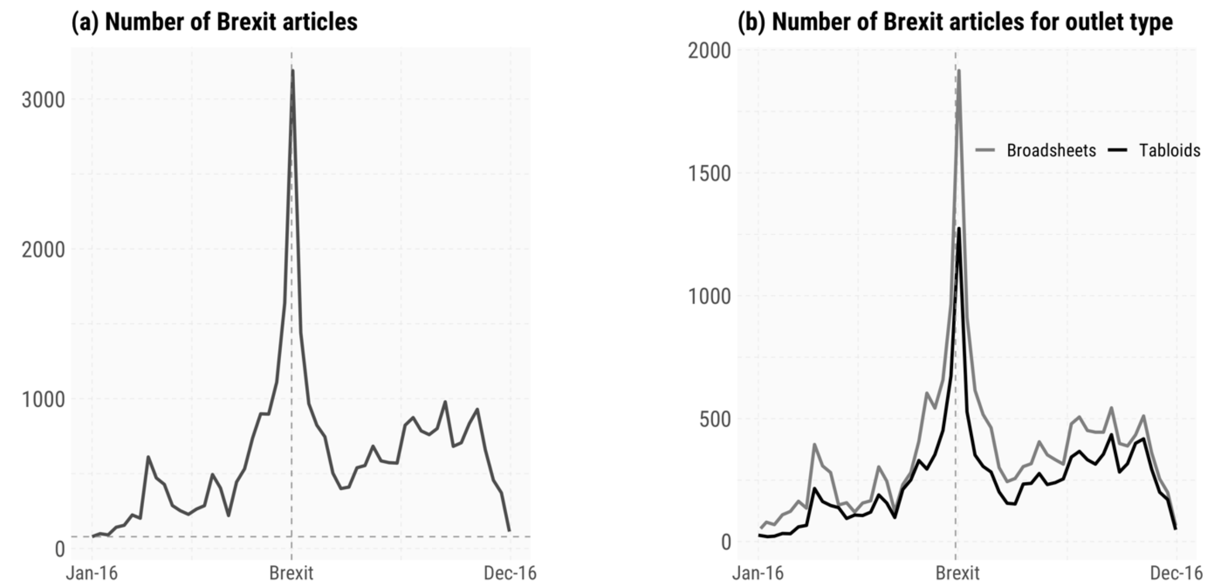
Online Supplementary Information

Supplementary materials intended for online publication: *Similar Citizen Portrayals? Converging Media-Based Othering in Tabloids and Broadsheets*

Supplementary Information 1: Brexit coverage, descriptive statistics, and search updates

Figure A1.1: Brexit coverage: centered around 01/01/2018 (to be 0)



Additional search updates based on qualitative inspection

We included as pertinent some specific unigrams, such as plural of specific nationalities without citizen mentions (e.g., “Hungarians”) and manually checked for false positives especially in cases where these do not necessarily refer to people, such as French or Dutch, often referring to businesses or products. In addition, we updated the list of pertinent hits by including a combination of the previously “second terms” followed by “from/of” or “from the/of the” and country of provenience listed for EU or U.K. citizens, so that combinations such as “citizens from the EU” count as EU citizen mentions. We are restrictive in terms of how far after the prepositions the locations can appear in order to avoid false positives and assure that people/citizen mentions have a close mention of origin/provenience. For all cases and combinations, we removed all negations, such as “non-EU citizens” for example in multiple different forms from the EU citizen and immigrant group and made sure they are in the Immigrant group, as non-EU immigrants.

Upon further qualitative inspection of contexts, we also apply a transformation for cases where we have for example “British immigrants” or “British emigrants/expats” relabeling them as “U.K. citizens”. Like our initial search, if we find any immigrant mention that is followed by “from/of/from the/of the” and an origin from the EU list, we place these in the EU citizen group since they are EU immigrants.

