Online intermediaries and the diversity of news content

Annex 2

Published 25 March 2024

Contents

Section

Overview

Online intermediaries (OIs) occupy an increasingly important place in news consumption.[1] OIs, and in particular social media platforms, have incentives to attract audience attention and to drive audience engagement, and have the ability to personalise the news that they show their users. These features raised concerns that news delivered via OIs may be narrowly focused on individuals' existing views and preferences and consequently could lead to news diets that lack a diversity of viewpoints.[2]

In principle an OI can affect a user's news diversity in different – and opposing – ways.[3] For example, a news feed recommender system might present news informed by a user's network and their past engagement. If the user is connected to similar or like-minded people and shows a preference for certain news topics or positions, then their feed could cover a narrow and conforming range of news. On the other hand, OIs can facilitate the discovery of news which the news consumer would otherwise not view. Search engines and news aggregators present news from a variety of sources, and social media sometimes features a degree of 'automated serendipity' of news articles to prevent monotony.[4] OI users can also 'stumble' upon news browsing through search results or their social media feed without having intended to look for news, for example through news articles recommended by people to whom the user is connected but does not know well ('weak ties').[5] These features of OIs' services can potentially increase the diversity of a person's news consumption. The empirical literature has mostly demonstrated that news consumption accessed through OIs is more diverse in the sense that it covers a larger number of news outlets.

In this research, and in contrast to most of the literature, we focus on the diversity of news *topics* consumed by individuals. We use a diversity measure of news consumption, referred to as 'Shannon entropy,' which encompasses both the number of topics and how much news consumption is focussed on some topics over others. A person's news consumption is therefore considered more diverse if it covers a wider range of topics, and if one or few topics do not dominate the total news consumption.

While this is not the first study to analyse topic diversity in relation to OIs, it is to our knowledge only the second one to do so through analysing the content of people's actual browsing data. This complements our understanding of news diversity and allows us to assess news diets more directly than previous approaches based on the number or range of outlets people use. A person might read news from different – possibly politically opposed – outlets but with very similar content or on the same topic. The news diet would be diverse in terms of news outlets but narrow in terms of news topics.

---

[1] Ofcom, 2023, News Consumption Survey.
[2] Ofcom, 2024, Online news: research update.
[3] Helberger, Karppinen & D'Acunto, 2018, *Exposure diversity as a design principle for recommender systems*. Information, Communication & Society.
[4] Möller et al., 2020, *Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity*. Information, Communication & Society.
[5] Barberá, 2015, *How social media reduces mass political polarization. Evidence from Germany, Spain, and the US.* Unpublished; Cardenal et al., 2019, *Digital technologies and selective exposure: How choice and filter bubbles shape news media exposure*. The International Journal of Press/Politics; andFletcher & Nielsen, 2018, *Automated serendipity: The effect of using search engines on news repertoire balance and diversity*. Digital Journalism.

To measure topic diversity, we collected the news headlines viewed in an internet browser by a sample of approximately 8,500 internet users based in the UK over a one-month period in autumn 2021 and used natural language methods to group similar news headlines into topics. We then computed the topic diversity viewed by each person and related this measure to the share of news that they consumed through different OIs.

What we have found – in brief

In line with the literature, we find that greater use of OIs to access news correlates with higher outlet diversity. However, for topic diversity we find the opposite; more reliance on OIs (in particular social media and search engines) is associated with lower topic diversity. This evidence is consistent with concerns around the impact of OIs on the diversity of users' news diets.

This finding comes with some limitations. While it is based on a sample which is representative of the UK population based on some demographic markers, the population of people willing to have their browsing and app usage tracked could be different from the general population in ways we cannot measure.

We also observe only a subset of the news that is consumed by the individuals within our sample. Like most of the literature, we do not have information about offline news consumption, and we do not observe what news articles are presented to and viewed by users within social media platforms and news aggregators. We can only infer – with some uncertainty – whether a person arrived at a news article through an OI. We also stress that our results only document associations between diversity and how news is being accessed. No causal conclusions can be drawn from the data and our research design.

Related work

The literature on news and media diversity distinguishes different types of diversity.[6] A traditional focus has been on diversity at the market level, e.g., the number and variety of news-producing organisations, and the risk of the news media being dominated by one owner or voice.

However, the digitisation of news and the increasingly significant role played by OIs have shifted attention towards another aspect of diversity, exposure diversity, defined as the extent to which audiences are exposed to a diverse array of news content and sources. News diversity at the market level (e.g., the number of news outlets) can conceivably result in less diverse individual news consumption since the OI can draw on a larger pool of news to build a news feed specifically tailored to an individual.[7]

We focus our research and literature review on the diversity of news consumption, and the challenges to news diversity presented by the algorithmic personalisation of news. While the news media market and certain individual news outlets might be diverse in both variety and balance, the news that a user on an OI is exposed to could potentially be narrow and skewed.

