Supplementary Material for Manuscript:

Using word embeddings to probe sentiment associations of politically loaded terms in news and opinion articles from news media outlets

David Rozado*, Musa al-Gharbi

*Corresponding autor: david.rozado@op.ac.nz

News articles and bias ratings from 47 news media outlets

The list of 47 news outlets analyzed in this paper as well as their human ratings of political bias were taken from the AllSides 2019 media bias chart v1.1 [1], see Figure S 1.

Figure S 1 The list of outlets analyzed in this paper as well as their human ratings of political bias were taken from the chart above produced by the AllSides organization. Printed with permission from http://www.allsides.com

The temporal coverage of articles availability in different news outlets is not uniform. For most media organizations, news articles availability in online domains or Internet cache copies becomes sparse as a function of article' age. This is not the case for some news outlets, where news articles

availability is excellent for articles as far back as the 1970s. Figure S 2 illustrates the time ranges of article data analyzed based on news outlets articles online availability.

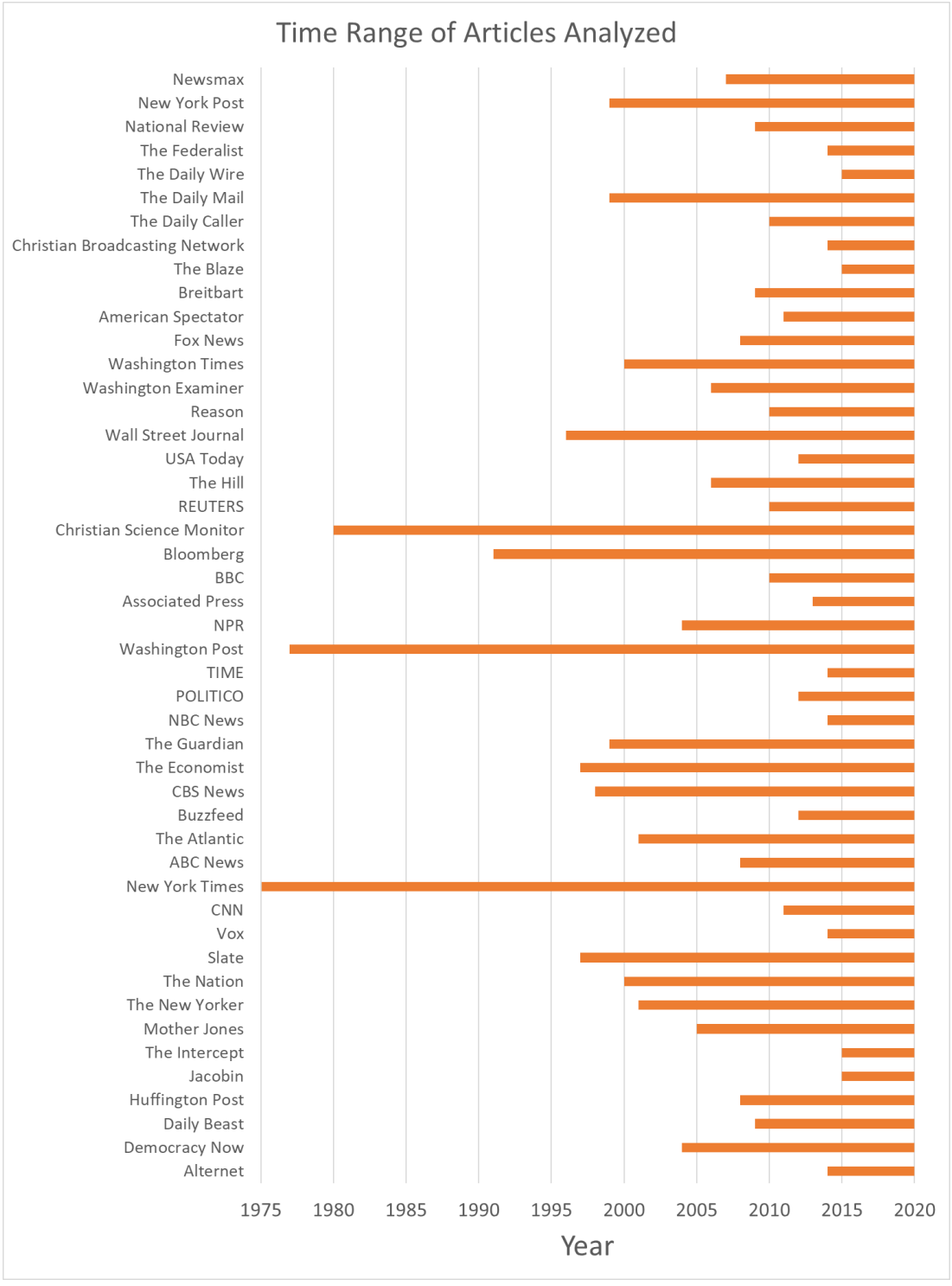

Figure S 2 Time range of articles analyzed based on articles online availability

Word embeddings

The success of word embeddings in language modelling emerges from their ability to map the statistical cooccurrence of words and their contexts in a training corpus into locations in vector space that represent the semantic and syntactic roles with which the words are used in the corpus. Thus, embedding spaces contain semantically meaningful spatial structure such that words with similar meaning, *cat* and *kitten* for instance, tend to be located in adjacent regions of embedding space.

Adjacency in vector space not only captures semantic similarity but also associations between word pairs. That is, words that tend to cooccur in similar contexts, even if they are not semantically interchangeable, are also positioned in nearby regions of vector space. For instance, related word pairs such as *car* and *fuel,* albeit not semantically swappable, tend to be adjacent in embedding models due to their relatedness. This feature has been shown to be useful for the sociological analysis of cultural associations contained in large text corpora [2].

Validating the quality of word embedding models derived from news outlets corpora

The accuracy of the embedding models trained on individual news media outlets corpora was validated by comparing their performance on similarity, association and word analogy tasks with well-known and popular word embedding models pre-trained on large corpora such as Wikipedia, Twitter or Common Crawl. The ability of each model to capture semantic similarity, relatedness (i.e. association) as well as morphological, lexical, encyclopedic and lexicographic analogies [3] was measured, see Table S 1.

The performance of the embeddings derived from individual news outlets was roughly similar to several popular pre-trained embedding models trained on large corpora such as Twitter or Google books. Performance of embeddings trained on news outlet-specific articles was slightly worse than some famous pre-trained embedding models such as word2vec trained on Google News. This is due to said popular pre-trained embedding models being trained on corpora at least 2 orders of magnitude larger in size than the individual news outlets corpora used in this work. The FastText model trained on Common Crawl outperformed all other embedding models probably due to its large training corpus size and its ability to model morphological relationships at the subword level.

