Multilingual Political Content Classification across Time and Space: An evaluation

Brian Boyle, Sebastian Popa & Zoltán Fazekas

May 2023 | COMPTEXT | Glasgow

Motivation & goals

1. Evaluate multi-context classifiers (time and space)
   - Does additional training data from different contexts aid (poor) content classification performance?
   - Which approaches work well and what are the sources of variation?

2. Explore classifier performance without context-specific training data
   - Can we combine translation and embedding approaches?
   - What is the overall performance loss and what the sources of variation?

Status: work-in-progress on a sub-sample of content labels and models with aim to maximize data labeling on collected data

*** *data collection part of the Pro-Con EU project*

Data collection

EP 2019 Political campaigning on Twitter dataset (Stier et al, 2020)

- All tweets by EP candidates (+ public replies, mentions, & retweets)
- 16 million tweets in 28 countries and 31 languages
- 500,000 MEP candidate tweets collected between 23 April and 30 May

Manual coding

- 17 research assistants hired to code tweets across 11 languages
- 9,000 tweets per coder: tweets split by candidate/public, then by country (for candidates), and language
- Random sample taken for each language, but: weighted so that 75% candidate tweets, 25% public

Note: for languages with more than one coder, 2,000 of the tweets were coded by both for inter-coder reliability checks

Data summary

		All tweets		Sampled tweets	
Country	**Language**	**Candidates**	**Public**	**Candidates**	**Public**
UK	English	131,332	5,113,760	13,500	4,500
France	French	62,403	2,911,611	13,500	4,500
Spain	Spanish	52,824	2,328,691	13,500	4,500
Italy	Italian	17,826	1,834,711	13,500	4,500
Poland	Polish	43,770	1,048,559	13,500	4,500
Netherlands	Dutch	13,793	433,309	7,500	2,750
Germany*	German	13,156	371,372	13,156	4,500
Greece*	Greek	4,349	72,301	4,349	32,000
Hungary*	Hungarian	326	2,118	326	2,118

*All candidate tweets were manually coded for these countries.

Coding summary

1. Political or personal content, sentiment, communication style
2. For political content: campaign messaging and/or political issues
3. If political issue, which issue (up to 3) :
   - Economy, Environment, EU, Brexit, Immigration, Support/opposition to democratic values, Anti-elitism, Crime and justice, Other
   - Open ended answers were inspected, re-grouped into larger categories or added to existing categories

Upon completion, coders were also asked to apply the same coding to political tweets from the 2014 EP campaign in the UK, Germany, and Spain (retro-coding, extending work by Theocharis et al, 2016)

Data structuring

(selected) Outcomes of interest, tweet mentions:

- political issue (1) *vs.* all other content (0)
- the environmental issue (1) *vs.* all other content (0)
- the economy (1) *vs.* all other issue content (0)

Training and test data

- DE-19, HU-19, GR-19 coded completely $\rightsquigarrow$ always training data only
- For each other context, 80%-20% split, stratified on outcome prevalence

  FR19, DE14, IR19, IT19, NL19, PL19, ES14, ES19, UK14, UK19

Important: training and test splits are kept identical across all experiments (until Question 2)

Data structuring

	Political issue		Environment		Economy (issue only)	
	Test	Train	Test	Train	Test	Train
DE14	0.13/728	0.13/2915	0.01/729	0.01/2914	0.16/94	0.16/373
ES14	0.23/453	0.22/1808	0.03/452	0.03/1809	0.3/101	0.30/407
ES19	0.35/2800	0.35/11201	0.02/2801	0.02/11200	0.21/989	0.21/3956
FR19	0.55/2699	0.55/10798	0.24/2699	0.24/10798	0.29/1478	0.29/5912
IR19	0.43/2939	0.43/11756	0.10/2939	0.10/11756	0.32/1255	0.32/5019
IT19	0.46/1508	0.46/6033	0.05/1509	0.05/6032	0.44/695	0.44/2780
NL19	0.40/1499	0.40/5993	0.10/1499	0.10/5993	0.32/598	0.32/2395
PL19	0.40/2505	0.40/10017	0.06/2504	0.06/10018	0.23/1001	0.23/4008
UK14	0.24/645	0.24/2579	0.03/645	0.03/2579	0.15/157	0.15/629
UK19	0.46/2699	0.46/10796	0.08/2699	0.08/10796	0.12/1242	0.12/4967

* Proportion of 1s followed by sample size.