Table A1.1: Detailed data summary

	Number of articles	Articles pre/post referendum	Type	Platform	Total EU mention	Total Immigrant mention	Total UK citizen mention
The Independent	5093	1080/4013	Broadsheet	Paper	2068	1116	2028
The Guardian	3475	1613/1862	Broadsheet	Paper+Online	1827	1318	1525
telegraph.co.uk	3270	1240/2030	Broadsheet	Online	1269	1139	2156
FT.com	2674	912/1762	Broadsheet	Online	411	268	370
The Daily Telegraph (London)	1446	452/994	Broadsheet	Paper	309	283	425
The Times	1342	357/985	Broadsheet	Paper	306	196	242
Financial Times	728	187/541	Broadsheet	Paper	245	149	154
thetimes.co.uk	584	156/428	Broadsheet	Online	154	100	121
Independent.co.uk	332	332/0	Broadsheet	Online	143	144	128
The Sunday Times	293	135/158	Broadsheet	Paper	78	49	67
The Sunday Telegraph (London)	267	87/180	Broadsheet	Paper	45	47	104
The Observer	67	59/8	Broadsheet	Paper	37	23	38
Express Online	4564	1211/3353	Tabloid	Online	1859	1656	2964
Mail Online	3402	1128/2274	Tabloid	Online	1508	1300	2169
mirror.co.uk	1942	784/1158	Tabloid	Online	446	457	788
thesun.co.uk	1247	401/846	Tabloid	Online	306	259	545
Daily Mirror	637	154/483	Tabloid	Paper	72	55	124
The Sun	567	184/383	Tabloid	Paper	54	73	114
The Express	397	119/278	Tabloid	Paper	75	52	132
Daily Mail (London)	380	118/262	Tabloid	Paper	84	86	123
Mail on Sunday (London)	99	31/68	Tabloid	Paper	15	22	27
Sunday Express	67	18/49	Tabloid	Paper	3	7	14
Sunday Mirror	56	17/39	Tabloid	Paper	2	4	20
Sunday Sun	17	8/9	Tabloid	Paper	2	4	5