	Popular pre-trained embedding models						Embedding models trained on news outlets articles							
Model index	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Word embedding algorithm	word2vec (Skip-Gram)	word2vec (Skip-Gram)	Glove	word2vec (Skip-Gram)	Glove	fastext	word2vec (CBOW)	word2vec (CBOW)	word2vec (CBOW)	word2vec (CBOW)	word2vec (CBOW)	word2vec (CBOW)	word2vec (CBOW)	word2vec (CBOW)
Vector dimensions	300	300	300	300	200	300	300	300	300	300	300	300	300	300
Training corpus name	Google News	Google Books N-grams 1990s	Wikipedia + Gigaword	Gigaword	Twitter	Common Crawl	New York Times (2015- 2019)	New York Times (1975- 1979)	Fox News (2015- 2019)	Breitbart (2015- 2019)	Reuters (2015- 2019)	BBC (2015- 2019)	Slate (2015- 2019)	Daily Mail (2015- 2019)
Corpus size in number of tokens	100B	NA	6B	NA	27B	600B	255M	230M	188M	100M	305M	143M	50M	838M
Model vocabulary size	3M	100K	400K	292K	1.2M	2M	200K	190K	184K	100K	191K	129K	82K	310K
WordSim-353	0.62	0.64	0.60	0.57	0.54	0.66	0.63	0.61	0.60	0.55	0.54	0.57	0.59	0.64
MEN simiarity dataset	0.68	0.68	0.74	0.59	0.61	0.71	0.71	0.70	0.64	0.54	0.50	0.66	0.60	0.73
SimLex-999	0.45	0.33	0.39	0.39	0.15	0.46	0.42	0.42	0.37	0.28	0.32	0.34	0.32	0.44
Google Semantic analogies	0.75	0.44	0.78	0.67	0.50	0.88	0.65	0.47	0.66	0.60	0.62	0.61	0.26	0.61
Google Syntactic analogies	0.74	0.39	0.67	0.67	0.60	0.84	0.60	0.57	0.56	0.48	0.46	0.55	0.41	0.62
BATS1 Inflectional Morphology analogies	0.68	0.36	0.60	0.60	0.51	0.85	0.50	0.46	0.44	0.41	0.39	0.41	0.40	0.48
BATS2 Derivational Morphology analogies	0.17	0.05	0.09	0.11	0.08	0.32	0.07	0.09	0.08	0.07	0.06	0.08	0.05	0.08
BATS3 Encyclopedic Semantics analogies	0.21	0.15	0.25	0.20	0.18	0.30	0.16	0.13	0.14	0.14	0.14	0.18	0.07	0.16
BATS4 Lexicographic Semantics analogies	0.06	0.08	0.07	0.05	0.07	0.10	0.04	0.05	0.04	0.02	0.03	0.03	0.02	0.05
AVERAGE	**0.48**	**0.35**	**0.47**	**0.43**	**0.36**	**0.57**	**0.42**	**0.39**	**0.39**	**0.34**	**0.34**	**0.38**	**0.30**	**0.42**

Table S 1 Comparison of performance between popular pre-trained embedding models (green) and word embedding models trained on outlet-specific news articles (yellow) across commonly used validation metrics in NLP.

Cultural axes and word vectors projections

A gender axis in an embedding space derived from a culturally representative corpus corresponds closely with the cultural category of the gender axis as humans use it in everyday language. That is, most humans have an intuitive understanding that words such as *necktie*, *T-shirt* and *skirt* fall in distinct locations of the gender axis as a result of the statistical frequencies with which those words are used around terms that describe males and females in everyday language. Word embedding models are able to capture these natural language regularities into vector space. Subsequently, vector algebra operations can be used to probe the embedding space in search of cultural associations latent in the corpus on which the embedding model was trained. Thus, by systematically projecting words onto a cultural axis, we can measure the connotations ingrained within the culture that produced the texts, see Figure S 3.

Usually, averaging several related vectors prior to estimating cultural dimensions improves the reliability of cultural axes estimation. That is, averaging the vector representations of *man* and *men* creates a better estimation of the maleness pole within the gender axis than either term in isolation. Similarly, the femaleness pole of the axis can be better estimated by averaging the vector representations of the words *woman* and *women*.

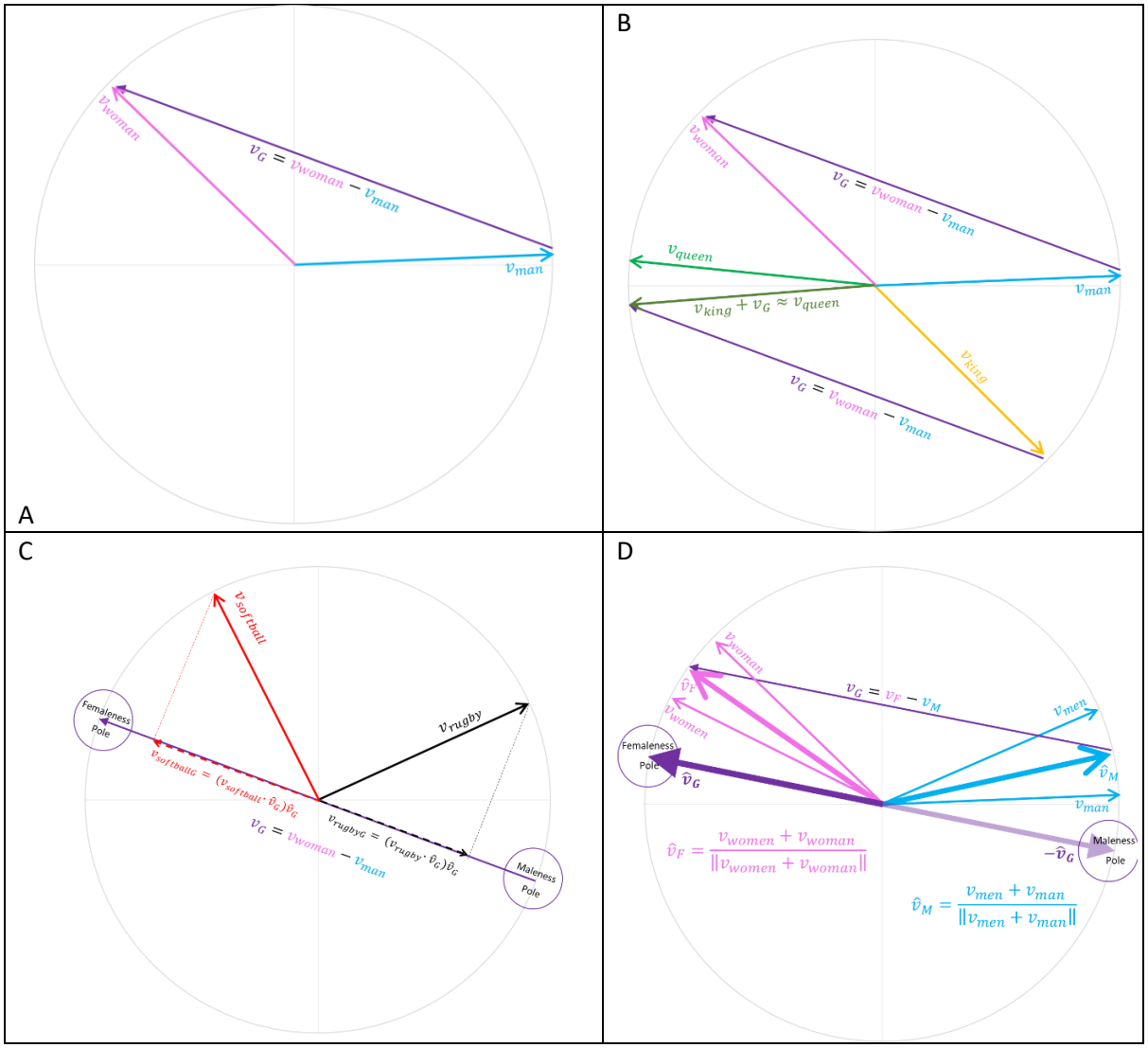

Figure S 3 Vector algebra operations in embedding space can produce vectors (i.e. cultural axis/dimensions) with culturally relevant meaning. Subtracting the vector representation of the word "man" from the vector representation of the word woman results in a gender vector that can be interpreted as a dimension in vector space tracing the space from maleness towards femaleness (A). The cultural relevance of the previous operation can be demonstrated by adding the gender axis to the vector representation of the word "king". That is, moving from "king" towards femaleness in vector space. This operation results in a vector whose closest neighbor is the vector representation for the word "queen" (B). Any term in the embedding model vocabulary can be projected onto a cultural axis such as gender. The value of the projections captures the latent associations of the term in the corpus with the poles of the cultural axis (C). To increase the reliability of cultural dimensions estimation, it is helpful to average several related words to create more robust poles (maleness and femaleness in the figure example) from which the cultural axis is estimated (D).

