Supplemental Material: Topic Diversity in Political Ads and Posts on Social Media – A Study of the 2022 Australian Federal Election Campaign

Methods

Data collection and preprocessing

Data cleaning. We started cleaning the ads data set. Fields containing text (title, link description, and body) were transferred into one column for further processing. We than added a column for party and candidate information and matched our senders list with the ads text list to fill in party and candidate data. The set with empty sender information was derived and manually coded again. Those ads were gained from additional senders of ads from the original list of organic postings. We than again added the sender information to the data set for the full list of text, party, and candidate including IDs for later identification of the documents. Then, html codes, urls, special characters, numbers, and stop words were removed from the text fields and rows with less than three words were excluded from further processing. We repeated this process of data cleaning for the post data. The only difference was that because the pageIDs from the meta ads library did not match the CrowdTangle ID we matched our sender information (party and candidate) based on page names. Also, to include only those senders that also had used ads during the campaign, we excluded those documents that were not assigned the party and candidate information from the full list of the ads' senders.

Topic modelling. To find the optimal number of topics for our data sets as described in the main article we first created a dtm (allowing unigrams and bigrams), excluding terms that occurred less than twice and plotted the coherence score for 50 topics including all relevant peaks for 16, 20, 24, and 28 topics. (see figure 1).

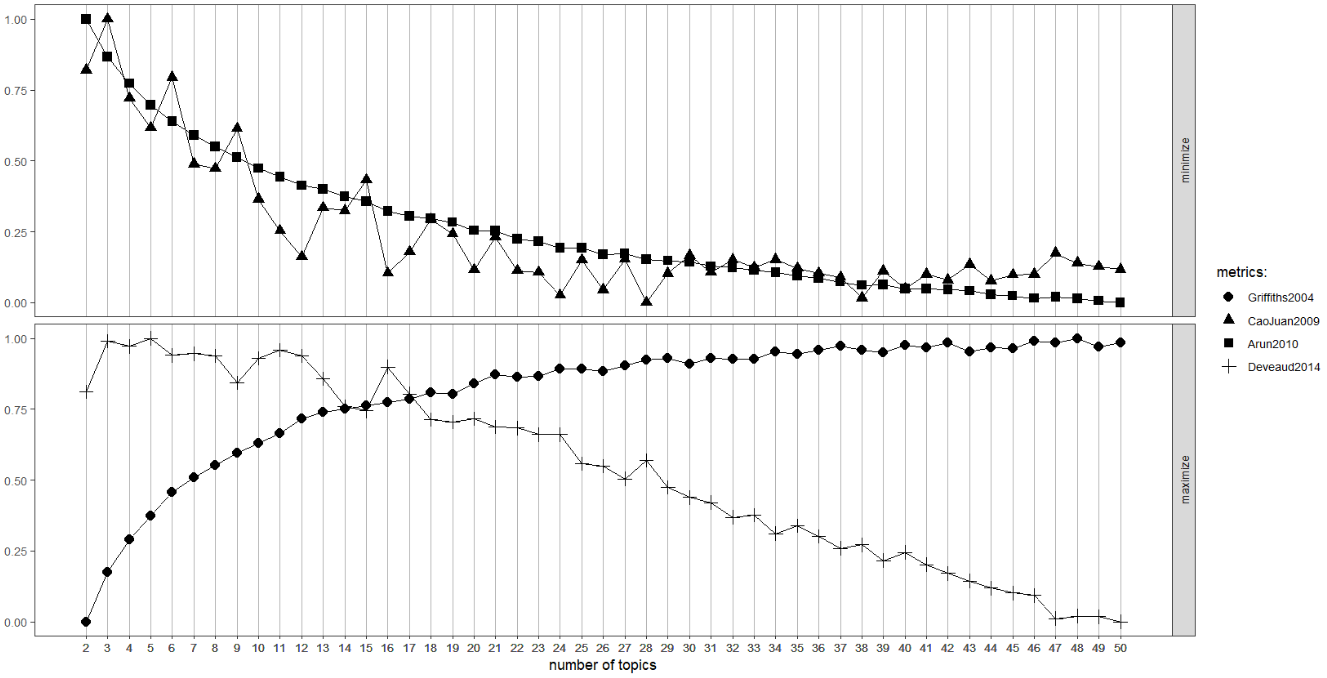

Figure 1. Plotted coherence score for ads data up to 50 topics (k).

The same was done for the posts data, in this case showing the coherence for 50 potential topics (see figure 2).

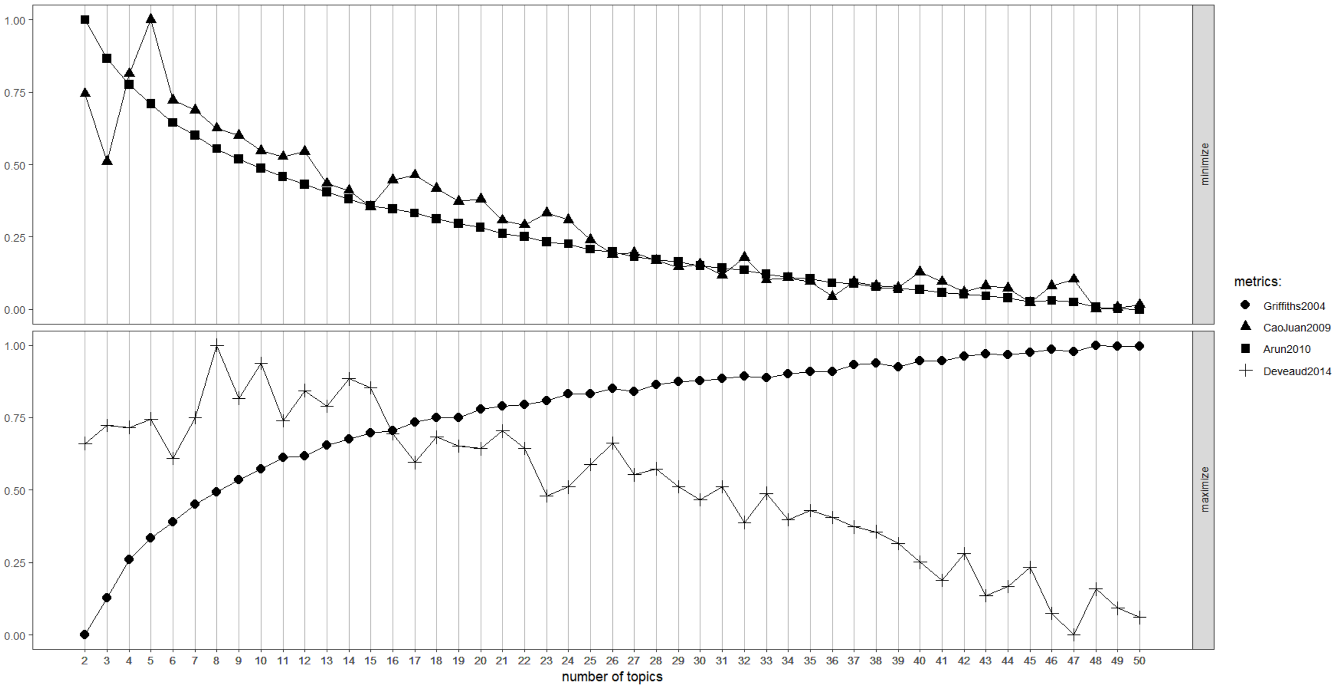

Figure 2. Plotted coherence score for posts data up to 50 topics (k).