Question 1:

Do joint (translated) text classifiers work better?

Question 1: approach

Baseline

- For each context, pretrained multilingual document embedding based on original language of tweet (cased Bert-base multilingual)
- Choice based on previous performance in political content coding of tweets (see for example Cross et al, 2022)
- Currently: 2014 is fitted within-language with year dummy

Classifiers

- Dichotomous classifier: xgBoost with grid-based parameter tuning using 5-fold cross-validation
- Choice based previous performance in political content coding of tweets (Fazekas et al, 2021) or in general (see Grinsztajn et al, 2022)
- Currently: comparisons with simpler regularized regressions have been carried out and xgBoost consistently outperformed these

Question 1: approach

Input for joint models

- Document-feature matrix based on translated tweets: unigrams with preprocessing, tf-idf transformation
- Multilingual pretrained document embedding based on original language of tweet (cased Bert-base multilingual)
- English language pretrained document embedding based on translated tweet (cased Distilbert-base)
- Included: year dummies (when applicable) and fitted with- or without country dummies

Evaluation: F1 score, which is weighted average of precision and recall, where *precision* is the proportion of relevant instances among the retrieved instances, or $\frac{tp}{tp+fp}$ and *recall* is the proportion of relevant instances that were retrieved, or $\frac{tp}{tp+fn}$