Figure A1.2: Salience of citizen mentions

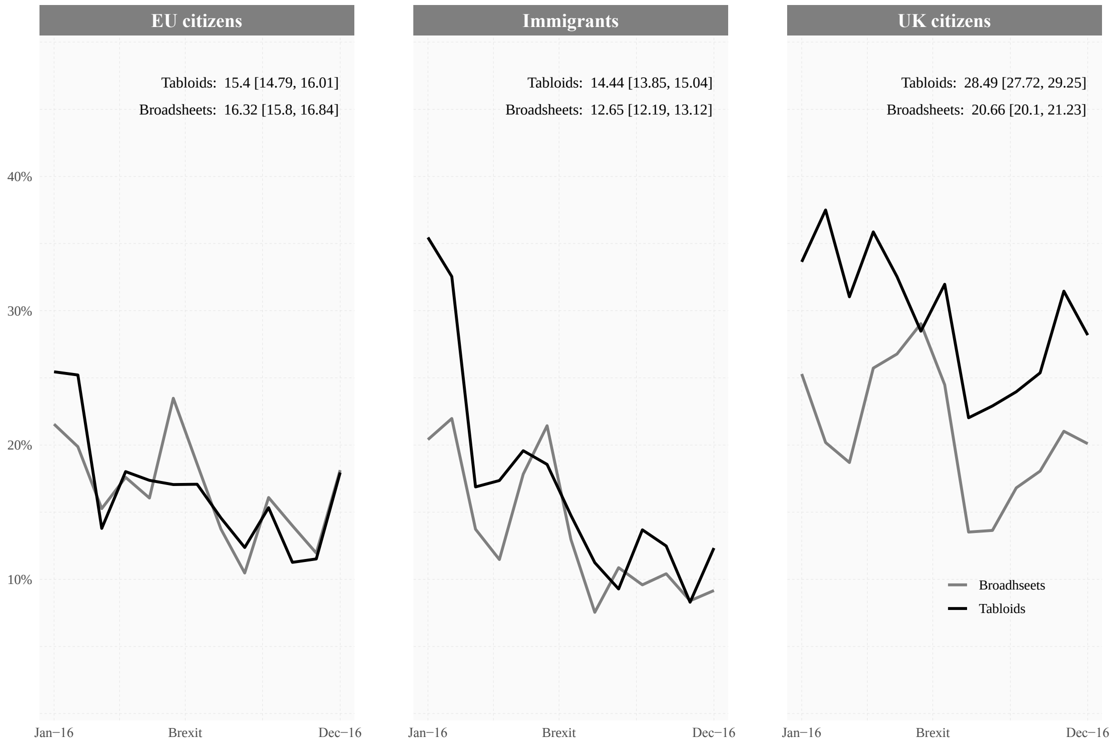
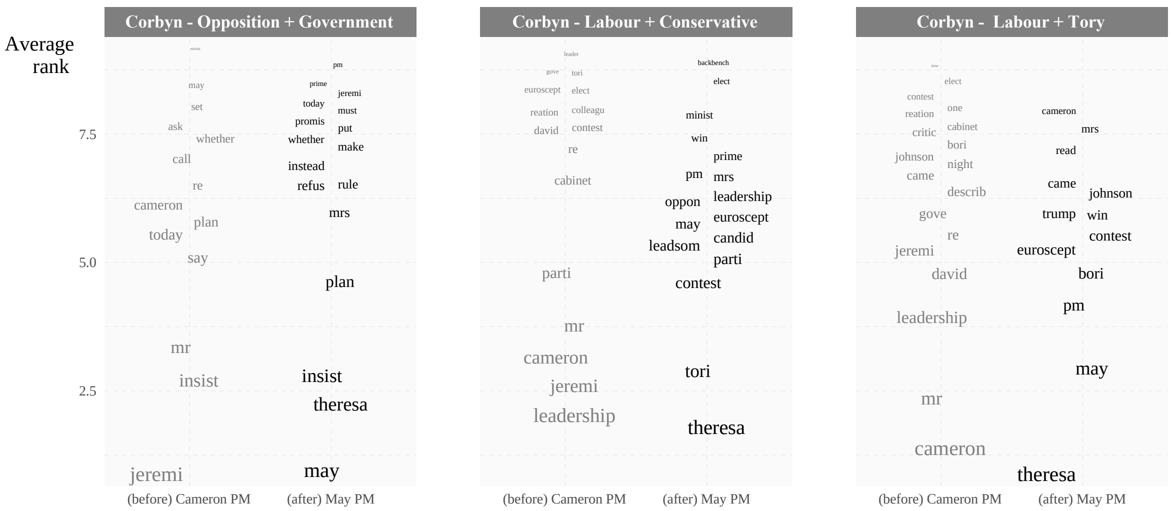


Figure A1.2 shows the proportion of the Brexit-related articles in broadsheets and tabloids mentioning immigrants, EU or U.K. citizens at least once. Consistent for both outlet types, the ingroup of U.K. citizens was the one that received most attention before and after the referendum. Furthermore, tabloids mentioned people significantly more often, especially immigrants, but the largest difference exists for references to the ingroup of U.K. citizens. When taking mentions of EU citizens and (non-EU) migrants together, they received more attention than the ingroup, but this difference is negligible in tabloids, whereas quite sizeable in broadsheets. Citizen and immigrant mentions were most frequent at the beginning of 2016, before and around the official announcement that a referendum will take place.

Supplementary Information 2: Custom analogy-based evaluation

Similar to traditional evaluation tasks, we use knowledge about political actors and their political relationship to check whether our model picks up on meaningful lexical relationships. We start with (Jeremy) Corbyn about whom we know that he was the leader of the Labour party and also leader of the opposition. Accordingly, changing these features can be included in various checks of lexical relationships where the aim is to recover the closest terms to, for example: v(“corbyn”) − v(“opposition”) + v(“government”). In other words, starting from Jeremy Corbyn as the Labour party leader or the opposition leader, then substituting the party from Labour to Conservative or the opposition to government, we should ideally get as a result something similar to the Prime Minister or the leader of the Conservative party. Furthermore, in July 2016, David Cameron resigned, hence, we will treat the before and after resignation periods separate, and this should influence the outcome of our exercise. Using the above-mentioned expressions or word operations, we can generate a hypothetical word vector and then we search for existing words in terms of closest in word vector.