Projections of sentiment lexicons onto political axes

This work examines latent sentiment associations within word embedding models derived from outlet-specific news articles. Specifically, we measure the strength of association between positive and negative labelled words from external sentiment lexicons and distinct political orientation groups represented by the poles of political axes. This methodology is used as a proxy to measure

how a specific news outlet associates in their textual content positive and negative words with terms that denote political orientation.

Since the words from a sentiment lexicon are labelled as positive (i.e. +1) or negative (i.e. -1), and their projections on, for example a political affiliation axis, are positive (if the projection is closer to the Democrat pole) and negative (if the projection is closer to the Republican pole), we can calculate the correlation between sentiment word projection values and their sentiment labels. A positive correlation illustrates a tendency on the training corpus to associated positive words with the positive pole of the axis (in this example Democrats) and/or negative words with the negative pole of the axes (in this case Republicans), see Figure S 4. The signs of the axis projections can be reversed by simply changing the direction of the vectors substraction from which the axis was derived $v_P = v_{democrat} - w_{republican}$ (positive projections are associated with democrats) to $v_P = v_{republican} - w_{democrat}$ (positive projections are associated with republicans).

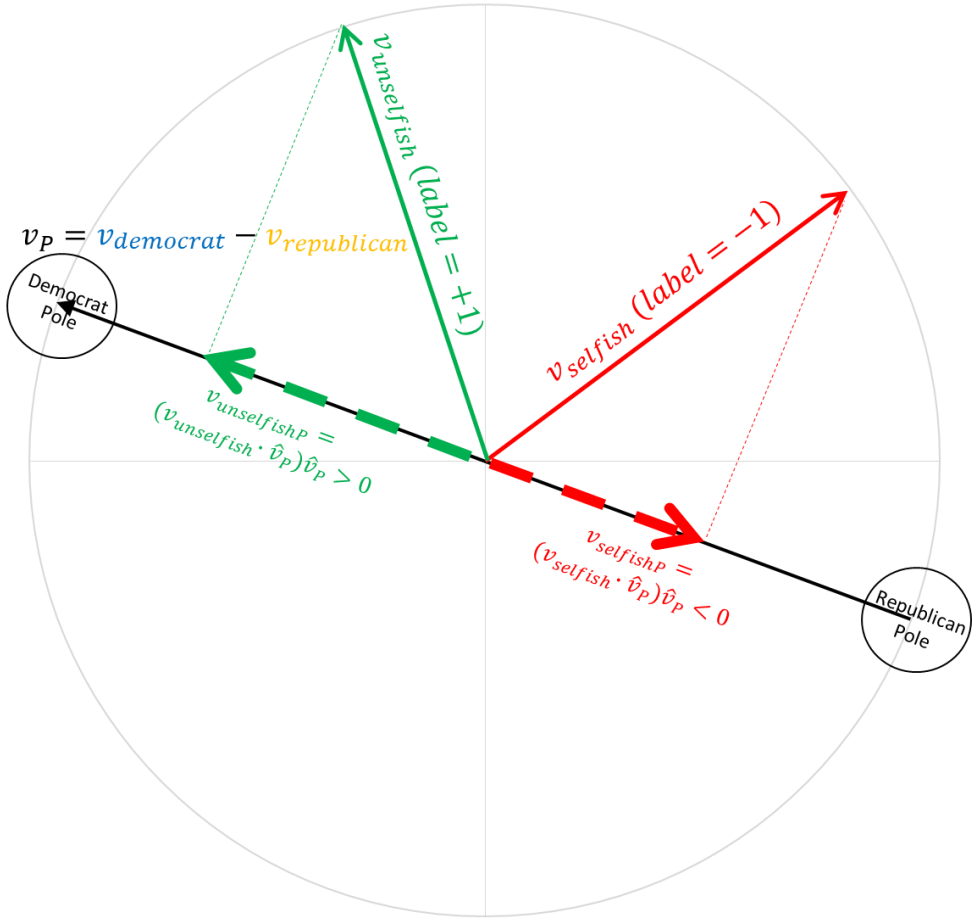

Figure S 4 Projecting a word with a positive or negative valence onto a cultural axis quantifies its degree of association with the terms used to build the poles (Democrat or Republican denoting terms in the figure above) of the cultural/demographic axis. The correlation between the projection values of sentiment words onto the political axis and their sentiment labels serves as a proxy to measure prevalent associations in the corpus of data on which the embedding model was trained.

Systematically projecting positive and negative labelled words from an external sentiment lexicon onto a political orientation cultural axis derived from a word embedding model allows to quantify the degree of association in the corpus on which the embedding model was trained between positive/negative terms and the distinct political orientations represented in the poles of the cultural

axis, see Figure S 4. This methodology has been successfully used previously to systematically test for sentiment associations in popular pre-trained embedding models [4].

Alignment of cultural axes representing political orientation with axes derived from Wordnet antonym pairs

In order to elucidate specific positive and negative pairs of terms that tend to more strongly project to either pole of a political axis, an additional experimental method can be used to estimate axes similarity in embedding space using pairs of antonyms, such as *unselfish-selfish* or *dry-wet* from the 3878 antonym pairs contained in Wordnet.

The cosine similarity between the set of axes generated using antonym pairs and the political axes analyzed in this work can then be estimated. A high degree of cosine similarity indicates alignment of the words in the antonym pair with the opposite poles of a political axis, See Figure S 5. For example, an axis derived from the antonym pair *maternal-paternal* should have a high degree of alignment (i.e. cosine similarity) with a gender axis. That is, in an embedding model trained on a normative corpus, the word *maternal* will be close to the feminine pole of the gender axis which is formed by words such as *woman*, *women* or *female*. In contrast, the word *paternal* will be closer to the masculine pole of the gender axis which is formed by words such as *man*, *men* or *male*. Thus, both of these axes will be similar in orientation. In contrast, an axis formed by a set of antonyms with no apparent relatedness to neither males nor females, such as centrifugal-centripetal, will be more orthogonal (i.e. dissimilar) to the gender axis. Thus, the degree of alignment between a cultural axis representing political orientation and a cultural axis representing an antonym pair can be used to measure the degree of association between the words in the antonym pair with the poles of the political axis.

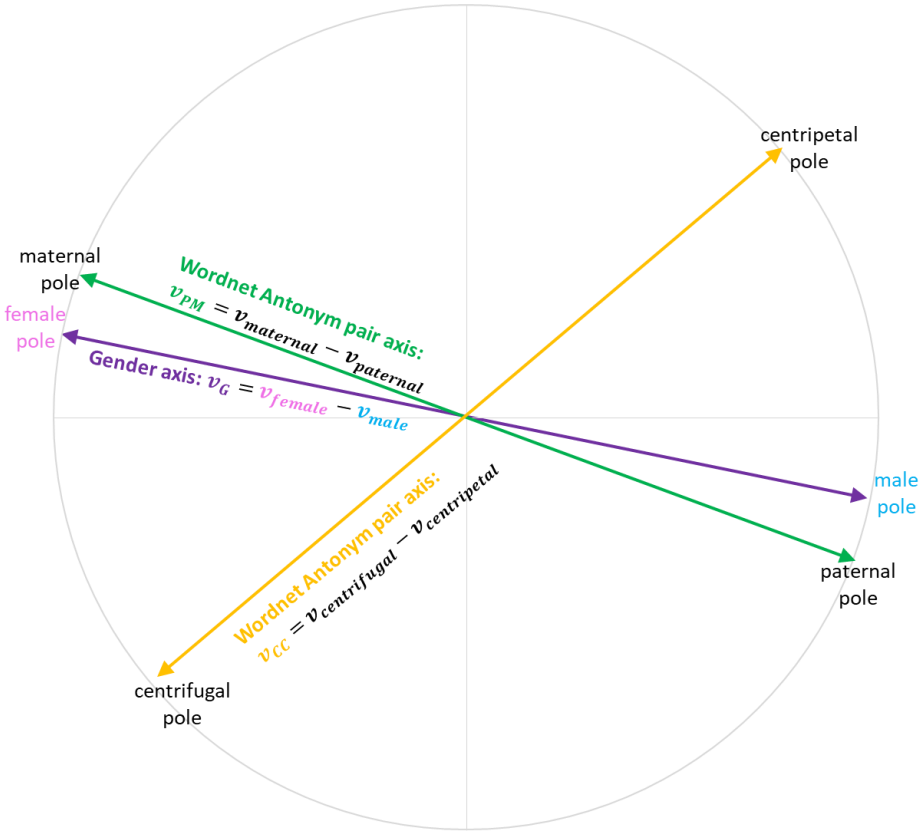

Figure S 5 The cosine similarity between axes created from WordNet antonym pairs and cultural axes representing demographic groups can be estimated to quantify the degree of alignment of the antonym pair with the poles of the cultural axis. An axis formed with words such as maternal and paternal that are clearly related to males and females will have a high degree of alignment with the gender axis. In contrast, antonym pairs with a low degree of relatedness or association with males and females (such as centrifugal-centripetal) will be more orthogonal to the gender axis.

