Media in Context Traditional Media Codebook

February 17, 2017

1 Data and methods

1.1 Data collection

The outlet-by-day dataset contains 135 variables and 2826 observations, corresponding to 27 media outlets - 17 newspapers, 9 TV and 1 radio. See variable `6 outlet_id` for the list of media outlets.

The analysed period is between 1 February 2015 and 30 May 2015.

Newspaper articles published between 1 February 2015 and 30 May 2015 were downloaded from the Nexis UK database. A subset of 11,000 articles were human-coded and labeled as either being about the UK General Elections or not. The labeled articles were used to train a supervised classifier in order to identify election articles in the rest of the corpus. A linear support vector machines classifier was trained using a stochastic gradient descent algorithm (F-score=0.95), which predicted 21,038 articles to be about the elections.

TV and radio news transcripts were downloaded using Box of Broadcats, the on-demand TV and radio service for education.

The total number of articles, TV and radio shows based on which this dataset was compiled is 22,947. These units were processed using quantiative text analysis techniques, and the results were aggregated at the outlet-day level.

1.2 Methods

Mentions of leaders and parties were measured using the party and leader names as keywords.

The tone, positive and negative emotions variables were measured using the psychological dictionaries developed by Pannebaker et al.(2015) and included in the Linguistic Inquiry and Word Count (LIWC2015) software.

The main topics were computed using a Latent Dirichlet Allocation (LDA) model (Blei et al. 2003; Griffiths and Steyvers, 2004). LDA is simple hierarchical Bayesian model of text based on the following assumptions: 1) each word in a text is exchangeable, each text in a corpus is a combination of a specific number of topics, and each specific topic is represented as a distribution of words over a fixed vocabulary. The generative structure that produces each document in a corpus is represented as random mixtures of latent topics and their associated distributions of words. We estimate the model using the sparse Gibbs sampler described in Yao et al. (2009) and the hyperparameter optimization routine utilized in Wallach et al. (2009).

Label	Keys
Business	govern busi energi power citi plan green compani work industri
Conservative	minist secretari tori cameron prime cabinet common parliament elect govern
Debates	debat leader cameron miliband bbc parti david question broadcast farag clegg
Economy	osborn economi econom govern budget chancellor elect growth market britain
Employment	work labour job tori countri vote live hour govern cameron
EU	cameron referendum britain vote govern european europ minist union power
ForeignAffairs	state islam govern countri syria presid report forc british britain
Green	vote parti elect polit green voter candid campaign poll gener
Housing	hous home buy properti rent london build council associ tenant
Immigration	immigr migrat rahman migrant elect britain net british power commun
Labour	labour parti leader leadership miliband union shadow elect secretari candid
LibDem	lib dem parti clegg liber nick democrat coalit seat tori
Media	block publish updat twitter cameron photograph guardian shapp conserv interview
NHS	health servic care women hospit patient fund doctor staff govern
Polls	poll labour cent tori seat parti vote conserv elect voter
Regions	labour seat candid vote london tori conserv elect west constitu
Schools	school wale educ welsh student govern univers plaid labour fee
SNP	snp scotland scottish labour sturgeon parti nicola vote leader murphi
SNPCoalition	labour snp govern parti miliband vote deal tori elect coalit
SocialIssues	famili life church children live sunday mother women father young
TaxSpend	tax cut spend plan labour billion budget pension incom benefit
UKIP	ukip farag parti nigel leader elect thanet south campaign support

2 Variable description

2.0.1 Metadata

1. day
   Day, integer 01-31.
2. month
   Month, integer 02-05.
3. date
   Date Stata format.
4. date_str
   Date, string.
5. weekday
   Day of the week, string, Monday-Sunday.
6. outlet_id
   MiC outlet id:

   - 1 Daily/Sunday Express
   - 2 Daily Mail/Mail on Sunday
   - 3 Daily Mirror/Sunday Mirror
   - 4 Daily Star/Daily Star Sunday
   - 5 Financial Times
   - 6 Guardian/Observer
   - 7 Independent/Independent on Sunday

- 9 The Sun
- 10 Times/Sunday Times
- 11 Birmingham Evening Mail
- 12 Daily Record/Sunday Mail
- 13 Evening Standard
- 14 Scotsman/Scotland on Sunday
- 15 Western Mail/Mail on Sunday
- 16 Western Morning News
- 17 BBC1 News at 10
- 18 Newsnight
- 19 ITV News at 10
- 20 Channel 4 News
- 21 Channel 5 News
- 22 Sky News at Ten
- 23 BBC Radio 4 Today
- 24 BBC London News
- 27 BBC Reporting Scotland
- 29 BBC Wales Today
- 71 Daily Telegraph/Sunday Telegraph
- 72 Yorkshire Evening Post

7. outlet_name
   Outlet name.
8. media_type
   Type of media (newspaper, TV, radio).
9. unit_of_analysis
   Unit of analysis.