Figure A2.1: Validation



Notes: GloVe, 100 dimensions, window of 6.
Size inverse proportional to rank. Constitutive terms ommitted from list.

As displayed in Figure A2.1, where we list the resulting top ranked terms, our model generates word embeddings with good face validity. Altering the opposition status to government status (leftmost panel), the top terms for Corbyn will contain the actual prime minister at the time. When we change the party from Labour to Conservative or Tory, we also see Cameron among the top results prior to his resignation as Prime Minister. In the post-resignation period, indeed, Theresa May appears highly ranked, and the other close results are prominent members of the Conservative party. Given that David Cameron continued to be a Conservative MP till September, it is not surprising that he appears even in the post-resignation period, however not in the case where we focus on the opposition vs government status.

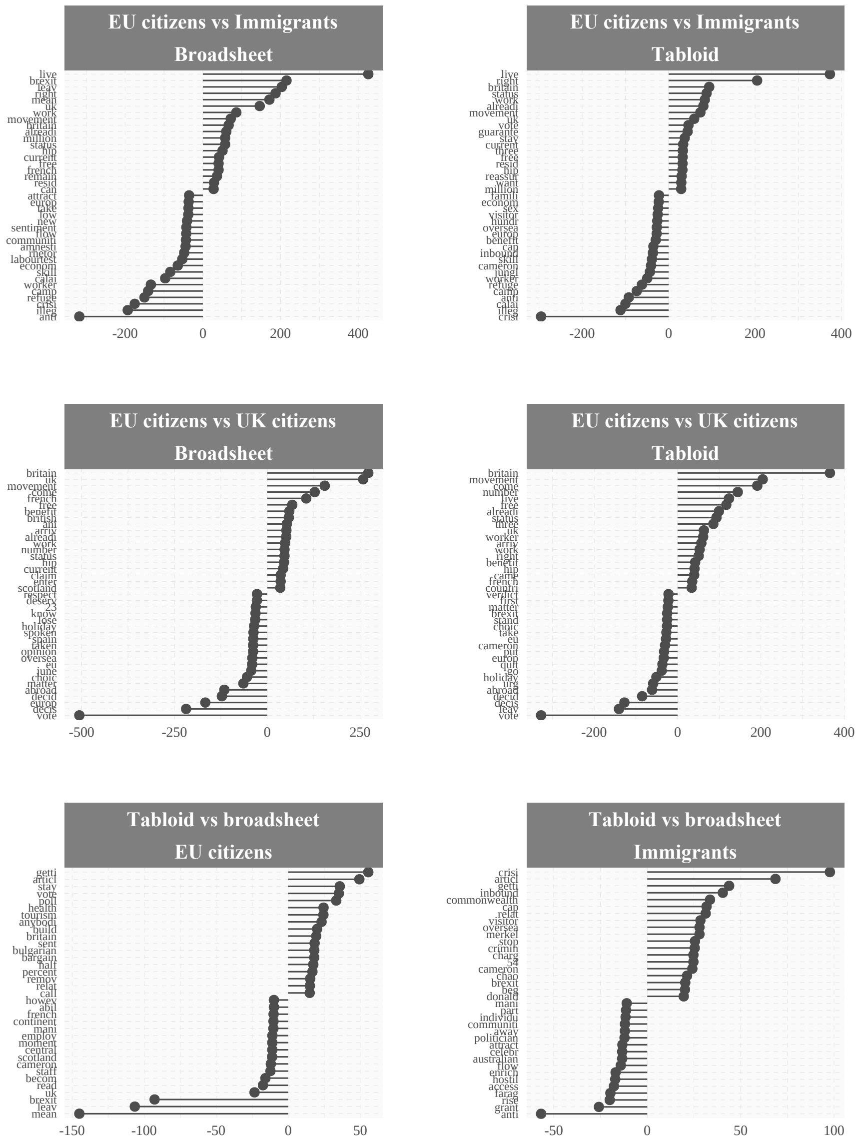
Supplementary Information 3: Context word comparisons