Latent Associations of political orientation with sentiment lexicons across media outlets

19 sentiment lexicons were tested for sentiment associations with political orientation in the 47 word embedding models derived from individual news media outlets content from the 2015-2019 timeframe. The average correlation between association measurements in outlet-specific embedding spaces using the different sentiment lexicons and the external human ratings of outlets political bias from the AllSides organization [1] are shown in Table S 2. Regardless of sentiment lexicons used (each varying in size from 50 to more than 7,000 terms), word embeddings models associations correlated substantially with human perception of outlet ideological bias. The Ideonomy Personality Traits lexicon (IPT), N=526, displayed the highest average degree of correlation with human ratings of outlets political leanings and hence it was used in subsequent analysis to measure political associations in embedding models trained on news outlets articles.

	Political orientation axis						
Sentiment lexicons	Personal ideology	Party affiliation and political parties	U.S. presidents	Ideologically Oriented Journalists	U.S. senators	Influential conservatives and liberals	**AVERAGE**
harvardGeneralInquirer3623	0.64	0.93	0.63	0.68	0.86	0.68	**0.74**
WEAT50	0.42	0.82	0.45	0.60	0.58	0.43	**0.55**
vaderLexicon7062	0.62	0.92	0.63	0.68	0.82	0.69	**0.73**
NRCEmotionLexicon5555	0.65	0.93	0.65	0.69	0.85	0.64	**0.73**
opinionLexicon6786	0.65	0.93	0.65	0.71	0.85	0.70	**0.75**
afinnLexicon2477	0.66	0.93	0.65	0.72	0.85	0.72	**0.75**
positiveNegativeAdjetives762	0.75	0.94	0.68	0.72	0.87	0.72	**0.78**
positiveNegativeAdjetives197	0.72	0.94	0.74	0.72	0.85	0.72	**0.78**
happySadAdjectives122	0.45	0.87	0.53	0.56	0.84	0.60	**0.64**
niceMeanAdjetives228	0.56	0.87	0.63	0.61	0.72	0.67	**0.68**
intelligentDullAdjetives75	0.67	0.80	0.60	0.57	0.67	0.55	**0.64**
inquirerViceVirtue1277	0.69	0.95	0.66	0.69	0.85	0.72	**0.76**
inquirerHostileAffiliation1176	0.57	0.87	0.63	0.62	0.79	0.61	**0.68**
inquirerPowerConflictCooperation294	0.44	0.73	0.42	0.50	0.82	0.40	**0.55**
inquirerAffectNegativePositive261	0.64	0.92	0.67	0.68	0.83	0.66	**0.74**
ideonomyPersonalityTraits526	0.76	0.94	0.72	0.72	0.86	0.73	**0.79**
EMOTEvalence985	0.71	0.93	0.63	0.72	0.85	0.67	**0.75**
EMOTELikeableness985	0.72	0.93	0.65	0.72	0.84	0.69	**0.76**
EMOTELikeableness554	0.72	0.93	0.69	0.69	0.82	0.69	**0.76**
AVERAGE	**0.63**	**0.90**	**0.63**	**0.66**	**0.81**	**0.65**	**0.71**

Table S 2 Pearson's r correlations between AllSides human ratings of media outlets political bias [1] and outlet-specific embedding model associations across six different political orientation axes (columns) using 19 different sentiment lexicons (rows).

Average correlation between the different sentiment lexicons tested and their measurements of political orientation associations in news media outlets

Average correlation between the results of using different sentiment lexicons to test for political associations in outlet specific embedding models are displayed in Table S 3. Correlation coefficients were in general very high (average r= 0.93). That is, all sentiment lexicons were similarly able to measure similar political sentiment associations in embedding models trained on news outlets corpora and they all correlate substantially with human perceptions of media bias.

	Sentiment lexicons																		
Sentiment lexicons	*harvardGeneralInquirer3623*	*WEAT1*	*vaderLexicon7062*	*NRCEmotionLexicon5555*	*opinionLexicon6786*	*afinnLexicon2477*	*positiveNegativeAdjetives762*	*positiveNegativeAdjetives197*	*happySadAdjectives122*	*niceMeanAdjetives228*	*intelligentDullAdjetives75*	*inquirerViceVirtue1277*	*inquirerHostileAffiliation1176*	*inquirerPowerConflictCooperation294*	*inquirerAffectNegativePositive261*	*ideonomyPersonalityTraits526*	*EMOTEvalence985*	*EMOTELikeableness985*	*EMOTELikeableness554*
harvardGeneralInquirer3623	1.00																		
WEAT1	0.85	1.00																	
vaderLexicon7062	0.99	0.88	1.00																
NRCEmotionLexicon5555	0.98	0.88	0.97	1.00															
opinionLexicon6786	1.00	0.89	1.00	0.98	1.00														
afinnLexicon2477	0.99	0.87	1.00	0.96	1.00	1.00													
positiveNegativeAdjetives762	0.98	0.82	0.95	0.98	0.96	0.96	1.00												
positiveNegativeAdjetives197	0.96	0.85	0.96	0.98	0.95	0.94	0.96	1.00											
happySadAdjectives122	0.97	0.77	0.97	0.94	0.97	0.96	0.90	0.92	1.00										
niceMeanAdjetives228	0.92	0.85	0.95	0.89	0.94	0.93	0.87	0.94	0.91	1.00									
intelligentDullAdjetives75	0.77	0.72	0.75	0.82	0.74	0.74	0.87	0.88	0.63	0.73	1.00								
inquirerViceVirtue1277	0.99	0.84	0.98	0.97	0.98	0.98	0.98	0.98	0.94	0.94	0.84	1.00							
inquirerHostileAffiliation1176	0.98	0.86	0.98	0.99	0.98	0.96	0.94	0.98	0.96	0.94	0.78	0.96	1.00						
inquirerPowerConflictCooperation294	0.92	0.72	0.88	0.94	0.89	0.87	0.90	0.86	0.91	0.73	0.67	0.86	0.91	1.00					
inquirerAffectNegativePositive261	0.98	0.88	0.98	0.99	0.98	0.97	0.96	0.99	0.95	0.94	0.81	0.98	1.00	0.90	1.00				
ideonomyPersonalityTraits526	0.96	0.81	0.95	0.97	0.94	0.94	0.99	0.98	0.89	0.90	0.91	0.99	0.95	0.87	0.97	1.00			
EMOTEvalence985	0.97	0.87	0.95	0.98	0.96	0.96	0.99	0.94	0.89	0.84	0.84	0.96	0.94	0.91	0.95	0.96	1.00		
EMOTELikeableness985	0.97	0.86	0.95	0.97	0.96	0.96	0.99	0.94	0.88	0.85	0.86	0.97	0.93	0.89	0.95	0.97	1.00	1.00	
EMOTELikeableness554	0.95	0.86	0.95	0.97	0.94	0.94	0.98	0.99	0.87	0.91	0.92	0.98	0.95	0.84	0.97	0.99	0.96	0.97	1.00

Table S 3 Correlation results between the political sentiment associations measured in news outlets when using different sentiment lexicons.

