British Election Longitudinal News Study 2015–2019: Print news coverage with validated topics and candidate sentiment

7 April 2021

1. Citation

Horvath, L., Arabaghatta Basavaraj, K., Banducci, S., Jones, A., Kolpinskaya, E., Malla, R., Stevens, D. (2021). British Election Longitudinal News Study 2015–2019: Print News coverage with validated topics and candidate sentiment. [Data collection]. DOI: 10.24378/exe.3223

2. Overview

Along with the broadcast news study (to be released separately), the British Election Longitudinal News Study 2015–2019 (BELNS) covers campaign coverage relating to three general elections: 2015, 2017, 2019. This project has received funding from the Economic and Social Research Council, awarded to the 'Media in Context' projects at each election period ES/T015675/1, ES/R005087/1, ES/M010775/1.

The print newspaper component in this release tracks coverage across 46 national and local sources, see Section 3.

The **outlet-day level data** tracks topic coverage relating some of the 'most important issues' facing the country, as asked in all waves of the British Election Study Internet Panels 2014–2023 (BESIP, Fieldhouse et al. 2019; variable `mii`). The unit of analysis is the news source with repeated measures for each day during the study period, with variables corresponding to election period (GE2015, GE2017, GE2019), and topics.

Please note, on each day of observation, source 'The Times' appears twice, as a raw topic count as well as an adjusted topic count 'The Times ADJUSTED,' see Section 4.2.

In the **candidate data**, the unit of analysis is the candidate standing for election and the variables relate to: election period (GE2015, GE2017, GE2019), number of stories in which candidate was mentioned across all sources, as well as sentiment per source using different measures (See Section 5 for details).

BES data linkage. The BESIP wave that overlaps with our data collection period in 2019 is Wave 18. Along with an earlier version of the GE2019 media data, we released two Python scripts: [1] predicting, from the open-ended BESIP response, the broad issue category that respondents in Waves 17 & 18 indicated was most important to them; and [2] linking our media coverage data to BESIP Wave 18 on `paperLastThree_multiple`, the newspapers that respondents read in the three days prior to their interview date, `starttimeW18`. This method can be adapted to link with the new study waves. Available at https://mediaeffectsresearch.wordpress.com/research-output/ under 'Media in Context 2019 Data Pre-release, 30 November 2020.'

Linked 2019 data. The dataset produced using the above script is released again with the new topic variables and consists of the BESIP Wave 18 respondent identifiers, the predicted most important issue for each respondent, and their predicted exposure to eight issues in the newspapers they have read in the three days prior to their interview. For a detailed description of this data see documentation at https://mediaeffectsresearch.wordpress.com/research-output/ under 'Media in Context 2019 Data Pre-release, 30 November 2020.'

3. Corpus

Across the three study periods (GE2015, GE2017, GE2019) we queried, daily, Lexis UK for archived news stories using a keyword search as shown in Table 1.

Table 1. Archival search parameters

Keywords (inclusive)	election, candidate, poll, tories, tory, conservative party, the conservatives, ukip, uk independence party, labour party, green party, libdem, lib dem, liberal democrat, snp, scottish national party, dup, democratic unionist party, plaid cymru, change uk, brexit party + 2019 party leaders
Sources	The (Sunday) Times, The Independent, The Guardian, The Sun, The Mirror, The (Daily/Sunday) Telegraph, The Daily Mail and Mail on Sunday, Scotsman, The Evening Standard, The (Sunday) Express, Daily Record and Sunday Mail, Belfast Telegraph, The Western Mail (Mail on Sunday), The (Sunday) Herald, Western Daily Press, Aberdeen Press and Journal, Daily Star Online, Chronicle, Stoke The Sentinel, Daily Post, Yorkshire Post, The Sunday Telegraph, The Northern Echo, South Wales Echo, Liverpool Echo, Birmingham Evening Mail, Hull Daily Mail, Derby Telegraph, Nottingham Post, Manchester Evening News, Evening Times, Leicester Mercury, The Plymouth Herald, Grimsby Telegraph, Aberdeen Evening Express, Evening Gazette, East Anglian Daily Times, Eastern Daily Press, The People, Coventry Telegraph, The Sunday Express, The Observer, Birmingham Post, Gloucestershire Echo, The Sunday Herald, Evening News, Sunday Mercury, Evening Star, Scunthorpe Telegraph
Source N	46
Time frame	17 March 2015 to 14 May 2015 18 April 2017 to 15 June 2017 29 October 2019 to 19 December 2019
News story N	154,681

Based on this text corpus, we release topic and sentiment data.