In addition to the model-based summaries, we can return to our text data and analyze the context or neighboring words for each of the citizen mentions. We extract three words before and after each occurrence of the citizen mentions and compare the frequency of these words, by relying on keyness representing relative frequencies. In Figure A3.1 we report a set of comparisons expressed through the chi-squared value of the word frequency comparisons, with higher values reflecting substantially and significantly more frequent use for the first element of the panel title. For example, in the first panel, we compare within all broadsheets the frequency of context words appearing around EU citizens *vs*. those around non-EU migrants, where: “live” is much more frequently use for EU citizens (first element of the title), and “anti” or “illeg” are much more frequently used around migrants.

We develop on a few selected comparisons that we deem important for better understanding outgroup representation nuances with special interest in EU citizens, but also between outlet differences. Accordingly, the first two rows show within-outlet type comparisons of different citizen groups. What sets apart EU citizens from migrants in both tabloids and broadsheets is a formulation of living, residing, and working rights in general *vs* benefit receiving, crisis or refugee related context (for non-EU or general migrants). Much of the U.K. citizen specificity in comparison to EU citizens is related to them being the group who will actually vote (decide) and references to potential implications for U.K. citizens being or going abroad.

We have seen convergence between outlet types, and the context term comparisons from the last row of Figure A3.1 reveal only minor differences as well but can help in understanding the broader framework of media representation and othering. Regarding EU citizens and immigrants, tabloids are slightly more interested in emphasizing the idea of health care tourism and refer to some Central Eastern European countries, whereas broadsheets are mentioning Western European (French) citizens more often in this context. Influx and criminality consideration are more frequent in the tabloids' immigrant mentions, whereas discussion of anti-immigrant policies and politicians appears to be slightly more frequent in broadsheets.

Figure A3.1: Context comparison (zoom for content)[1]