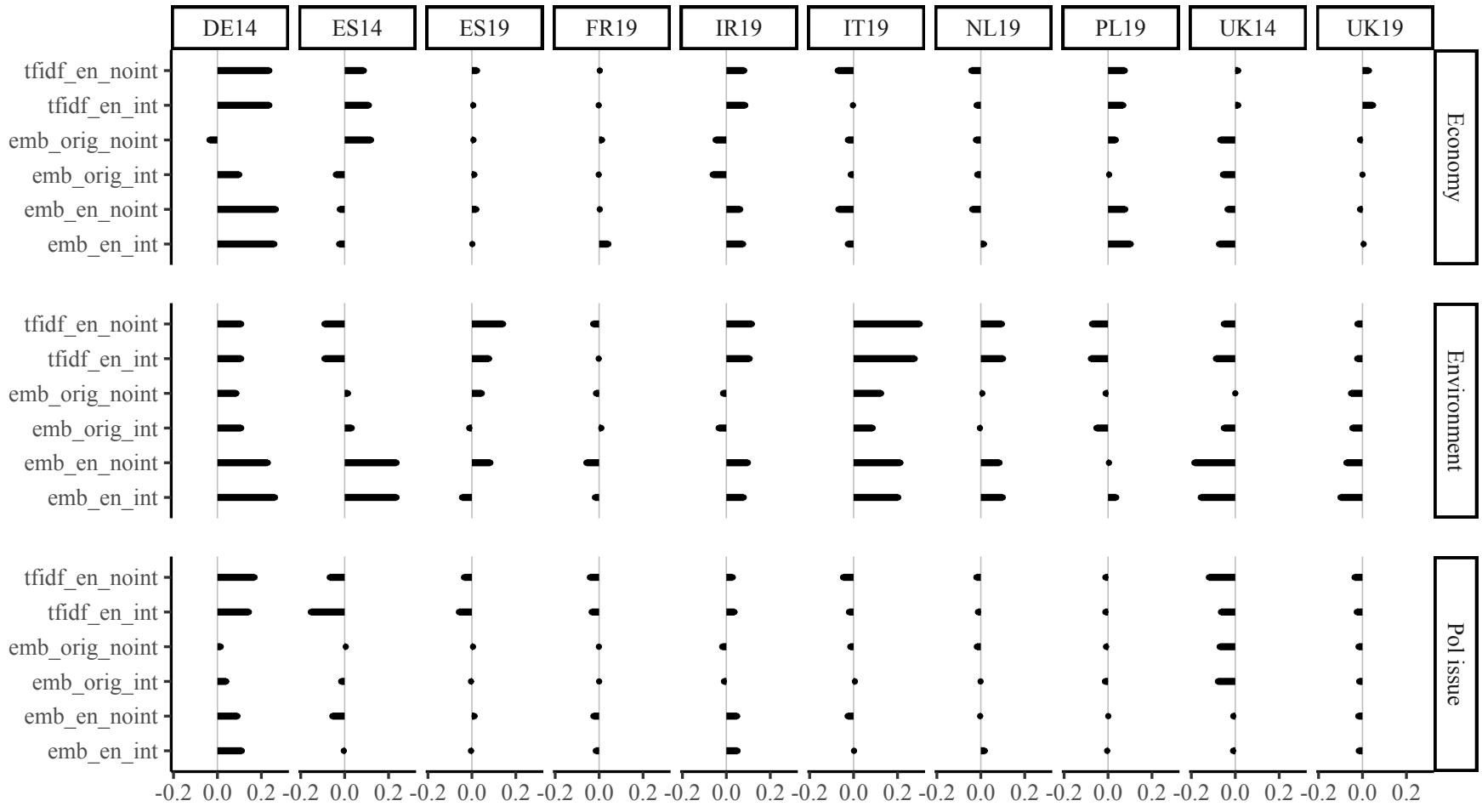
Overall between-approach differences

	Political issue		Environment		Economy (issue only)	
	$F1_{diff}$	P_{high}	$F1_{diff}$	P_{high}	$F1_{diff}$	P_{high}
Emb_{en} wo cntry	0.001	0.4	0.063	0.7	0.025	0.5
Emb_{orig} wo cntry	-0.013	0.3	0.019	0.5	-0.004	0.4
Tf-idf_{en}wo cntry	-0.018	0.2	0.05	0.5	0.043	0.8
Emb_{en}	0.013	0.4	0.06	0.6	0.038	0.7
Emb_{orig}	-0.009	0.2	0.004	0.4	-0.007	0.3
Tf-idf_{en}	-0.02	0.2	0.039	0.5	0.054	0.7

*F1 difference averaged across contexts, + alternative is better.

*P_{high} = proportion of contexts where alternative outperforms baseline.

Context level F1 differences compared to baseline F1



Overall context-level differences

	Political issue		Environment		Economy (issue only)	
	$F1_{base}$	$F1_{diff}$	$F1_{base}$	$F1_{diff}$	$F1_{base}$	$F1_{diff}$
DE14	0.328	0.095	0.200	0.150	0.286	0.177
ES14	0.568	-0.049	0.235	0.056	0.64	0.038
ES19	0.656	-0.014	0.449	0.049	0.673	0.011
FR19	0.855	-0.020	0.767	-0.018	0.706	0.010
IR19	0.729	0.023	0.510	0.058	0.573	0.033
IT19	0.784	-0.017	0.231	0.201	0.744	-0.035
NL19	0.750	-0.006	0.527	0.064	0.778	-0.021
PL19	0.722	-0.008	0.564	-0.029	0.571	0.061
UK14	0.624	-0.059	0.571	-0.089	0.500	-0.034
UK19	0.725	-0.022	0.649	-0.053	0.569	0.010

*F1 difference averaged across 6 scenarios.

Answer 1: no, mostly small benefits

Small and inconsistent differences

- Exceptions related to DE-14, IT-19
- Systematically better performance for specific issue content
- Tentative: low performing benchmark context benefit from joint classifiers, thus potential ranking and use of only "better" contexts can be explored

Between-approach, issue, and context variation

- Possible sources: sample size, balance in outcome, baseline performance
- Non-independence of these features and possible flooring/ceiling effects

F1 gains correlated with data features

	r_{prop1}	r_{n}	r_{f1base}
Political issue	-0.24	0.02	-0.43
Environment	-0.22	-0.19	-0.63
Economy (issue only)	-0.29	-0.19	-0.59

	r_{prop1}	r_{n}	r_{f1base}
Emb_{en} wo cntry	-0.39	-0.14	-0.78
Emb_{orig} wo cntry	-0.27	-0.10	-0.40
Tf-idf_{en}wo cntry	-0.39	-0.08	-0.62
Emb_{en}	-0.30	-0.24	-0.71
Emb_{orig}	-0.20	-0.17	-0.56
Tf-idf_{en}	-0.27	-0.11	-0.53

Question 2:

Can we use (translated) text classification to label political content in contexts without human coded data?