Most papers on exposure diversity have looked at ideological or political diversity, reflecting the American context of a two-party system (and thus straightforward definition of diversity and related concepts) and high polarisation.

Flaxman, Goel & Rao (2016) analyse the browsing histories of American internet users and conclude that a user is more likely to consume cross-cutting news[8] through search and social media than through news accessed directly from the news outlet.[9] However, the ideological distance of news between individuals on social media is greater. The study emphasises that most instances of direct (and overall) access to news is through mainstream media. This would also explain why directly accessed news is less likely to be cross-cutting, but also less segregated. Fletcher, Kalogeropoulos, & Nielsen (2023) replicate this finding for a British panel of internet users; news accessed directly from the news outlet is more centrist, and at the same time less likely to include cross-cutting content.[10] The study also finds that the diversity of news outlets increases with the users' reliance on social media and search engines compared to directly accessed news.

Cardenal et al. (2019) analyse a Spanish panel of internet users. They find that news accessed directly from news outlets and news accessed through Facebook exhibit similar levels of cross-cutting exposure while news accessed through Google search engine increases the probability of cross-cutting exposure. Similarly, Wojcieszak et al. (2022) conclude for a panel of American internet users that search engines and social media are significantly more likely to expose people to cross-

[6] Voakes et al., 1996, *Diversity in the news: a conceptual and methodological framework*. Journalism & Mass Communication Quarterly; and Napoli, 1999, *Deconstructing the diversity principle*. Journal of Communication.
[7] Levy, 2021, *Social media, news consumption, and polarization: evidence from a field experiment*. American Economic Review, p. 851.
[8] We follow most of the literature and define 'cross-cutting news' is news that challenge or oppose the position of a reader, and 'like-minded news' as news that conform with the position of a reader.
[9] Flaxman, Goel & Rao, 2016, *Filter bubbles, echo chambers, and online news consumption*. Public Opinion Quarterly.
[10] Fletcher, Kalogeropoulos & Nielsen, 2023, *More diverse, more politically varied: How social media, search engines, and aggregators shape news repertoires in the United Kingdom*. New Media & Society.

cutting news than direct access.[11] Fletcher & Nielsen (2018) use survey data from the UK, the USA, Spain and Germany to demonstrate that people who use search engines for news discovery use more news sources and are more likely to use news sources from both ends of the political spectrum.

Finally, other studies considered the number of distinct news outlets as an outcome.[12] The emerging consensus among these articles is that outlet diversity for news accessed through OIs is at least as high as news accessed directly.[13]

Another strand of research has looked at news diversity on a specific platform, in particular Facebook. Interestingly, their findings on news diversity are often in contrast to the literature cited above.[14] González-Bailón et al. (2023) and Levy (2021) both report that news articles visited on Facebook are more segregated than news sites accessed directly. However, we note that their segregation measures capture the audience diversity of a news article, while our research focuses on the diversity of news articles viewed by individuals. Levy also finds that Facebook seems to promote more articles from like-minded than cross-cutting sources, even if the user follows both.

Bakshy, Messing, & Adamic (2015) and González-Bailón et al. (2023) both find that cross-cutting news content on Facebook goes through a 'funnel': a randomly picked news article has a good chance of being cross-cutting for a user, but a news article shared by the user's connection is less likely to be cross-cutting.[15] The likelihood of being exposed to a cross-cutting news article in Facebook's news feed and engaging with such a news article is lower still. Nyhan (2023) also finds considerable exposure of Facebook users to like-minded sources: 50.4% of a user's Facebook content comes from like-minded sources as opposed to 14.7% from cross-cutting sources.[16]

Very few papers have analysed topic diversity – the type of diversity that is the focus of this paper. Haim, Graefe & Brosius (2018) create artificial Google accounts with different preferences and browsing histories to compare the topic distribution on Google News across these accounts.[17] They find that these fake accounts were presented with news articles aligned with their preferences in their news feed, but that a news search containing the same search words produced a nearly identical selection and ranking of news articles across the different accounts. Möller et al. (2020) compare different recommender systems applied to news articles from a Dutch broadsheet newspaper and conclude that recommendation algorithms present a more diverse range of topics than human editors. Both these studies consider topic diversity in a stylised setting (e.g., using artificial Google accounts, and simulating article recommendations) rather than in the context of actual news consumption, leaving open the question of how diversity in consumed news differs

---

[11] Wojcieszak et al., 2022, *Avenues to news and diverse news exposure online: comparing direct navigation, social media, news aggregators, search queries, and article hyperlinks*. The International Journal of Press/Politics.

