

British Election Longitudinal News Study 2016–2019: Broadcast news coverage with validated topics and candidate/actor sentiment

30 March 2022

1. Citation

Malla, R., Stevens, D., Horvath, L., Banducci, S., Jones, A., Kolpinskaya (2022). British Election Longitudinal News Study 2016–2019: Broadcast News coverage with validated topics and candidate sentiment. [Data Collection].

2. Overview

Along with the print news study (Horvath et al. 2021) the British Election Longitudinal Broadcast News Study 2016–2019 (BELBNS) covers campaign coverage relating to two general elections: 2017, 2019 and the EU Referendum in 2016. Unlike the BELNS data set (2021), we have not been able to access enough broadcast news content for GE2015 so we have replaced GE2015 with the Brexit referendum news coverage (EUREF2016). 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 broadcast news component in this release tracks coverage across 24 broadcast news 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 **mi**). The unit of analysis is the news source with repeated measures for each day during the study period, with variables corresponding to election period (EUREF2016, GE2017, GE2019), and topics.

In the **candidate data**, the unit of analysis is the candidate standing for election and the variables relate to: the election period (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). For EUREF, the unit of analysis is the political figure in the referendum and the variables relate to the referendum campaign.

3. Corpus

Across the three study periods (EUREF2016, GE2017, GE2019) we queried, daily, Box of Broadcasts for archived news stories using a keyword search as shown in Table 1.

Table 1. Archival search parameters

	'election* OR candidate* OR poll* OR "jeremy corbyn" OR "theresa may" OR "nicola sturgeon" OR "tim farron" OR "arlene foster" OR "gerry adams" OR "leanne wood" OR "caroline lucas" OR "jonathan berry" OR "paul nuttall" OR "tories" OR "tory" OR "conservative party" OR "the conservatives" OR "ukip" OR "uk independence party" OR "labour party" OR "green party" OR "libdem" OR "lib dem" OR "liberal democrat*" OR "snp" OR "scottish national party" OR "dup" OR "democratic unionist party" OR "plaid cymru"
	EUREF2016 referendum* OR poll* OR brexit* OR "european union" + David Cameron, George Osborne, Jeremy Corbyn, Alan Johnson, Tim Farron, Boris Johnson, Michael Gove, Nigel Farage, Gisella Stuart, Nicola Sturgeon, Priti Patel, Liz Kendall, Penny Mordaunt, Amber Rudd, Ruth Davidson, Angela Eagle, Kenneth Clarke, Philip Hammond, Douglas Carswell
Keywords (inclusive)	GE2017 election* OR candidate* OR poll* OR "ed miliband" OR "david cameron" OR "nicola sturgeon" OR "nick clegg" OR "peter robinson" OR "gerry adams" OR "leanne wood" OR "natalie bennett" OR "nigel farage" OR "tories" OR "tory" OR "conservative party" OR "the conservatives" OR "ukip" OR "uk independence party" OR "labour party" OR "green party" OR "libdem" OR "lib dem" OR "liberal democrat*" OR "snp" OR "scottish national party" OR "dup" OR "democratic unionist party" OR "plaid cymru"
	GE2019 election* OR candidate* OR poll* OR "tories" OR "tory" OR "conservative party" OR "the conservatives" OR "ukip" OR "uk independence party" OR "labour party" OR "green party" OR "libdem" OR "lib dem" OR "liberal democrat*" OR "snp" OR "scottish national party" OR "dup" OR "democratic unionist party" OR "plaid cymru" OR "change uk" OR "brexit party" + 2019 party leaders
Sources	BBC News 24 BBC Parliament BBC Scotland BBC1 Cambridgeshire BBC1 Channel Islands BBC1 East Midlands BBC1 London North East and Cumbria BBC1 North West BBC1 Scotland BBC1 South East

	BBC1 Wales
	BBC1 West
	BBC1 Yorks and Lincs
	BBC1 Yorkshire
	BBC2 England
	BBC2 Scotland
	BBC4
	Channel 4
	Channel 5
	ITV London
	ITV2
	More4
	Sky News
Source N	24
Time frame	3 May 2016 to 30 June 2016 18 April 2017 to 15 June 2017 7 November 2019 to 26 December 2019
Broadcasts	EUREF 362 GE2017 428 GE2019 487

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 mentioned in the BELNES (Horvath et al. 2021) single most important issues” facing the country, and thus code in our media data:

- Europe,
- Economy,
- Environment,
- Health,
- Immigration,
- Inequality

4.3 Validation

The first task is to predict the topics of the transcripts (broadcasts). For pre-processing the text, we have used SnowballStemmer by NLTK for stemming the text. We have then created a scikit-learn pipeline in which the whole text is vectorized using TfidfVectorizer followed by a multinomial logistic regression classifier with maximum iterations of 2000. The training data comes from the validated lexis dataset with labels and is passed into the learn pipeline created. The various performance metrics can be seen below:

Table 2: Classification Report

topic	precision	recall	f1_score	support
economy	0.43	0.34	0.38	208
environment	0.72	0.74	0.73	198
europe	0.48	0.52	0.5	189

health	0.61	0.59	0.6	199
immigration	0.64	0.59	0.62	202
inequality	0.64	0.63	0.64	195
None	0.65	0.8	0.72	209
accuracy			0.6	1400
macro avg	0.6	0.6	0.6	1400
weighted avg	0.6	0.6	0.6	1400

5. Sentiment Analysis

5.1 Method

We determined overall sentiment relying on the full text of the broadcast transcript, using two methods. First, we trained a binary sentiment classifier using the labelled NLTK Twitter sentiment dataset¹ 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. **BBC1 London_Positive** is the proportion of positive stories across all stories mentioning the candidate in the BBC London broadcasts specifically.

Second, we predicted sentiment using the VADER sentiment dictionary², 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 thresholding³. 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 stories, and scores in-between as neutral stories. Thus e.g. **BBC1 London_Vader_Positive** is the proportion of positive news stories across all stories about the candidate in the BBC London broadcasts.

References:

Horvath, L., Banducci, S., Kolpinskaya, E., Malla, R., Stevens, D. (2020). Media in Context and the 2019 General Election: Newspaper content (data pre-release). [Data collection]. Available from: <https://mediaeffectsresearch.wordpress.com/>

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