Question 2: overall approach

Baseline: (joint) classifiers on previously used training data (three versions of input), with year and country dummies

Comparisons

1. Without 2014: removed from training all 2014 and evaluated on full 2014 data
2. Without FR19 and NL19 (removed simultaneously from training), evaluated on full FR19 and NL19 data
3. Without FR19 and NL19 (removed simultaneously from training), evaluated on the previously used test set from FR19 and NL19

Classifiers and evaluation: identical (xgBoost and F1 scores)

- Changes regarding the left-out samples
- *(not included)* Changes regarding the average performance, once training left out

Time: F1 scores comapred

		Political issue		Environment		Economy (issue only)	
		$F1_{base}$	$F1_{wo2014}$	$F1_{base}$	$F1_{wo2014}$	$F1_{base}$	$F1_{wo2014}$
DE14	Emb_{en}	0.439	0.459	0.462	0.362	0.545	0.605
	Emb_{orig}	0.368	0.441	0.308	0.216	0.385	0.508
	$Tf\text{-}idf_{en}$	0.471	0.503	0.308	0.519	0.522	0.623
ES14	Emb_{en}	0.564	0.546	0.471	0.346	0.615	0.664
	Emb_{orig}	0.552	0.471	0.267	0.237	0.600	0.556
	$Tf\text{-}idf_{en}$	0.414	0.516	0.143	0.216	0.750	0.630
UK14	Emb_{en}	0.614	0.620	0.414	0.359	0.426	0.429
	Emb_{orig}	0.545	0.594	0.519	0.382	0.444	0.337
	$Tf\text{-}idf_{en}$	0.559	0.627	0.483	0.378	0.513	0.532

* Models without country intercepts in appendix.

Time: overall F1 score differences

	Political issue		Environment		Economy (issue only)	
	$F1_{base}$	$F1_{diff}$	$F1_{base}$	$F1_{diff}$	$F1_{base}$	$F1_{diff}$
DE14	0.426	0.042	0.350	0.012	0.463	0.128
ES14	0.512	0.013	0.293	0.013	0.682	-0.062
UK14	0.570	0.039	0.482	-0.111	0.461	-0.023

*Baseline F1 and difference averaged across all models.

*F1 difference is $F1_{without}$ - $F1_{baseline}$.

Countries: F1 scores compared

		Political issue			Environment			Economy (issue only)		
		$F1_{base}$	$F1_{wo1}$	$F1_{wo2}$	$F1_{base}$	$F1_{wo1}$	$F1_{wo2}$	$F1_{base}$	$F1_{wo1}$	$F1_{wo2}$
FR19	Emb_{en}	0.840	0.771	0.757	0.748	0.475	0.488	0.748	0.569	0.564
	Emb_{orig}	0.855	0.586	0.602	0.777	0.302	0.330	0.704	0.526	0.536
	$Tf\text{-}idf_{en}$	0.821	0.782	0.771	0.765	0.620	0.624	0.704	0.626	0.633
NL19	Emb_{en}	0.768	0.700	0.723	0.627	0.484	0.502	0.792	0.675	0.713
	Emb_{orig}	0.749	0.503	0.510	0.523	0.259	0.279	0.762	0.529	0.588
	$Tf\text{-}idf_{en}$	0.737	0.705	0.707	0.628	0.594	0.642	0.759	0.699	0.719

*Models without country intercepts in appendix.

*$F1_{wo1}$ = all FR19 and NL19 is test, $F1_{wo2}$ = original FR19 and NL19 is test.