Testing for political associations in outlet-specific word embeddings using the IPT sentiment lexicon - Detailed results

Table S 4 provides specific numerical results and statistical significance tests, Bonferroni adjusted for multiple comparisons, of the association experiments for political bias in outlet-specific embedding models summarized in Figure 3 of the main manuscript .

		Personal ideology		Party affiliation		U.S. presidents		Ideologically oriented journalists		U.S. Senators		Influential conservatives and liberals	
	Outlet name	r	p	r	p	r	p	r	p	r	p	r	p
1	Alternet	0.43	3.99E-18	0.44	1.56E-19	0.50	8.89E-26	0.03	1.00E+00	0.37	1.17E-12	0.10	1.00E+00
2	Democracy Now	0.33	4.01E-07	0.36	1.14E-08	0.10	1.00E+00	0.32	7.95E-07	0.18	2.96E-01	0.27	2.99E-04
3	Daily Beast	0.11	1.00E+00	0.34	1.64E-11	0.34	8.33E-12	0.09	1.00E+00	0.30	5.73E-09	0.07	1.00E+00
4	Huffington Post	0.54	6.88E-35	0.50	1.24E-28	0.57	2.30E-39	0.20	3.75E-03	0.23	2.06E-04	0.38	2.11E-15
5	The Intercept	0.25	2.64E-04	0.24	1.39E-03	0.18	1.68E-01	0.12	1.00E+00	0.23	3.32E-03	0.01	1.00E+00
6	Jacobin	0.39	3.48E-14	0.38	2.88E-13	0.39	8.69E-14	0.26	3.31E-05	0.33	6.19E-10	0.21	6.74E-03
7	Mother Jones	0.23	9.47E-04	0.27	1.66E-05	0.16	5.00E-01	0.19	4.24E-02	0.17	1.75E-01	0.11	1.00E+00
8	The New Yorker	0.37	1.29E-14	0.46	8.35E-24	0.31	1.11E-09	-0.01	1.00E+00	0.24	4.21E-05	0.06	1.00E+00
9	The Nation	0.40	2.32E-16	0.50	1.99E-27	0.41	3.88E-17	0.39	7.97E-15	0.50	7.12E-27	0.29	5.64E-08
10	Slate	0.37	2.85E-14	0.33	3.97E-11	0.13	1.00E+00	0.25	1.65E-05	0.12	1.00E+00	0.26	4.24E-06
11	Vox	0.25	1.03E-05	0.38	6.68E-15	0.23	9.11E-05	-0.02	1.00E+00	0.27	1.39E-06	0.26	4.74E-06
12	CNN	0.14	8.41E-01	0.26	3.64E-06	0.16	2.22E-01	0.24	3.47E-05	-0.09	1.00E+00	0.29	7.31E-08
13	New York Times	0.42	3.53E-20	0.39	9.73E-17	0.29	2.84E-08	0.24	2.32E-05	0.14	5.25E-01	0.29	2.61E-08
14	ABC News	0.31	3.19E-09	0.18	3.76E-02	-0.01	1.00E+00	0.09	1.00E+00	-0.01	1.00E+00	0.14	6.27E-01
15	The Atlantic	0.28	6.17E-08	0.34	2.18E-12	0.12	1.00E+00	0.12	1.00E+00	0.25	9.16E-06	0.31	4.95E-10
16	Buzzfeed	0.39	2.28E-15	0.11	1.00E+00	0.17	1.35E-01	0.09	1.00E+00	0.13	1.00E+00	0.19	1.31E-02
17	CBS News	0.08	1.00E+00	0.10	1.00E+00	0.20	8.86E-03	0.12	1.00E+00	-0.02	1.00E+00	-0.04	1.00E+00
18	The Economist	0.14	1.00E+00	0.17	1.48E-01	0.08	1.00E+00	0.31	7.14E-09	0.23	5.38E-04	0.05	1.00E+00
19	The Guardian	0.59	6.76E-44	0.44	2.42E-22	0.65	4.49E-58	0.33	1.02E-11	0.39	1.17E-16	0.53	7.41E-35
20	NBC News	0.24	7.74E-05	0.22	5.07E-04	0.10	1.00E+00	0.26	8.38E-06	0.08	1.00E+00	0.33	1.48E-10
21	POLITICO	0.32	2.26E-09	0.13	1.00E+00	0.17	1.76E-01	0.05	1.00E+00	-0.16	2.05E-01	0.03	1.00E+00
22	TIME	0.26	3.90E-06	0.21	1.23E-03	0.10	1.00E+00	0.25	2.15E-05	0.13	1.00E+00	0.30	8.53E-09
23	Washington Post	0.29	1.06E-08	0.41	2.09E-18	0.27	2.18E-07	0.07	1.00E+00	0.07	1.00E+00	0.30	3.49E-09
24	NPR	0.36	1.86E-13	0.06	1.00E+00	0.03	1.00E+00	-0.06	1.00E+00	0.00	1.00E+00	0.15	2.92E-01
25	Associated Press	0.43	4.47E-19	0.11	1.00E+00	0.05	1.00E+00	0.25	1.15E-05	0.05	1.00E+00	0.13	1.00E+00
26	BBC	0.36	4.97E-13	0.04	1.00E+00	0.17	5.68E-02	0.40	2.81E-16	0.21	2.05E-03	0.17	5.67E-02
27	Bloomberg	0.19	1.10E-02	-0.01	1.00E+00	0.03	1.00E+00	0.11	1.00E+00	-0.18	4.44E-02	0.08	1.00E+00
28	Christian Science	0.24	8.54E-05	0.18	4.18E-02	0.23	3.12E-04	-0.01	1.00E+00	0.10	1.00E+00	0.19	2.05E-02
29	REUTERS	0.34	5.63E-11	0.01	1.00E+00	0.01	1.00E+00	0.15	2.93E-01	-0.02	1.00E+00	0.12	1.00E+00
30	The Hill	0.14	1.00E+00	0.14	1.00E+00	-0.06	1.00E+00	-0.15	5.70E-01	-0.05	1.00E+00	0.07	1.00E+00
31	USA Today	0.33	5.11E-11	0.20	5.44E-03	0.13	9.61E-01	0.21	1.54E-03	0.01	1.00E+00	0.17	6.44E-02
32	Wall Street Journal	-0.08	1.00E+00	-0.06	1.00E+00	0.04	1.00E+00	-0.10	1.00E+00	-0.23	1.48E-04	0.02	1.00E+00
33	Reason	0.21	1.68E-02	-0.10	1.00E+00	-0.02	1.00E+00	0.13	1.00E+00	-0.17	3.61E-01	0.05	1.00E+00
34	Washington Examir	0.00	1.00E+00	-0.18	2.12E-02	-0.09	1.00E+00	-0.33	1.67E-11	-0.37	1.51E-14	-0.19	8.56E-03
35	Washington Times	-0.14	5.98E-01	-0.07	1.00E+00	-0.27	3.83E-07	0.02	1.00E+00	-0.18	3.20E-02	-0.07	1.00E+00
36	Fox News	-0.30	1.09E-08	-0.27	1.13E-06	-0.26	2.05E-06	-0.41	1.66E-18	-0.37	5.83E-14	-0.25	1.06E-05
37	American Spectato	0.12	1.00E+00	-0.22	1.05E-03	-0.15	7.27E-01	-0.23	5.80E-04	-0.21	6.05E-03	-0.20	1.27E-02
38	Breitbart	-0.19	8.74E-03	-0.14	5.27E-01	-0.09	1.00E+00	-0.26	6.20E-06	-0.33	3.60E-11	-0.13	1.00E+00
39	The Blaze	0.02	1.00E+00	-0.17	1.21E-01	0.10	1.00E+00	-0.11	1.00E+00	-0.13	1.00E+00	0.02	1.00E+00
40	Christian Broadcast	0.03	1.00E+00	-0.36	2.17E-08	-0.31	4.40E-06	-0.02	1.00E+00	-0.48	4.66E-17	-0.41	1.83E-11
41	The Daily Caller	0.09	1.00E+00	-0.26	6.04E-06	-0.25	2.67E-05	-0.29	6.17E-08	-0.29	6.01E-08	-0.02	1.00E+00
42	The Daily Mail	-0.10	1.00E+00	-0.16	9.88E-02	0.14	4.40E-01	-0.12	1.00E+00	-0.19	8.84E-03	-0.06	1.00E+00
43	The Daily Wire	0.05	1.00E+00	-0.37	1.65E-12	-0.03	1.00E+00	-0.35	3.22E-11	-0.34	2.22E-10	-0.06	1.00E+00
44	The Federalist	-0.11	1.00E+00	-0.29	1.35E-07	0.12	1.00E+00	-0.41	3.86E-17	-0.41	1.04E-17	-0.22	7.43E-04
45	National Review	-0.20	2.94E-03	-0.33	4.20E-11	-0.27	7.63E-07	-0.53	6.13E-34	-0.46	2.54E-24	-0.41	2.95E-18
46	New York Post	-0.11	1.00E+00	-0.20	2.29E-03	-0.08	1.00E+00	-0.11	1.00E+00	-0.11	1.00E+00	-0.04	1.00E+00
47	Newsmax	-0.19	1.57E-02	-0.24	9.66E-05	-0.05	1.00E+00	-0.02	1.00E+00	-0.36	2.30E-13	-0.11	1.00E+00