[12] Fletcher, Kalogeropoulos & Nielsen (2023); Scharkow et al, 2020, *How social network sites and other online intermediaries increase exposure to news*. PNAS; Stier et al, 2022, *Post post-broadcast democracy? News exposure in the age of online intermediaries*. American Political Science Review; andUlloa & Kacperski, 2023, *Search engine effects on news consumption: Ranking and representativeness outweigh familiarity in news selection*. New Media & Society.

[13] Ross Arguedas et al., 2022, *Echo chambers, filter bubbles, and polarisation: a literature review*. Oxford: Reuters Institute for the Study of Journalism, p. 17.

[14] González-Bailón et al., 2023, *Asymmetric ideological segregation in exposure to political news on Facebook*. Science.

[15] Bakshy, Messing & Adamic, 2015, *Exposure to ideologically diverse news and opinion on Facebook*. Science.

[16] Nyhan et al., 2023, *Like-minded sources on Facebook are prevalent but not polarizing*. Nature.

[17] Haim, Graefe and Brosius, 2018, *Burst of the filter bubble? Effects of personalization on the diversity of Google News*. Digital Journalism.

across different access and discovery modes. Closer to our own research, Jürgens & Stark (2022) use content analysis to classify news articles into topics and analyse how OI use relates to news topic diversity.[18] They look at a panel of German news consumers and find that the more often a person uses OIs the more diverse their news diet is (the 'within effect'). On the other hand, people who use search engines and certain social media more have less diverse news consumption when compared to people who use them less (the 'between effect').

We conclude this literature review with two notes of caution. First, the literature cited above is for the most part correlational. Users might be using different access channels for different types of news, leading to spurious differences in diversity across discovery channels. Some papers have experimental designs, but their outcome of interest is not diversity. Instead, they manipulate the news diet as an experimental treatment and focus their attention on outcomes such as polarisation. Second, the news media and OIs operate in a highly dynamic environment, and any research finding may be contingent on region, platform, and time.

[18] Jürgens & Stark, 2022, *Mapping exposure diversity: The divergent effects of algorithmic curation on news consumption*. Journal of Communication.

Data

Our main source of data is the Ipsos Iris web tracking panel, which tracks the web activity of a representative sample of UK adults over time.[19] Ofcom purchased one month of web tracking data covering the period between 15 September and 15 October 2021. The dataset comes as a table with one row for each visit to a website by an individual on desktop or a mobile device. The dataset does not record any content viewed on a social media feed or on an app. Thus, we do not observe news consumed directly on social media or on any app.

We filtered the dataset to only include visits to a pre-defined list of web domains that correspond to 23 news outlets in the UK. These are the same outlets as those included in Fletcher, Kalogeropoulos & Nielsen (2023), with the addition of iNews and CNN. The final sample contains close to 58,000 article headlines, and close to 230,000 article views (as some articles are read by several people).

Each visit to a news article on a provider website is categorised according to the route individual took to get to that article ('access mode'). We distinguish between the following access modes for a news article: direct; social media; search engine; news aggregator; and other. We infer the access mode for an article from the user's browsing and app usage history using the following algorithm:

- If a user accesses a homepage of a news outlet and afterwards opens a news article on that outlet's website, then we consider the access mode for this article to be direct.
- If the access mode cannot be classified as direct using the above approach, we proceed to assess whether it can be classified as an OI. If a user visits an OI website or uses an OI app and subsequently opens a news article on their browser (e.g., through clicking a hyperlink), then the access mode for this news article is OI (which we categorise as either social media / search engine / news aggregator).[20]
- To allow for the possibility that the user does not access the news article immediately after accessing the news outlet homepage or an OI – for example by opening a new tab on their browser before opening the news article – we also use this classification if the news article access is at most five steps after the visit to the news outlet homepage.[21] If more than one access mode is detected within these five steps, then we use the most recent (least distant in terms of steps) access. For example, if a user opens Google's search engine, then the BBC homepage, and two steps later an article on the BBC, then the access to this article is classified as direct.

If a first visit to a news article on an outlet's website is followed by a chain of other news article visits within the same outlet's website, then the access mode for the subsequent visits can be indeterminate. For example, if a user exhibits a browsing history of (social media -> news 1 on outlet A -> news 2 on outlet A) then news 1 has social media as access mode, but it is unclear whether news 2 should be classified as direct access, or as social media access. We thus follow Fletcher,

---

[19] Ofcom, 2022, Media Plurality and Online News Annex 5: Ipsos Iris passive monitoring data analysis.

[20] To classify the access mode as social media after using a social media app (rather than visiting a social media website) we also require the news article visit to be within five minutes of using this social media app. This is because accessing social media via using an app – unlike using a web browser to visit a website – does not allow for the possibility of leaving a tab open and coming back to it at a later stage to continue browsing; therefore, the delay reduces our confidence that the news visit originates from the social media app.