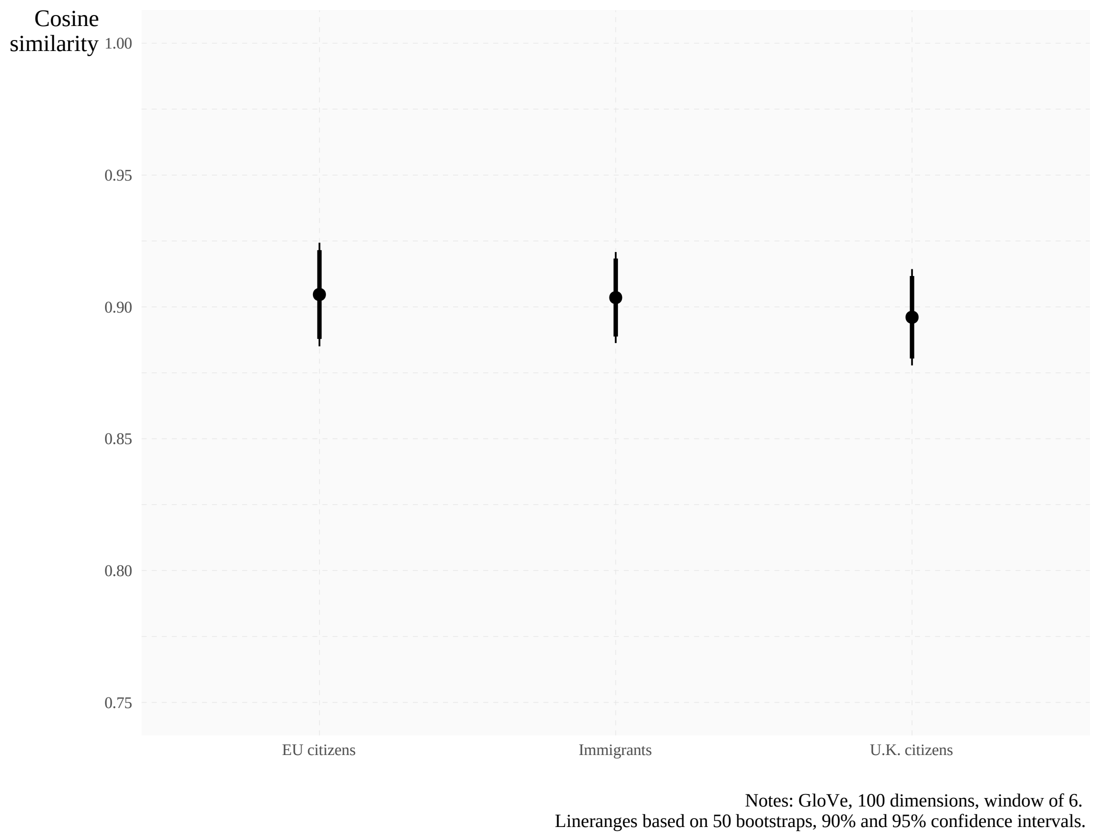


Notes: Terms use in the +/- 3 word window. Other citizen mentions ommitted from list.
Chi-squared measure for text keyness, with positive numbers reflecting use in first element of panel title.

[1] In these figures we highlight differences in shared terms, omitting those terms that appear to a similar extent in both contexts, and terms used exclusively in one context.

Supplementary Information 4: Between-outlet type cosine similarity of citizen groups

Figure A4.1: Between-outlet type cosine similarity of citizen groups



Supplementary Information 5: Additional information on robustness checks

Word embedding models perform best on large corpora, which for media coverage analysis of specific events is not (always) achievable (although see longitudinal news coverage analysis by Kroon et al., 2019). We have introduced a specific evaluation and validation task and rely throughout our paper on uncertainty measures derived by bootstrap. Furthermore, while the corpus is specific, and we do not have that many unique words and documents in our corpus, the terms of interest appear sufficiently frequently in our data. In our validation exercise that performed well, we have for example 11,782 for Corbyn before Brexit mentions, or 13,014 occurrences of Conservatives overall. Given the specific interest in these terms and bigram (or trigram) based approach built, we cannot rely on input from pre-trained embeddings (Rodriguez & Spirling, 2021) or a la carte embedding (Khodak et al., 2018; Rodriguez et al., 2021), but we are confident that we have enough occurrences that our method selection does not contribute to any bias (see for comparison Rodriguez et al., 2021).

Nevertheless, we also report and discuss additional checks in the paper. Figure A.3.1 summarizes the results from 4 different robustness checks. Each panel displays the cosine similarity between a pair of citizen mentions, with colors representing the outlet types. Along our main results reported above as leftmost entries in each panel, the different shaped “dots” are cosine similarities from alternative models, which are also displayed on the x-axis labels.

The robustness tests and the motivation behind these are described in detail in the main paper. Two additional details are relevant for the context window and the reduced overlap models. A narrower context window (3 words) can mean that the similarity scores will be slightly more based on syntactic, rather than semantic similarity. More importantly, we wanted to assure our main choice of context (6 words) is not too wide. We also report results based on a more restricted article subset, because journalists might share source documents and know from research on intermedia agenda-setting that news outlets also influence each other's coverage. This could lead to an inflated similarity score, if a large portion of the materials is shared or quoted across different outlet types especially regarding the content the mentions the citizen groups of interest. Thus, we compare each article from a broadsheet to each article from tabloids within one week and keep only those where at most 30% of the content overlaps, and then fit our embedding model.

Figure A.3.1: Robustness tests