Table S 4 Pearson correlation coefficients between positive/negative labels in the IPT lexicon (N=526) and the projection values of the IPT terms on 6 axes (columns) representing political orientation in 47 word embedding models (rows) each trained on a corpus of news and opinion articles from individual news media outlets for the time interval 2015-2019. The r columns are the Pearson correlation coefficients between sentiment lexicon annotations (positive or negative) and the projection value of the sentiment lexicon terms on the political axis represented by the column. Positive values of r indicate preferential association of positive words with terms that denote left-leaning individuals and negative words with terms that denote right-leaning individuals. A negative value of r indicates the opposite, preferential association in the outlet embeddings of positive words with terms that denote right-leaning individuals and negative words with terms that denote left-leaning individuals. The p columns display the (multiple comparisons Bonferroni corrected) p-values of the correlation coefficient r. The complete list of words used to build the poles of each political axis is provided in this SM.

Correlation between the six different political orientation axis in embedding space and their ability to detect political orientation in news media outlets corpora

The six political orientation axes used in Figure 3 of the main manuscript have a high degree of correlation among their measurements of political orientation bias across different news media outlets in Table S 4 (average r= 0.75, see Table S 5). This suggests the ability of different political orientation axes to consistently align themselves in embedding space to trace the spectrum of political orientation despite being constructed with different sets of words whose only commonality is their factual political affiliation.

	Political orientation axis					
Political orientation axis	Personal ideology	Party affiliation	U.S. presidents	Ideologically oriented journalists	U.S. senators	Influential conservatives and liberals
Personal ideology	1.00					
Party affiliation	0.77	1.00				
U.S. presidents	0.62	0.81	1.00			
Ideologically oriented journalists	0.71	0.75	0.61	1.00		
U.S. senators	0.80	0.92	0.74	0.77	1.00	
Influential conservatives and liberals	0.72	0.80	0.76	0.75	0.77	1.00

Table S 5 Correlations between the experimental results reported in Table S 4 and Figure 3 of the main manuscript for the six different political orientation axes used to probe for political bias in news media outlets.

Alternative ratings of news media political bias

There are a number of publicly available human ratings of news outlets bias. However, their bias annotations are often similar. In this work, we have used the AllSides media bias chart 2019 v1.1 [1] human ratings of outlets bias as the reference *ground truth* since it aggregates thousands of human opinions about media political bias and also because we also used that source to specify the list of 47 outlets analyzed in this work. Nonetheless, we also retrieved a number of other human ratings of news outlets bias to compare them to AllSides ratings.

The list of alternative human ratings of news outlets bias analyzed and their ideological bias ratings correlations with AllSides ideological bias ratings are provided below. The high degree of correlation suggests that the results reported in this work are similar regardless of human ratings of outlet ideological bias source used.

Ad Fontes media bias chart: r=0.91 https://www.adfontesmedia.com/interactive-media-bias-chart

Media bias fact check: r=0.97 https://mediabiasfactcheck.com

Fair reporters: r=0.93 http://www.fairreporters.org/news-media-bias-ratings

Long-term temporal changes in media associations

To measure temporal dynamics of political associations in news media content, outlet-specific embedding models were generated for the following intervals: 1975-1979, 1980-1984, 1985-1989, 1990-1994, 1995-1999, 2000-2004, 2005-2009, 2-10-2014 and 2015-2019. An embedding model was also generated for 1970-1974 New York Times articles, but since this was the only outlet with significant number of articles available for the time period, it was not used in the analysis to prevent a single outlet characterizing an entire five year time range.

The results in Figure 7 of the main manuscript aggregate a growing set of news outlets due to larger availability of news media content in recent times. We replicate the results of Figure 7 in the main

manuscript using a fixed set of 11 popular outlets with full article availability since at least the year 2000. For each ideological axis in Figure S6 from left to right, two-sided t-tests with null hypotheses zero slope for left-leaning media were: t(2)=8.35, p= .01, t(2)=1.10, p= .384 and t(2)=2.35, p= .014. A Fisher combination of p-values was significant (Fisher statistic=14.34, p=0.026). For right-leaning outlets however, the results are not significant and thus inconclusive.

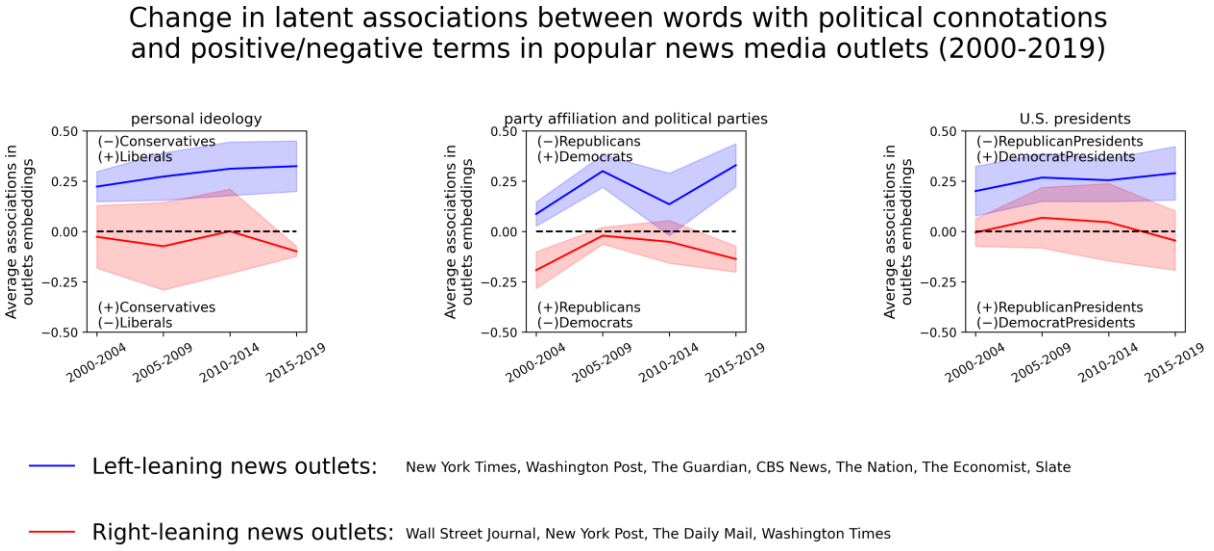

Figure S 6 Long-term analysis spanning two decades of latent sentiment associations in a fixed set of popular left-leaning and right-leaning news media outlets across several political orientation categories. The dotted black line in each plot denotes a level of no preferential association of positive/negative words with either pole of the political axis. The shade of the trends shows the 95% confidence interval. The top left and bottom left of each plot indicate the positive/negative sentiment associations displayed by embedding models in that area of the chart.