Countries: overall F1 score differences

Table: Test set: Full FR19 and NL19 data

	Political issue		Environment		Economy (issue only)	
	$F1_{base}$	$F1_{diff}$	$F1_{base}$	$F1_{diff}$	$F1_{base}$	$F1_{diff}$
FR19	0.835	-0.152	0.750	-0.290	0.714	-0.159
NL19	0.744	-0.142	0.593	-0.146	0.760	-0.131

*Baseline F1 and difference averaged across all models.

Table: Test set: original shared test set (20%)

	Political issue		Environment		Economy (issue only)	
	$F1_{base}$	$F1_{diff}$	$F1_{base}$	$F1_{diff}$	$F1_{base}$	$F1_{diff}$
FR19	0.835	-0.155	0.750	-0.279	0.714	-0.155
NL19	0.744	-0.131	0.593	-0.114	0.760	-0.099

*F1 difference is $F1_{without}$ - $F1_{baseline}$.

Answer 2: maybe, with some limitations

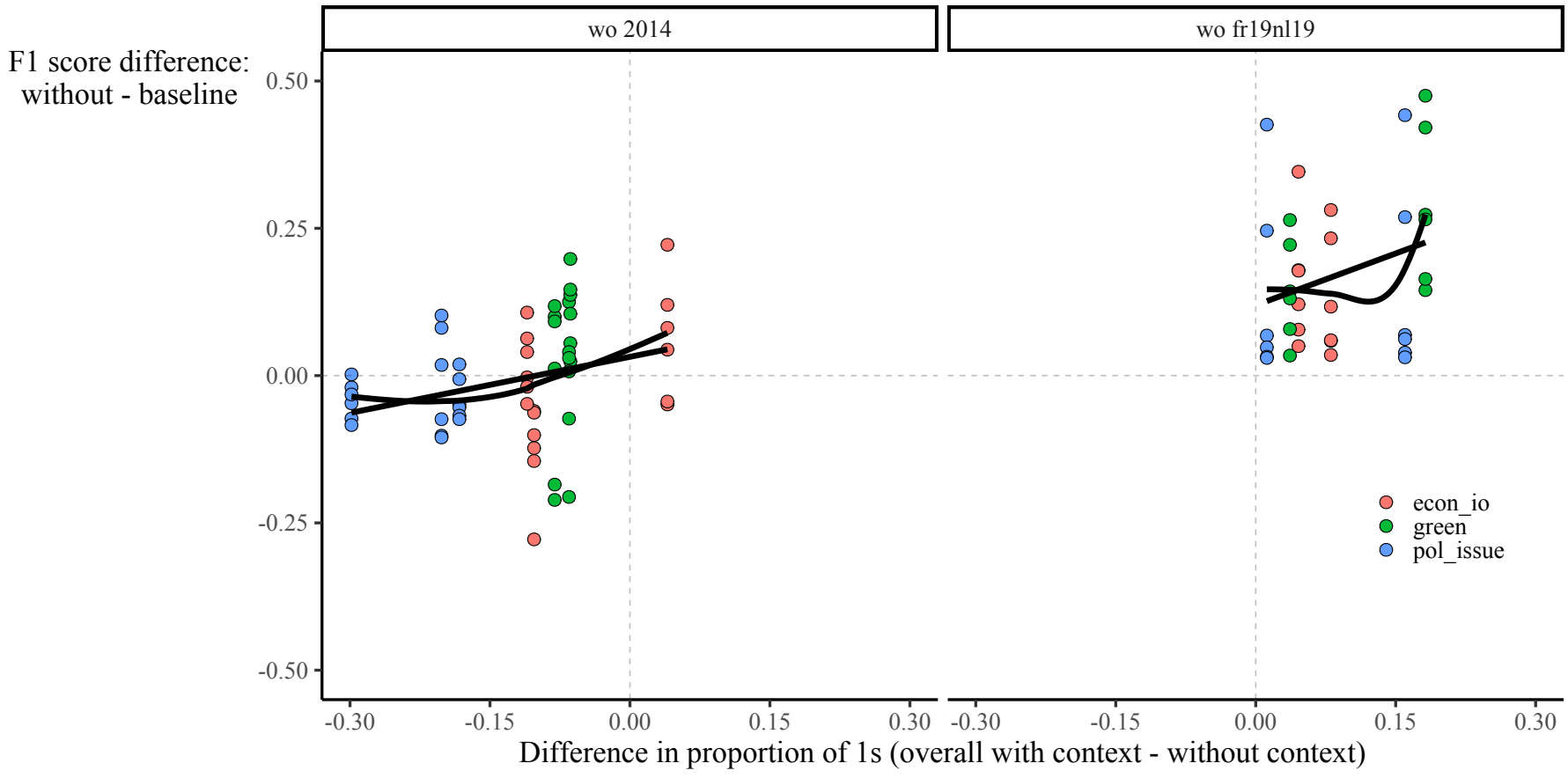
Within-country, across-time:

- Promising results, but overall weak classifiers
- Caveat 1: unclear which modeling approach is most reliable, although *tf-idf* seems particularly good
- Caveat 2: data quality and sample size might factor in

Between-country:

- Expected results, with an average drop of 0.14 (NL19) to 0.22 (green issue) in F1 scores, with some specific combinations being even worse
- Across model and content heterogeneity
- Caveat 1: larger samples and better performing classifiers can contribute to ceiling effects

F1 differences and proportion differences



Overall (tentative) conclusions

Mixed results

- Minor and varying gains from joint classifiers
- English translation + embedding combination works acceptably
- Stronger context effects between countries, rather than within, but conflated by data and classifier quality

Next steps

1. Extend outcome list + model selection
2. Complete leave-one-out approach
3. Alternative LLM use?

Thank you for your attention!

Time appendix: F1 scores comapred

		Political issue		Environment		Economy (issue only)	
		$F1_{base}$	$F1_{wo2014}$	$F1_{base}$	$F1_{wo2014}$	$F1_{base}$	$F1_{wo2014}$
DE14	Emb_{en}	0.419	0.466	0.429	0.311	0.552	0.615
	Emb_{orig}	0.343	0.427	0.286	0.274	0.250	0.528
	$Tf\text{-}idf_{en}$	0.517	0.515	0.308	0.493	0.522	0.667
ES14	Emb_{en}	0.512	0.586	0.471	0.431	0.618	0.662
	Emb_{orig}	0.573	0.471	0.250	0.243	0.760	0.538
	$Tf\text{-}idf_{en}$	0.455	0.560	0.154	0.360	0.750	0.669
UK14	Emb_{en}	0.615	0.596	0.385	0.361	0.465	0.402
	Emb_{orig}	0.553	0.607	0.571	0.373	0.432	0.392
	$Tf\text{-}idf_{en}$	0.531	0.605	0.519	0.373	0.488	0.536

* Models without country intercepts.

Countries appendix: F1 scores compared

		Political issue			Environment			Economy (issue only)		
		$F1_{base}$	$F1_{wo1}$	$F1_{wo2}$	$F1_{base}$	$F1_{wo1}$	$F1_{wo2}$	$F1_{base}$	$F1_{wo1}$	$F1_{wo2}$
FR19	Emb_{en}	0.829	0.767	0.758	0.709	0.444	0.448	0.709	0.588	0.579
	Emb_{orig}	0.854	0.412	0.409	0.753	0.332	0.348	0.719	0.373	0.389
	$Tf\text{-}idf_{en}$	0.814	0.783	0.783	0.749	0.585	0.589	0.698	0.648	0.651
NL19	Emb_{en}	0.747	0.699	0.714	0.612	0.481	0.536	0.739	0.680	0.694
	Emb_{orig}	0.732	0.306	0.323	0.534	0.312	0.354	0.756	0.475	0.524
	$Tf\text{-}idf_{en}$	0.734	0.704	0.704	0.634	0.555	0.559	0.749	0.714	0.728

*Models without country intercepts.

*$F1_{wo1}$ = all FR19 and NL19 is test, $F1_{wo2}$ = original FR19 and NL19 is test.