[21] Example: A person opens the website of news outlet on their browser tab X. They then open a new browser tab Y to look for holiday destinations. Then they go back to tab X and click on a link to a news article. If the person has spent up to five steps (websites) on browser Y, then this article will be classified as 'direct access.' Otherwise, it will be 'none' (see below).

Kalogeropoulos & Nielsen (2023) in that we only classify the access mode for the first news visit according to the rules above, but not the following news visits in a chain of news visits within the same outlet and within one hour (the access mode for these news visits is recorded as 'None'). We refer to such a chain within an outlet as a *news session*.

If a user reads a news article from an outlet and we cannot determine the access mode through one of the above classification rules, then the access mode of the news article is 'other' – for example, a link to a news article received via e-mail – unless the outlet has been visited within the past 24 hours. In the latter case, the access mode is again indeterminate ('None') as we cannot confidently interpret the news visit as a continuation of the past news session or as a new news session. Finally, we also do not classify a news visit's access mode if the same user has visited the same article in the past.

Importantly, even for news visits for which we cannot determine the access mode, we still classify the topic and the outlet of the news article. This information enters our computations for the news diversity measures for users. Of the articles for which we can identify an access mode, the distribution of access modes is shown in Table 1.

Table 1: Distribution of news sessions across access modes

Access Mode	Share
Direct	43.3%
Social	5.5%
Search	6.7%
Aggregator	0.5%
Other (unidentified)	44.0%

The final dataset used for this report therefore consists of all unique visits to the domains of major news outlets by members of the Ipsos Iris panel, tagged with the most likely mode of access. In total we identified 57,648 unique news articles, and 322,660 visits to news outlet domains. Of the 322,660 total visits, we were able to determine the access mode for 230,000.

Within this context, the analysis is focused on article headlines. This choice was made for both methodological and conceptual reasons. Firstly, we could more easily collect the article headlines than the article bodies due to access restrictions. Furthermore, article lengths vary considerably across and within news outlets and we are concerned that topic classifications might vary systematically by article length – especially if the article length is a feature of data truncation (e.g., where a paywall restricts access to some of the text of an article).

Secondly, article headlines are more likely to contain words and expressions which capture the gist of the news content since publishers select them to do so, and less likely to contain expressions which might make it more difficult for a topic model to determine clusters of similar articles such as

filler words. We therefore think that for our topic analysis, headlines are better suited than full article texts.[22]

Once all articles have been tagged with a topic (or as an outlier), we can begin to measure the diversity of the news diets of all individuals in the dataset. We have chosen individuals as the unit of analysis instead of all news visits for each mode of access (i.e., direct, social, etc.) for several reasons. Most importantly, OIs might show users a wide range of viewpoints collectively, but this might still result in low diversity at the individual level. Consider a news aggregator with two users, one who likes reading about politics and another who likes reading about sports. Even if the news aggregator only shows each person articles that they are interested in, the overall set of articles it recommends will cover a diverse set of topics. Consequently, we choose to measure the diversity of topics that individuals are exposed to – regardless of access mode – and relate this to the proportion of their news diet that comes from each access mode.

Following from Fletcher, Kalogeropoulos & Nielsen (2023) we measure diversity using entropy, specifically Shannon's H. Entropy can be regarded as a measure of the unpredictability of the topic of a randomly picked news article. Consider an individual who reads 10 articles. If all the articles they read come from the same topic, then Shannon's H is 0 (we can predict the topic with 100% certainty). If all articles are about different topics, then Shannon's H will be higher to reflect the greater unpredictability.[23] Formally,

$$Entropy = -\sum_{i} p_i \log_2 p_i$$

Where p_i is the proportion of an individual's news diet that comes from topic *i*. We have also included other measures of diversity, including Simpson's D which is equivalent to the Herfindahl-Hirschman-Index, as part of the robustness checks.

An ideal entropy measure for a person would be calculated over their probability distribution over topics. For example, if a person's probability of reading news on topic A is 1/3, and about topic B is 2/3, then these probabilities would enter the calculation in the entropy equation. In the sample we only observe the number of articles actually read over the sample period: the *sample* distribution of read articles across topics. In the limit, with the number of articles going towards infinity, this sample distribution would approach the 'true' probability distribution. However, the entropy calculated over few articles can exhibit severe bias. In the extreme, the entropy computed for any person who reads only one article is always 0. We therefore limit our sample to individuals who have accessed at least 10 news articles over the observation period, where we count any article from any of the 23 news outlets listed above as a news article. Out of a total of 8,592 individuals in the original dataset, we calculate entropy for 3,807 of them (the reason for this large drop is that a very large subset of the individuals only read a small number of articles).