4. Topics and validation

4.1 Method

We concentrate on linkages with the issues commonly mentioned in BESIP's "single most important issues" facing the country, and thus code in our media data:

- Europe,
- Economy,
- Environment,
- Crime
- Health,
- Immigration,
- Inequalities (predominantly economic inequalities such as poverty, housing, homelessness, cost of living),
- Terrorism,

- Negativity
- Ageing

To capture these, we used the data archive's narrow topic labels, which are high precision labels capturing the specific topic addressed in the news story, for example "air pollution" or "employment growth". After normalising these labels (snowball stemming), we found 6,736 unique topic labels. We then conducted an extensive keyword search to sort these into our broad topics in addition to manually coding the first 500 most frequent topic labels. We were able to identify 98,297 stories that contained a reference to at least one of our broad topics or 63%, the distribution shown in Table 2 below.

Table 2. Number of news stories with reference to topic (topics can overlap)

Election	Topic	Unadjusted count*	Adjusted count*	Rank within year
GE2015	Ageing	1235	1014	10
	Crime	5654	4099	3
	Economy	16739	13420	1
	Environment	2500	2130	8
	Europe	3978	3017	6
	Health	4478	3858	4
	Immigration	2610	2168	7
	Inequality	6529	5561	2
	Negativity	4211	2903	5
	Terrorism	1496	775	9
GE2017	Ageing	1574	1340	10
	Crime	7160	5331	3
	Economy	14290	11519	1
	Environment	2648	2221	8
	Europe	14101	11780	2
	Health	4837	4163	5
	Immigration	2594	2155	9
	Inequality	4431	3808	7
	Negativity	5149	4037	4
	Terrorism	4684	3494	6
GE2019	Ageing	1114	848	10
	Crime	7911	5063	4
	Economy	13508	9612	2
	Environment	3864	3008	7
	Europe	15694	12426	1
	Health	4871	3931	5
	Immigration	1555	1145	9
	Inequality	4602	3631	6
	Negativity	11249	7480	3
	Terrorism	2314	1413	8

See Section 4.2

4.2 Recommended adjustment for topic counts

Via Lexis Nexis, the Times and Sunday Times archival queries return systematically more stories than other comparable broadsheets. We investigated this by manually counting the number of stories in a set nine of print copies of the (Sunday) Times as well as the (Sunday) Telegraph for comparison. The number of digitally archived stories were in all cases higher than our calculation, on average by factor 1.63 for the Telegraph and by factor 4.42 for the Times. While we can treat the former as an acceptable and/or baseline

deviation reflecting differences in calculation methods, latter stands out as a characteristic of Times stories uniquely.

We calculated the adjusted topic counts on each day of observation as a separate source called 'The Times ADJUSTED.' The adjustment was made using the quotient of the two factors above, i.e. 4.42 / 1.63 = 2.71. We divided each raw topic count with this number and converted the result to its ceiling, to avoid 0s. We recommend that any analysis be run on this source instead of 'The Times' for a more conservative estimate of topic coverage. In Table 2 above, we show the full distribution of articles using both the raw and adjusted topic counts.

4.3 Validation

Three human coders cross-coded a random sample of 250 news stories from 2019 across six broad topics (a set of topics coded in the previous release), which we checked against our method of prediction as described in Section 4.1. Our results show high levels of agreement between human coders and machine classifications, and acceptable levels of inter-rater reliability.

Table 3: Validation results

Topic	Accuracy[1]	Cohen's κ
Europe	0.92	0.93
Economy	0.87	0.77
Health	0.96	0.84
Environment	0.96	0.81
Inequality	0.94	0.74
Immigration	0.97	0.85

[1]*Proportion of machine predictions that are identical to the median topic label across three human coders.*

5. Sentiment and validation

5.1 Method

We determined overall sentiment relying on the full text of the news stories, using two methods. First, we trained a binary sentiment classifier using the labelled NLTK Twitter sentiment dataset[1] and defined actor-level sentiment (general election candidates) as the predicted sentiment in the story containing given actor. In the candidate dataset, `prop_positive` is the proportion of positive stories across all stories mentioning the candidate during the general election period. Similarly, for each source e.g. `Bath_Chronicle_Positive` is the proportion of positive stories across all stories mentioning the candidate in the Bath Chronicle specifically.