Building Axes to test for bias in popular word embedding models

A comprehensive set of cultural/demographic axes have been used in this work. Some cultural axes were designed for illustration purposes while others were designed specifically to detect sentiment associations for political categories. The list of axes created and the poles used to build them are detailed below:

Axis name: Politics - Personal ideology
Pole 1 (conservative): conservative, conservatives, right_winger, rightwinger, right_wingers, rightwingers, right_leaning, right_wing, rightwing, right_leaning
Pole 2 (liberal): liberal, liberals, progressive, progressives, left_winger, leftwinger, left_wingers, leftwingers, left_leaning, left_wing, leftwing, left_leaning

Axis name: Politics - Party affiliation
Pole 1 (Republican): Republican, Republicans, GOP, Republican_Party
Pole 2 (Democrat): Democrat, Democrats, Democratic_Party

Axis name: Politics – U.S. presidents
Notes: U.S. presidents since World War II.

Pole 1 (Republicans): Dwight_Eisenhower, Richard_Nixon, Gerald_Ford, Ronald_Reagan, George_Bush, Donald_Trump
Pole 2 (Democrats): Franklin_Roosevelt, Harry_Truman, John_Kennedy, Lyndon_Johnson, Jimmy_Carter, Bill_Clinton, Barack_Obama

Axis name: Politics – Journalists
Notes: left-wing and right-wing journalists according to https://www.politico.com/blogs/media/2015/04/twitters-most-influential-political-journalists-205510
Pole 1 (right wing journalists): Jake_Tapper,Megyn_Kelly,Sean_Hannity,Michelle_Malkin,Dana_Perino,Bret_Baier,Greta_Van Susteren, Glenn_Beck, Bill_Reilly, Andrew_Malcolm, Matt_Drudge, Charles_Krauthammer, Ann_Coulter, Ed_Henry, Dana_Loesch, Brit_Hume, Sarah_Elizabeth_Cupp, Major_Garrett ,Greg_Gutfeld, Tucker_Carlson, Andrea_Tantaros, Andrew_Napolitano, Erick_Erickson, Stephen_Hayes, Kimberly_Guilfoyl, Jonah_Goldberg, Neil_Cavuto, Peggy_Noonan, Monica_Crowley, Kirsten_Powers, Robert_Costa, Larry_Sabato, Mary_Katharine_Ham, Eric_Bolling, Rich_Lowry
Pole 2 (left wing journalists): Anderson_Cooper, Rachel_Maddow, Ezra_Klein, Arianna_Huffington, Nate_Silver, George_Stephanopoulos, Christiane_Amanpour, Paul_Krugman, Ann_Curry, Chris_Hayes, Glenn_Greenwald, Melissa_Harris_Perry, Fareed_Zakaria, Donna_Brazile, Nicholas_Kristof, John_Dickerson, David_Corn, Robert_Reich, Katrina_vanden_Heuvel, Jim_Roberts, Matt_Taibbi, Matthew_Yglesias, Lawrence_Donnell, Andy_Borowitz, Chris_Matthews, Diane_Sawyer, Don_Lemon, Markos_Moulitsas, Thomas_Friedman, Ana_Marie_Cox, Chris_Cuomo, Al_Sharpton, Andrew_Sullivan, Bill_Keller, Charles_Blow

Axis name: Politics – U.S. Senators (as of March 2020)
Notes: Republican and Democratic senators in the U.S. according to https://en.wikipedia.org/wiki/List_of_current_United_States_senators

Pole 1 (Republican senators): Mike_Pence,Richard_Shelby,Dan_Sullivan,Lisa_Murkowski,Martha_McSally,Tom_Cotton, John_Boozman, Cory_Gardner, Rick_Scott, Marco_Rubio, David_Perdue, Kelly_Loeffler, Jim_Risch, Mike_Crapo, Todd_Young, Mik_Braun, Chuck_Grassley, Joni_Ernst, Pat_Roberts, Jerry_Moran, Rand_Paul, Mitch_McConnell, John_Kennedy, Bill_Cassidy, Susan_Collins, Roger_Wicker, Cindy_Hyde_Smith, Josh_Hawley, Roy_Blunt, Steve_Daines, Ben_Sasse, Deb_Fischer, Thom_Tillis,Richard_Burr, John_Hoeven, Kevin_Cramer, Rob_Portman, James_Lankford, Jim_Inhofe, Pat_Toomey, Tim_Scott, Lindsey_Graham, John_Thune, Mike_Rounds, Marsha_Blackburn, Lamar_Alexander, Ted_Cruz, John_Cornyn, Mitt_Romney, Mike_Lee, Shelley_Moore_Capito, Ron_Johnson, Mike_Enzi, John_Barrasso

Pole 2 (Democrat senators): Doug_Jones, Kyrsten_Sinema, Kamala_Harris, Dianne_Feinstein, Michael_Bennet, Chris_Murphy, Richard_Blumenthal, Chris_Coons, Tom_Carper, Brian_Schatz, Mazie_Hirono, Dick_Durbin, Tammy_Duckworth, Chris_Van_Hollen, Ben_Cardin, Elizabeth_Warren, Ed_Markey, Debbie_Stabenow, Gary_Peters, Tina_Smith, Amy_Klobuchar, Jon_Tester, Jacky_Rosen, Catherine_Cortez_Masto, Jeanne_Shaheen, Maggie_Hassan, Bob_Menendez, Cory_Booker, Tom_Udall, Martin_Heinrich, Chuck_Schumer, Kirsten_Gillibrand, Sherrod_Brown, Ron_Wyden, Jeff_Merkley, Bob_Casey, Sheldon_Whitehouse, Jack_Reed, Patrick_Leahy, Mark_Warner, Tim_Kaine, Patty_Murray, Maria_Cantwell, Joe_Manchin, Tammy_Baldwin, Bernie_Sanders,

Axis name: Politics – famous/influential liberals and conservatives

Note: list taken from the top 20 conservatives and liberals at "Top 100 US liberals and conservatives". We exclude names already included in axes listed above to avoid redundant double counting. https://www.telegraph.co.uk/news/worldnews/northamerica/usa/6951961/Top-100-US-liberals-and-conservatives.html
Pole 1 (famous/influential conservatives): Dick_Cheney, Rush_Limbaugh, Sarah_Palin, Robert_Gates, Roger_Ailes, David_Petraeus, Paul_Ryan, Tim_Pawlenty, John_Roberts, Haley_Barbour, Eric_Cantor, John_McCain, Bob_McDonnell, Newt_Gingrich Mike_Huckabee
Pole 2 (famous/influential liberals): Hillary_Clinton, Nancy_Pelosi, Rahm_Emanuel, Al_Gore, Oprah_Winfrey, Tim_Geithner, David_Axelrod, Harry_Reid, Michelle_Obama, Arianna_Huffington, Sonia_Sotomayor, Denis_McDonough, Janet_Napolitano, Mark_Warner, Robert_Gibbs, Barney_Frank, John_Kerry, Eric_Holder

Axis name: Gender - occupations
Pole 1 (males): man, men, male, males
Pole 2 (females): woman, women, female, females