For each of these individuals, we then calculate the share of their news sessions that come from each mode of access as defined in the previous section. We can then use this to relate the topic diversity of each user i's online news diet to the share of their news from each access mode:

(1) $$Entropy_i = \beta_{So} Social_i + \beta_{Se} Search_i + \beta_{Ag} Aggregator_i + \beta_{Ot} Other_i + \varepsilon_i$$

[22] Headlines are of course not fully immune against a confounding of topics. In an exploratory stage of this research, we observed that articles referring to a boxing *fight* match and articles referring to a certain court *fight* were often categorised under the same topic.

[23] In this example the index *i* runs from 1 to 10 (topics are numbered 1 to 10). Each of the ten topics will have $p_i = 0.1$ (10% of the articles are about any particular topic). Applying these numbers to the entropy formula, the resulting entropy is $-10 \times (0.1 \times \log_2 0.1) = 3.3$.

where $Social_i$ is person *i*'s share of news sessions from social media, and the remaining variables are defined analogously. The share of direct news sessions is the reference category. The estimated coefficients β therefore tell us how much a 1 percentage point increase in the share of an individual's online news sessions coming from each online intermediary at the expense of direct access is associated with a change in the diversity of their overall news consumption in terms of topics.

Topic modelling

The methodology for this analysis can be split into two components: natural language processing (NLP) and topic modelling. We first need to turn the news article headlines into a useable format for quantitative analysis. For this we use state of the art tools for extracting numeric features from text language models. Once we have numeric representations of each of the headlines in the dataset, we then proceed to using statistical techniques to identify clusters of similar headlines which we use as topics. We then identify the distribution of topics for each individual and summarise the diversity of topics using the entropy measure described above. This makes it possible to relate the diversity of topics in everyone's news diet to the share of online intermediaries in their news browsing sessions.

Natural Language Processing

The first point of departure between this work and other work in the area is the use of NLP to understand differences in the content of news. Information about the outlet that produced a news article can only tell us so much information about it, especially when it comes to the diversity of views to which users have access.

Traditionally, NLP has tended to involve analysing raw word counts or word counts weighted by their frequency in each document relative to the whole corpus of documents (we refer to this technique as 'tf-idf'). These methods can be successful for simple tasks, but they do not consider word order or context. Sentences with similar meanings that share few words in common will have very different representations and vice versa. Additionally, if the number of words in the corpus is very large then the word counts for individual sentences with 10 or so words may have many zeros. When we compare the similarity of sentences later, this can introduce measurement error by distorting the measured distance between sentences.[24]

Consider the following two sentences:

1. She likes biscuits.
2. He enjoys cookies.

NLP methods using word counts alone will fail to capture the similarity between these two sentences, because they do not share any words in common. On the opposite extreme, sentences that contain the same words but have different meanings will mistakenly be seen as similar:

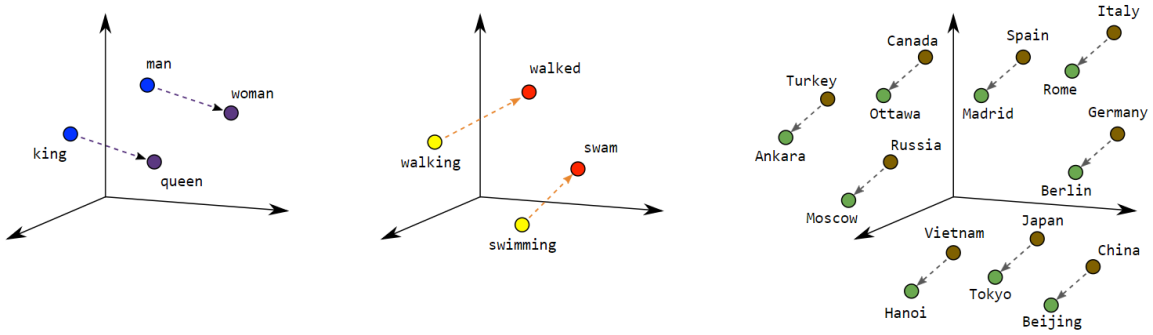
1. I sat on the sand by the bank.
2. I sat in the waiting room at the bank.

These issues can be partly addressed by making use of word embeddings. Word embedding models represent every word as a vector of numbers. The vectors are learned by deleting words from a sentence and then training a neural network to predict the missing word using the surrounding words. The model will learn that words that often appear together should be represented by vectors that are close together. This makes it possible to easily identify analogies like king is to queen as man is to woman or Paris is to France as London is to the UK. Additionally, words that have similar

[24] For example, if (dis)similarity is measured using the Euclidean distance between the tf-idf vectors, then having many entries equal to 0 will make a pair of sentences appear to be more similar than they actually are.

meanings are likely to be used in similar contexts and will therefore also be close together. Figure 1 below illustrates how this looks if words are represented by a 3-dimensional vector.

