precision.

- Using additional classifiers allows annotation of nuanced protest event characteristics, enhancing analysis depth: (1) ethnic/religious minorities involvement, (2) demonstrative vs violent protests.

- Increased portability: our models can be applied to other languages (e.g., English and Italian) and several types of texts (e.g., tweets, press releases).

Thank you very much!

🤗

Return to Data

Appendix 1: Data Summary by Task and Country

Task 1: Data summary	
Country	**Texts**
Germany	1,930
France	599
Spain	471
Sweden	394
Netherlands	260
Austria	190
Belgium	158
Total	**4,002**

Task 2a: Data summary	
Country	**Texts**
Germany	913
France	294
Spain	225
Sweden	185
Netherlands	128
Austria	91
Belgium	77
Total	**2,546**

Task 2b: Data summary	
Country	**Texts**
Germany	1,019
Spain	686
France	308
Sweden	216
Netherlands	132
Austria	98
Belgium	82
Total	**1,911**

Return to Data

Appendix 2: Average Texts Lengths by Country

Task 1: Average text lengths	
Country	**Texts**
Germany	665
France	835
Spain	612
Sweden	704
Netherlands	986
Austria	424
Belgium	423
Overall	**688**

Task 2a: Average text lengths	
Country	**Texts**
Germany	572
France	903
Spain	501
Sweden	756
Netherlands	904
Austria	389
Belgium	472
Overall	**642**

Task 2b: Average text lengths	
Country	**Texts**
Germany	581
Spain	226
France	903
Sweden	820
Netherlands	901
Austria	396
Belgium	488
Overall	**511**

Appendix 3 - Task 1: Protest Event Identification

Binary classification to identify articles explicitly discussing relevant protest events.

Labels distribution:

- Non-protest: 2,092 texts
- Protest: 1,910 texts

Speed [using top classifier]: 401 texts coded in 2min49sec (2.37it/s)

Model	Seed	Split	Accuracy	Average F-score	Individual F-scores	AUC
XLM roBERTa	449	60% training 20% validation 20% testing	78%	77%	75% vs 80%	0.77
	257		76%	76%	77% vs 75%	0.76
	861		74%	74%	72% vs 76%	0.74
	385	70% training 15% validation 15% testing	78%	78%	76% vs 80%	0.79
	206		73%	72%	70% vs 75%	0.75
	920		76%	76%	77% vs 75%	0.76
	102	**80% training 10% validation 10% testing**	73%	73%	72% vs 74%	0.73
	835		79%	79%	79% vs 79%	0.79
	373		**80%**	**80%**	**78% vs 81%**	**0.80**
mBERT	493	60% training 20% validation 20% testing	77%	77%	77% vs 77%	0.78
	89		75%	75%	78% vs 72%	0.75
	759		74%	74%	73% vs 75%	0.74
	501	70% training 15% validation 15% testing	76%	76%	78% vs 72%	0.75
	895		76%	75%	73% vs 78%	0.76
	946		80%	80%	77% vs 82%	0.80
	477	80% training 10% validation 10% testing	76%	76%	74% vs 78%	0.77
	832		73%	72%	67% vs 77%	0.73
	50		74%	74%	73% vs 75%	0.74
Frozen XLM roBERTa	532	80% training 10% validation 10% testing	79%	79%	80% vs 76%	0.78
	987		71%	70%	74% vs 67%	0.71
	270		70%	70%	75% vs 64%	0.69
	431		74%	73%	76% vs 70%	0.74

Appendix 4 - Task 2a: Protest Event Issue (ethnic /religious minorities vs others)

Binary classification to identify articles explicitly discussing ethnic /religious minorities.

Labels distribution:

- Non-ethnic/religious issue: 1,010 texts
- Ethnic/religious issue): 903 texts

Speed [using top classifier]: 287 texts coded in 1min57sec (2.44it/s)

Model	Seed	Split	Accuracy	Average F-score	Individual F-scores	AUC
XLM roBERTa	129	60% training 20% validation 20% testing	69%	69%	70% vs 67%	0.69
	624		71%	70%	63% vs 76%	0.69
	917		70%	69%	72% vs 68%	0.71
	516	**70% training 15% validation 15% testing**	**75%**	**75%**	**72% vs 77%**	**0.75**
	387		72%	72%	63% vs 78%	0.70
	789		69%	68%	61% vs 74%	0.68
	122	80% training 10% validation 10% testing	70%	70%	68% vs 71%	0.70
	8		74%	74%	68% vs 79%	0.73
	804		68%	67%	60% vs 73%	0.66
mBERT	28	60% training 20% validation 20% testing	67%	67%	66% vs 67%	0.67
	661		70%	70%	67% vs 73%	0.70
	499		67%	66%	60% vs 72%	0.67
	714	70% training 15% validation 15% testing	71%	71%	68% vs 73%	0.71
	564		70%	70%	65% vs 74%	0.70
	135		67%	67%	66% vs 68%	0.67
	75	80% training 10% validation 10% testing	73%	73%	75% vs 72%	0.73
	327		74%	74%	71% v 77%	0.74
	613		69%	68%	61% vs 74%	0.67
Frozen XLM roBERTa	907	70% training 15% validation 15% testing	61%	54%	30% vs 73%	0.57
	752		56%	46%	22% vs 69%	0.55
	623		53%	36%	00% vs 69%	0.50

Appendix 5 - Task 2b: Demonstrative vs Violent

Binary classification to identify articles explicitly discussing demonstrative vs violent

Labels distribution:

- Demonstrative: 1,126 texts
- Violent: 785 texts

Speed [using top classifier]: 287 texts coded in 1min48sec (2.64it/s)

Return to results

Model	Seed	Split	Accuracy	Average F-score	Individual F-scores	AUC
XLM roBERTa	718	60% training 20% validation 20% testing	64%	65%	67% vs 62%	0.66
	65		73%	72%	79% vs 62%	0.70
	541		68%	68%	71% vs 64%	0.68
	374	**70% training 15% validation 15% testing**	66%	66%	68% vs 64%	0.67
	129		70%	69%	76% vs 60%	0.68
	973		**75%**	**75%**	**78% vs 71%**	**0.77**
	386	80% training 10% validation 10% testing	68%	69%	72% vs 63%	0.69
	650		75%	74%	81% vs 64%	0.72
	812		69%	69%	76% vs 57%	0.66
mBERT	826	60% training 20% validation 20% testing	73%	72%	79% vs 62%	0.70
	901		70%	70%	75% vs 64%	0.70
	541		69%	68%	75% vs 60%	0.67
	553	70% training 15% validation 15% testing	61%	62%	65% vs 57%	0.62
	471		60%	60%	62% vs 56%	0.62
	270		68%	68%	69% vs 68%	0.70
	747	80% training 10% validation 10% testing	69%	70%	74% vs 63%	0.70
	75		66%	66%	71% vs 59%	0.65
	457		70%	70%	77% vs 60%	0.68
Frozen XLM roBERTa	626	70% training 15% validation 15% testing	59%	44%	74% vs 00%	0.50
	358		59%	44%	74% vs 00%	0.50
	191		55%	39%	71% vs 00%	0.50