Second, we predicted sentiment using the VADER sentiment dictionary[2], relying on the proportion of 'negative', 'positive' or 'neutral' words featured in the news story text. The actor-level sentiment is measured as the *relative proportion* of these valence categories in the story featuring the actor (general election candidate), expressed in a single 'compound score' variable ranging -1 (extreme negative) to 1 (extreme positive). In the candidate data, we draw on this compound score to express sentiment across all stories during the general election period for each source, using thresholding[3]. Stories with a compound score of less than -0.05 were counted as negative stories, those with a score larger than +0.05 as positive

[1] http://www.nltk.org/nltk_data/
[2] https://github.com/cjhutto/vaderSentiment
[3] Explanation on compound scores and thresholds suggested in documentation, see footnote above

stories, and scores in-between as neutral stories. Thus e.g. `Bath_Chronicle_VADER_Positive` is the proportion of positive news stories across all stories about the candidate in the Bath Chronicle.

5.2 Adjustment to baseline sentiment

Users of the VADER sentiment data may find it useful to adjust actor-level sentiment to baseline source sentiment. For example, a high proportion of positive articles about a particular actor might be associated with a source that is overwhelmingly positive in tone.

For each source, we drew a random sample of *N* = 500 and found the following sentiment proportions:

Table 4: Baseline sentiment per source

Source	Negatives	Neutral	Positives
Aberdeen Evening Express	0.46	0.01	0.51
Aberdeen Press and Journal	0.43	0.02	0.54
Bath Chronicle	0.22	0.00	0.77
Belfast Telegraph	0.36	0.00	0.63
Birmingham Evening Mail	0.30	0.03	0.66
Birmingham Post	0.22	0.01	0.76
Bristol Post	0.18	0.01	0.80
Daily Post (North Wales)	0.24	0.01	0.74
Daily Record and Sunday Mail	0.35	0.02	0.63
Daily Star Online	0.42	0.00	0.56
Derby Telegraph	0.26	0.01	0.71
East Anglian Daily Times	0.16	0.00	0.83
Eastern Daily Press	0.21	0.01	0.77
Evening Gazette	0.20	0.03	0.76
Evening News (Norwich)	0.23	0.00	0.75
Evening Star	0.12	0.00	0.87
Evening Times (Glasgow)	0.26	0.01	0.72
Exeter Express and Echo	0.20	0.00	0.79
Gloucestershire Echo	0.15	0.03	0.81
Grimsby Telegraph	0.22	0.01	0.76
Hull Daily Mail	0.26	0.00	0.72
Leicester Mercury	0.26	0.01	0.72
Liverpool Echo	0.29	0.01	0.70
Manchester Evening News	0.25	0.02	0.72
Nottingham Post	0.21	0.02	0.75
Scotsman	0.23	0.04	0.72
Scunthorpe Telegraph	0.11	0.03	0.85
South Wales Echo	0.25	0.02	0.72
South Wales Evening Post	0.28	0.02	0.69
Stoke The Sentinel	0.29	0.02	0.68
Sunday Mercury	0.30	0.02	0.67
Sunderland Echo	0.18	0.01	0.79
The Citizen Gloucester	0.15	0.02	0.82
The Daily Mail and Mail on Sunday	0.39	0.00	0.60

The Daily/Sunday Telegraph	0.31	0.01	0.67
The Evening Standard	0.30	0.00	0.69
The Express	0.34	0.00	0.65
The Guardian	0.26	0.00	0.73
The Herald	0.30	0.01	0.68
The Independent	0.40	0.00	0.59
The Mirror	0.44	0.02	0.53
The Northern Echo	0.24	0.00	0.75
The Observer	0.32	0.00	0.67
The People	0.43	0.01	0.55
The Plymouth Herald	0.27	0.01	0.71
The Sun	0.35	0.02	0.61
The Sunday Express	0.28	0.01	0.70
The Sunday Herald	0.30	0.00	0.69
The (Sunday) Times	0.36	0.00	0.63
The Western Mail	0.23	0.01	0.76
Torquay Herald Express	0.14	0.01	0.84
Wales on Sunday	0.27	0.01	0.71
Western Daily Press	0.30	0.00	0.69
Western Morning News	0.19	0.01	0.79
Yorkshire Evening Post	0.16	0.02	0.81
Yorkshire Post	0.20	0.02	0.77

5.3 Validation

We completed an initial validation exercise to evaluate the binary sentiment classifier's performance on our news data, with two independent coders. On a sample of 50 news stories, one of the coders was in 82% agreement with the binary classifier assigning the same sentiment label to 41 stories; while the other coder was in 78% agreement assigning the same label to 39 stories. The Cohen's κ inter-rater reliability is 0.63.