Figure 1: Illustration of word embeddings



Male-Female Verb Tense Country-Capital

Source: Google.[25]

Returning to the pairs of sentences above, word embeddings solve the problem with the first pair, but not the second. This is because the word embeddings for 'he' and 'she' are likely to be close to each other, as will the embeddings for 'likes' and 'enjoys' and 'cookies' and 'biscuits'. On the other hand, the word embedding for "bank" is always the same regardless of whether it has a different meaning in context. This is where NLP models that make use of the transformer architecture come in.[26] They combine word embeddings within each sentence together by computing a weighted average that takes the meanings of the other words in the sentence into account. Transformers can pick up, for example, that the use of the word 'sand' in the same sentence as the word "bank" suggests we are talking about a riverbank rather than a financial institution. This distinguishes transformer models from more commonly used language models such as LDA.

Since they address our theoretical concerns about using count-based models and have shown state of the art performance in identifying similar sentences, we have chosen to use sentence transformers for our analysis.[27] We chose the all-MiniLM-L12-v2 model given its strong performance and small size.

The usefulness of sentence embeddings for content analysis of news is best illustrated with a simple example. Table 2 below shows the similarity in headlines based on the Euclidean distance between each pair of sentence embeddings for 5 fictitious headlines. Scores are scales to range from 0 to 1. A value of 1 indicates that two sentences are exactly the same and a value of 0 indicates that they are completely unrelated.

[25] Google Developers, *Embeddings: Translating to a lower-dimensional space*.
[26] Devlin et al., 2018, *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805.
[27] Reimers & Gurevych, 2019, *Sentence-bert: Sentence embeddings using siamese bert-networks*. arXiv:1908.10084.

Table 2: Example of distances between news headlines

	Rising fuel prices are causing households hardship	Anger at expansion of low-traffic neighbourhoods	Russian army advancing on Kharkiv	Two soldiers killed in explosion in Kabul
Rising fuel prices are causing households hardship	1.00	0.21	0.02	0.02
Anger at expansion of low-traffic neighbourhoods	0.21	1.00	0.03	0.00
Russian army advancing on Kharkiv	0.02	0.03	1.00	0.14
Two soldiers killed in explosion in Kabul	0.02	0.00	0.14	1.00

Naturally, every sentence achieves a perfect similarity score with itself, but there are a couple of other patterns that emerge. The two headlines 'Rising fuel prices are causing households hardship' and 'Anger at expansion of low-traffic neighbourhoods' have the highest similarity score (0.21), presumably because they are both related to traffic and transport. Similarly, the two headlines 'Russian army advancing on Kharkiv' and 'Two soldiers killed in explosion in Kabul' have a relatively high similarity scores of 0.14 because they both relate to events around armed conflict. However, the headlines relating to traffic show little similarity to the headlines relating to armed conflict. If we were to crudely partition this set of five headlines into topics using their similarity scores, we would therefore end up with two topics, 'traffic' and 'armed conflict'. There are however much more sophisticated ways of doing this called topic modelling.

Topic modelling

Topic modelling broadly consists of three steps. First, raw text must be turned into a vector representation. In our case this means calculating sentence embeddings using sentence transformers. The next step is to reduce the dimensionality of the resulting embeddings. This is not strictly necessarily, but especially with the more complicated language models the sentence embeddings can be as large as 768-dimensional vectors. Some of these values will be important for distinguishing individual headlines from each other, but others will not.

Dimensionality reduction algorithms collapse as much of the important sources of variation between sentence embeddings as possible onto a small number of dimensions (in our research we chose five dimensions). The simplest way to do this is with Principal Component Analysis, which finds linear combinations of the input vectors that explain most of the variance. Along with most current work on topic modelling, we opt instead to use a non-linear method called UMAP that aims to preserve

the distances between individual sentences and has achieved state of the art results in identifying clusters in high-dimensional data.[28] This is standard practice among embedding-based topic modelling methods.[29]

The value of dimensionality reduction for topic modelling is that it substantially reduces the computational burden to identify clusters. Instead of computing the similarity between two 768-dimensional vectors, we simply do it with 5-dimensional ones without losing too much of the original information.

Once we have applied dimensionality reduction to the sentence embeddings, we then use cluster analysis to identify groups of headlines that are most like each other. We used a hierarchical clustering algorithm called *hdbscan* for this.[30] It has two advantages over alternatives for our purposes: it classifies headlines that do not clearly belong in any of the topics as outliers and it does not try to ensure that the clusters it identifies are all the same size. This means that some clusters of similar headlines that cover more popular news stories (such as the coronavirus epidemic) can be larger than ones that cover less popular stories or stories which attract less coverage (such as the volcanic eruption in the Canary Islands).