Axis name: Countries - economic development
Pole 1 (poverty): poor, poverty, underdeveloped
Pole 2 (rich): wealth, rich, wealthy, prosperous, developed

Axis name: Death to life
Pole 1 (death): death, dying, decease
Pole 2 (life): alive, life, living

Axis name: Historical figures
Pole 1 (malevolent historical figures): Hitler, Stalin, Bin_Laden, Pol_Pot, Heinrich_Himmler, Saddam_Hussein, Joseph_Goebbels
Pole 2 (respected historical figures): Gandhi, MLK, Nelson_Mandela, Mother_Teresa, Abraham_Lincoln

Axis name: Health status
Pole 1 (disease): disease, sick, sickness, illness
Pole 2 (health): health, healthy, well_being

Axis name: Government type
Pole 1 (dictatorship): dictatorship, dictator, dictators, autocracy, authoritarianism, totalitarianism, tyranny, despotism
Pole 2 (democracy): democracy, democratic_leader, democratic_leaders, representative_government

Sentiment lexicons used in this work

Table S 6 contains the list of 19 sentiment lexicons tested in this work. The lexicons contain terms externally annotated for positive and negative polarity. The ensemble of sentiment lexicons includes several lexicons often used in the scholarly literature for content and sentiment analysis, several online lists of positive and negative character traits, lists of positive and negative adjectives as well

as several specialized lexicons from the Harvard General Inquirer (HGI) that measure constructs with clear positive and negative dichotomies such as vice/virtue, conflict/cooperation or hostility/affiliation.

Lexicons were preprocessed to remove invalid entries such as for instance emoticons in the Vader lexicon since they are not present in the word embeddings models analyzed. All lexicons were lowercased. Preprocessing occasionally resulted in lexicon sizes slightly smaller than the original lexicons size. In the case of HGI, the smaller lexicon size is due to entries in HGI having multiple annotations for different senses. In those cases, this work used the annotation corresponding to the most frequent sense of the word according to HGI metadata.

Lexicon ID	Lexicon Name	Preprocessing lexicon size	Postprocessing Lexicon size	Lexicon description and location
1	WEAT	50	50	Lexicon used by (8) for testing for bias in word embedding models https://science.sciencemag.org/content/sci/suppl/2017/04/12/356.6334.183.DC1/Caliskan-SM.pdf
2	Harvard General Inquirer IV-4 dictionary Positiv Negativ	4206	3623	Positiv 1,915 words of positive outlook. Negativ 2,291 words of negative outlook http://www.wjh.harvard.edu/~inquirer/homecat.htm
3	Vader Lexicon	7500	7062	Empirically validated by multiple independent human judges. The VADER sentiment lexicon is sensitive both the polarity and the intensity of sentiments expressed in social media contexts, and is also generally applicable to sentiment analysis in other domains. Sentiment ratings from 10 independent human raters (all pre-screened, trained, and quality checked for optimal inter-rater reliability). Over 9,000 token features were rated on a scale from "[–4] Extremely Negative" to "[4] Extremely Positive", with allowance for "[0] Neutral (or Neither, N/A)". https://github.com/cjhutto/vaderSentiment
4	NRC Emotion Lexicon	5555	5555	The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing. http://saifmohammad.com/WebPages/lexicons.html
5	Opinion Lexicon	6800	6786	A list of English positive and negative opinion words or sentiment words. This list was compiled over many years starting from our first paper (Hu and Liu, KDD-2004). https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
6	Afinn Lexicon	2477	2477	AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
7	Positive/Negative Character Traits	762	762	A list of Positive/negative character traits and personality characteristics https://www.listofcharactertraits.com/positive.php https://www.listofcharactertraits.com/negative.php
8	Positive/Negative Adjetives	200	197	Lists of common adjectives that describe personality. 100 common personality adjectives that describe people negatively. 100 common personality adjectives that describe people positively https://www.englishclub.com/vocabulary/adjectives-personality.htm
9	Happy/Sad Adjectives	122	122	List Of Happy/Sad Character Traits & Personality Characteristics https://www.listofcharactertraits.com/happy.php https://www.listofcharactertraits.com/sad.php
10	Nice/Mean Adjetives	228	228	List Of Nice/Mean Character Traits & Personality Characteristics https://www.listofcharactertraits.com/nice.php https://www.listofcharactertraits.com/mean.php
11	intelligent/Dull Adjetives	75	75	List Of Intelligent/Dull adjectives https://www.thesaurus.com/browse/intelligent https://www.thesaurus.com/browse/unintelligent
12	Inquirer Vice/Virtue	1404	1277	719 words indicating an assessment of moral approval or good fortune, especially from the perspective of middle-class society. 685 words indicating an assessment of moral disapproval or misfortune http://www.wjh.harvard.edu/~inquirer/homecat.htm
13	Inquirer Hostile/Affiliation	1390	1176	833 words words indicating an attitude or concern with hostility or aggressiveness. 557 words are also tagged Affil for words indicating affiliation or supportiveness. http://www.wjh.harvard.edu/~inquirer/homecat.htm
14	Inquire Conflict/Cooperation	346	294	228 words for ways of conflicting. 118 words for ways of cooperating http://www.wjh.harvard.edu/~inquirer/homecat.htm
15	Inquirer Affect Negative/Positive	319	261	193 words of negative affect "denoting negative feelings and emotional rejection. 126 words of positive affect "denoting positive feelings, acceptance, appreciation and emotional support." http://www.wjh.harvard.edu/~inquirer/homecat.htm
16	ideonomy PersonalityTraits 526	638	526	638 Primary personality traits (234 positive, 292 neutral, 292 negative) http://ideonomy.mit.edu/essays/traits.html
17	EMOTEvalence985	985	985	The English Word Database of EMOtional TErms (EMOTE; Grühn, 2016) is a database of 1287 nouns and 985 adjectives. The goal of the database is to provide an easily accessible and comprehensive word pool that are relevant for research in the socio-emotional domain and for research on the interface between cognition and emotion, such as emotional memory. https://acelab.wordpress.ncsu.edu/material/emote/
18	EMOTELikeableness985	985	985	The English Word Database of EMOtional TErms (EMOTE; Grühn, 2016) is a database of 1287 nouns and 985 adjectives. The goal of the database is to provide an easily accessible and comprehensive word pool that are relevant for research in the socio-emotional domain and for research on the interface between cognition and emotion, such as emotional memory. https://acelab.wordpress.ncsu.edu/material/emote/
19	EMOTELikeableness554	554	554	The English Word Database of EMOtional TErms (EMOTE; Grühn, 2016) is a database of 1287 nouns and 985 adjectives. The goal of the database is to provide an easily accessible and comprehensive word pool that are relevant for research in the socio-emotional domain and for research on the interface between cognition and emotion, such as emotional memory. https://acelab.wordpress.ncsu.edu/material/emote/

Table S 6 External sentiment lexicons used to test for sentiment associations in word embeddings models trained on news media outlets news and opinion articles

Supplemental Information References

[1] "AllSides Media Bias Ratings," *AllSides*. https://www.allsides.com/blog/updated-allsides-media-bias-chart-version-11 (accessed May 10, 2020).

[2] A. C. Kozlowski, M. Taddy, and J. A. Evans, "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings," *Am. Sociol. Rev.*, vol. 84, no. 5, pp. 905–949, Oct. 2019, doi: 10.1177/0003122419877135.

[3] A. Gladkova, A. Drozd, and S. Matsuoka, "Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't.," in *Proceedings of the NAACL Student Research Workshop*, San Diego, California, Jun. 2016, pp. 8–15. Accessed: Mar. 25, 2019. [Online]. Available: http://www.aclweb.org/anthology/N16-2002

[4] D. Rozado, "Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types," *PLOS ONE*, vol. 15, no. 4, p. e0231189, Apr. 2020, doi: 10.1371/journal.pone.0231189.