2.0.2 Proportions of party and leader mentions by day

10. total_size
    Average number of words post-processing. Each unit (TV show, radio show or article) was processed by removing punctuation and common English stopwords. The number of words in clean units was averaged by outlet-day.
11. con_mention
    Proportion Conservative mentions out of variable `10 total_size`.
12. lab_mention
    Proportion Labour mentions out of variable `10 total_size`.
13. ld_mention
    Proportion Lib Dem mentions out of variable `10 total_size`.
14. snp_mention
    Proportion SNP mentions out of variable `10 total_size`.
15. pc_mention
    Proportion Plaid Cymru mentions out of variable `10 total_size`.
16. ukip_mention
    Proportion UKIP mentions out of variable `10 total_size`.
17. green_mention
    Proportion Green mentions out of variable `10 total_size`.

18. cameron
    Proportion Cameron mentions out of variable `10 total_size`.
19. miliband
    Proportion Miliband mentions out of variable `10 total_size`.
20. clegg
    Proportion Clegg mentions out of variable `10 total_size`.
21. sturgeon
    Proportion Sturgeon mentions out of variable `10 total_size`.
22. wood Proportion
    Wood mentions out of variable `10 total_size`.
23. farage Proportion
    Farage mentions out of variable `10 total_size`.
24. bennett Proportion
    Bennett mentions out of variable `10 total_size`.

2.0.3 Party and leader sentiment

25. words_overall
    Average number of words in the original text.

26. tone_overall
    Average overall tone, measured using LIWC2015.
27. posemo_overall
    Average positive emotion, measured using LIWC2015.
28. negemo_overall
    Average negative emotion, measured using LIWC2015.
29. con_words
    Average number of words in Conservative segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
30. con_tone
    Average tone (LIWC2015) in conservative segment.
31. con_posemo
    Average positive emotion (LIWC2015) in conservative segment.
32. con_negemo
    Average negative emotion (LIWC2015) in conservative segment.
33. lab_words
    Average number of words in Labour segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
34. lab_tone
    Average tone (LIWC2015) in segment.
35. lab_posemo
    Average positive emotion (LIWC2015) in segment.
36. lab_negemo
    Average negative emotion (LIWC2015) in segment.
37. libdem_words
    Average number of words in Lib Dem segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
38. libdem_tone

Average tone (LIWC2015) in Lib Dem segment.

39. libdem_posemo
    Average positive emotion (LIWC2015) in Lib Dem segment.
40. libdem_negemo
    Average negative emotion (LIWC2015) in Lib Dem segment.
41. snp_words
    Average number of words in SNP segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
42. snp_tone
    Average tone (LIWC2015) in SNP segment.
43. snp_posemo
    Average positive emotion (LIWC2015) in SNP segment.
44. snp_negemo
    Average negative emotion (LIWC2015) in SNP segment.
45. pc_words
    Average number of words in Plaid Cymru segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
46. pc_tone
    Average tone (LIWC2015) in Plaid Cymru segment.
47. pc_posemo
    Average positive emotion (LIWC2015) in Plaid Cymru segment.
48. pc_negemo
    Average negative emotion (LIWC2015) in Plaid Cymru segment.
49. ukip_words
    Average number of words in UKIP segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
50. ukip_tone
    Average tone (LIWC2015) in UKIP segment.
51. ukip_posemo
    Average positive emotion (LIWC2015) in UKIP segment.
52. ukip_negemo
    Average negative emotion (LIWC2015) in UKIP segment.
53. green_words
    Average number of words in Green segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
54. green_tone
    Average tone (LIWC2015) in Green segment.
55. green_posemo
    Average positive emotion (LIWC2015) in Green segment.
56. green_negemo
    Average negative emotion (LIWC2015) in Green segment.
57. cameron_words
    Average number of words in Cameron segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
58. cameron_tone
    Average tone (LIWC2015) in Cameron segment.
59. cameron_posemo
    Average positive emotion (LIWC2015) in Cameron segment.