The number of clusters and the number of elements in each cluster are determined by the clustering algorithm. The user can specify parameters which govern how strict the algorithm is in considering a group of points to be a cluster, such as the minimum number of points in a cluster, and how close two points would need to be to be considered part of the same cluster. The user can also specify the exact number of clusters. If this number is smaller than the number of initial clusters, then the algorithm starts merging the most similar clusters until the desired number of clusters is achieved.

In this instance our baseline model requires a cluster to have at least 50 points and uses the default settings for the remaining parameters. We then inspected the resultant topics visually and we examine several different restrictions on the number of topics as part of our robustness checks.

---

[28] McInnes, Healy & Melville, 2018, *Umap: uniform manifold approximation and projection for dimension reduction*. arXiv:1802.03426.

[29] Grootendorst, 2022, *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv:2203.05794; Angelov. *Top2Vec: Distributed representations of topics*. arXiv:2008.09740, 2020.

[30] Campello, Moulavi & Sander, 2013, *Density-based clustering based on hierarchical density estimates*. In Pei et al., Advances in Knowledge Discovery and Data Mining, pp. 160-172, Springer.

Results

Overall, we find that greater use of OIs to access news correlates with higher outlet diversity. However, for topic diversity we find the opposite: more reliance on OIs (in particular social media and search engines) is associated with lower topic diversity. The rest of this section sets out the details of our results and their findings.

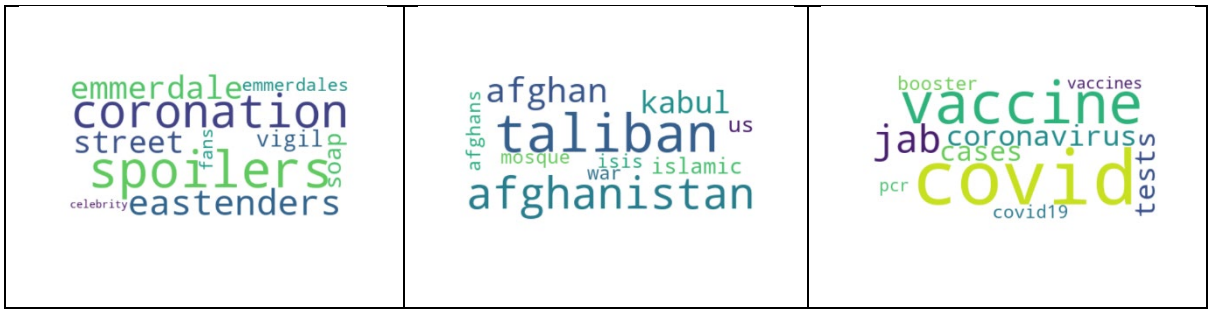
We first outline some of the major descriptive features of the topic modelling. In total, the baseline version of our model identified 106 topics, the largest of which contained articles about the reality TV show Married at First Sight with 2,606 articles and the smallest of which contained articles about broadband with 50 articles. Table 3 shows a snapshot of the five largest topics, the number of articles belonging to each, representative words identified by the model, and an example headline from the topic. The most popular topics for our sample of news consumers in Autumn 2021 were Married at First Sight, the petrol crisis, the Sarah Everard murder case, Westminster politics/Brexit, and assorted book/TV/film reviews.

Table 3: Top 5 topics and their representations

Topic number/name	Number of articles	Representative words	Example headline
1: Married at First Sight	2,606	Katie, married, she, her, sight, Kardashian, first, Stacey, price, at.	Married At First Sight UK: Morag is asked why the 'old Luke' wasn't good enough for her.
2: Energy crisis/driver shortage	1,874	Energy, petrol, fuel, crisis, gas, climate, drivers, shortage, bills, driver.	Energy crisis UK: Which energy suppliers have gone bust and why?
3: Sarah Everard murder case	1,441	Sarah, Couzens, Everard, Wayne, murder, police, jailed, Everard's, killer, man.	Police officer Wayne Couzens charged with murder of Sarah Everard appears in court.
4: Politics/Brexit	1,365	Brexit, Starmer, Keir, Boris, EU, Labour, Johnson, conference, Ireland, Johnson's.	Labour conference 2021: Sir Keir Starmer takes fight to Boris Johnson in deeply personal speech.
5: Book/Film/TV reviews	1,355	Review, the, of, books, Netflix, comedy, music, and, best.	The week in theatre: A Number; The Visit; Alone in Berlin review.