60. cameron_negemo
    Average negative emotion (LIWC2015) in Cameron segment.
61. miliband_words
    Average number of words in Miliband segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
62. miliband_tone
    Average tone (LIWC2015) in Miliband segment.
63. miliband_posemo
    Average positive emotion (LIWC2015) in Miliband segment.
64. miliband_negemo
    Average negative emotion (LIWC2015) in Miliband segment.
65. clegg_words
    Average number of words in Clegg segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
66. clegg_tone
    Average tone (LIWC2015) in Clegg segment.
67. clegg_posemo
    Average positive emotion (LIWC2015) in Clegg segment.
68. clegg_negemo
    Average negative emotion (LIWC2015) in Clegg segment.
69. sturgeon_words
    Average number of words in Sturgeon segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
70. sturgeon_tone
    Average tone (LIWC2015) in Sturgeon segment.
71. sturgeon_posemo
    Average positive emotion (LIWC2015) in Sturgeon segment.
72. sturgeon_negemo
    Average negative emotion (LIWC2015) in Sturgeon segment.
73. wood_words
    Average number of words in Wood segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
74. wood_tone
    Average tone (LIWC2015) in Wood segment.
75. wood_posemo
    Average positive emotion (LIWC2015) in Wood segment.
76. wood_negemo
    Average negative emotion (LIWC2015) in Wood segment.
77. farage_words
    Average number of words in Farage segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.
78. farage_tone
    Average tone (LIWC2015) in Farage segment.
79. farage_posemo
    Average positive emotion (LIWC2015) in Farage segment.
80. farage_negemo
    Average negative emotion (LIWC2015) in Farage segment.
81. bennett_words

Average number of words in Bennett segments. Segments were extracted out of the clean text by taking the 20 words before and after the party/leader mention.

82. bennett_tone
    Average tone (LIWC2015) in Bennett segment.
83. bennett_posemo
    Average positive emotion (LIWC2015) in Bennett segment.
84. bennett_negemo
    Average negative emotion (LIWC2015) in Bennett segment.

2.0.4 Sums of leader and party mentions by day

85. total_units
    Total units (articles, TV or radio shows) per day.
86. con_units
    Total units mentioning Conservatives.
87. lab_units
    Total units mentioning Labour.
88. ld_units
    Total units mentioning Lib Dems.
89. snp_units
    Total units mentioning SNP.
90. pc_units
    Total units mentioning Plaid Cymru.
91. ukip_units
    Total units mentioning UKIP.
92. green_units
    Total units mentioning the Greens.
93. cameron_units
    Total units mentioning Cameron.
94. miliband_units
    Total units mentioning Miliband.
95. clegg_units
    Total units mentioning Clegg.
96. sturgeon_units
    Total units mentioning Sturgeon.
97. wood_units
    Total units mentioning Wood.
98. farage_units
    Total units mentioning Farage.
99. bennett_units
    Total units mentioning Bennett.
100. sum_words_day
     Sum of the number of words in the clean text for all units (articles, TV and radio shows) per day.
101. con_ment
     Total number of times Conservatives mentioned in an outlet per day.
102. lab_ment
     Total number of times Labour mentioned in an outlet per day.

103. ld_ment
Total number of times Lib Dems mentioned in an outlet per day.
104. snp_ment
Total number of times SNP mentioned in an outlet per day.
105. pc_ment
Total number of times Plaid Cymru mentioned in an outlet per day.
106. ukip_ment
Total number of times UKIP mentioned in an outlet per day.
107. green_ment
Total number of times Greens mentioned in an outlet per day.
108. cameron_ment
Total number of times Cameron mentioned in an outlet per day.
109. miliband_ment
Total number of times Miliband mentioned in an outlet per day.
110. clegg_ment
Total number of times Clegg mentioned in an outlet per day.
111. sturgeon_ment
Total number of times Sturgeon mentioned in an outlet per day.
112. wood_ment
Total number of times Wood mentioned in an outlet per day.
113. farage_ment
Total number of times Farage mentioned in an outlet per day.
114. bennett_ment
Total number of times Bennett mentioned in an outlet per day.

2.0.5 Topics

115. topic_conservatives
Average Conservative topic proportion per day.
116. topic_labour
Average Labour topic proportion per day.
117. topic_libdem
Average Lib Dem topic proportion per day.
118. topic_snp
Average SNP topic proportion per day.
119. topic_ukip
Average UKIP topic proportion per day.
120. topic_green
Average Green topic proportion per day.
121. topic_business
Average Business topic proportion per day.
122. topic_nhs
Average NHS topic proportion per day.
123. topic_taxspend
Average Taxation and Spending topic proportion per day.
124. topic_housing
Average Housing topic proportion per day.
125. topic_polls

Average Polls topic proportion per day.

126. topic_regions
Average Regions topic proportion per day.
127. topic_socialissues
Average Social Issues topic proportion per day.
128. topic_economy
Average Economy topic proportion per day.
129. topic_debates
Average Debates topic proportion per day.
130. topic_eu
Average EU topic proportion per day.
131. topic_media
Average Media topic proportion per day.
132. topic_snpcoalition
Average SNP Coalition topic proportion per day.
133. topic_schools
Average Schools topic proportion per day.
134. topic_employment
Average Employment topic proportion per day.
135. topic_foreign
Average Foreign Policy topic proportion per day.