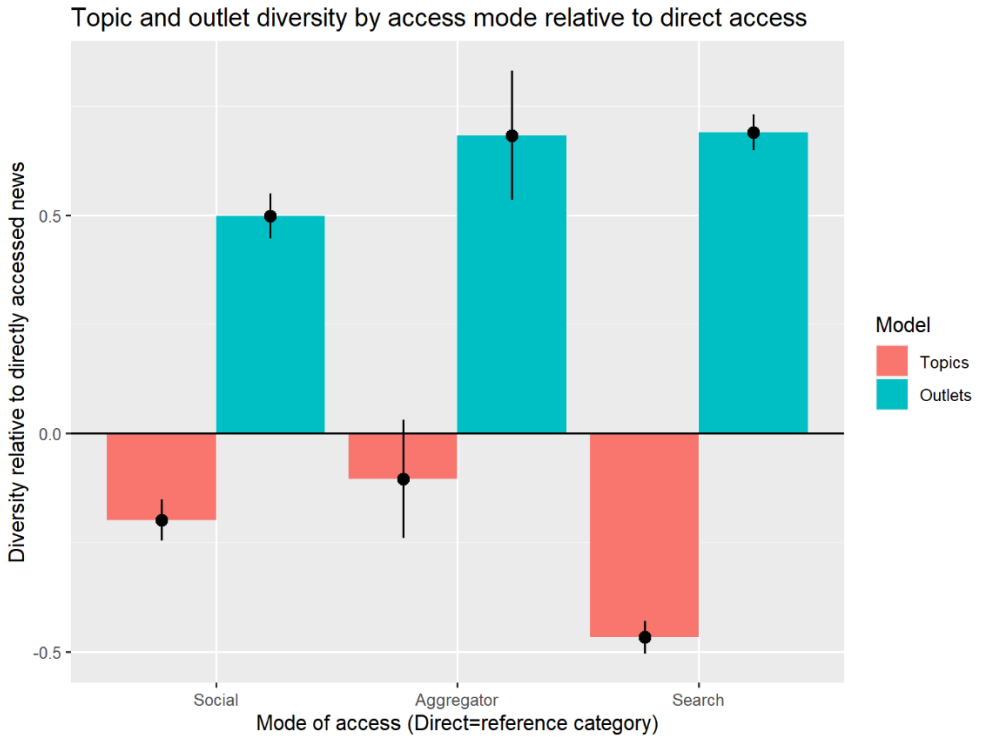
To give a clearer picture of the performance of the topic modelling, we have also generated word clouds showing the relative importance of the key words for each topic. Figure 2 below shows word clouds for three exemplary topics, one about the TV show Coronation Street, one about the war in Afghanistan, and one about the covid vaccine. The word clouds demonstrate that the characteristic words identified for a topic align with our intuition. For example, the word cloud about the war in Afghanistan groups together the words 'Afghanistan,' 'Taliban,' 'Kabul,' and 'war':

Figure 2: Examples of word clouds from topic models



Once we have tagged every article with a topic (or as an outlier), we then proceed to calculate the diversity of each individual user's news diet as described above. We do this for both the topics of news articles that they accessed and for the outlets that published those articles. This allows us to compare against the baseline of other research that has focused on the diversity of outlets. Figure 3 and Table 4 report the headline regression results. The y-axis in figure 3 represents the expected value of diversity for someone who gets all their news from that source, compared to someone who gets all their news from direct access. For ease of interpretation, we have rescaled the entropy values to range between 0 (for the lowest entropy in the sample) and 1 (for the highest entropy in the sample). Following previous findings, more intermediated news sessions are associated with greater diversity of outlets (green columns). But we see the opposite finding when we focus on the diversity of topics (red columns).

Figure 3: Topic and outlet diversity across access modes



The higher the user's share of a news sessions from social media or search, the less diverse the set of topics they are exposed to relative to the reference category of direct access. For news aggregators we do not find a significant negative association between news aggregator sessions and topic diversity.

These results are consistent with concerns about how different modes of news access curate and present news. A person who goes on a news outlet's homepage visits only one outlet but will see a variety of headlines on a wider range of topics, much as they would looking at the front page of a physical newspaper.

A social media platform on the other hand might identify the interests of the user and try to find articles to satisfy those interests, and in doing so cover a wider range of news outlets, but ultimately a narrower range of topics. This finding is consistent with growing concerns that social media drives echo chambers. We discuss these issues in more detail in our main report[31] and in our 2022 Discussion Document.[32]

The result for search engines may be impacted by personalisation, but there is a separate factor at play; on a search engine, a user indicates in the search term the topics in which they are interested. This could explain why we observe the lowest diversity scores for search-based news sessions.

The effect of news aggregators on topic diversity is negative, but not significant. The lack of significance combines a smaller point estimate (-0.10) compared to the estimates for social media and search engines (-0.20 and -0.47 respectively) and a wider confidence interval which reflects the small share of news aggregators' news sessions in the overall sample – our baseline algorithm only identified 0.6% of news sessions as being from aggregators. News aggregators tend to use a

[31] Ofcom, 2024, Online news: research update.
[32] Ofcom, 2022, Discussion document: Media plurality and online news, ('Discussion Document, 2022').

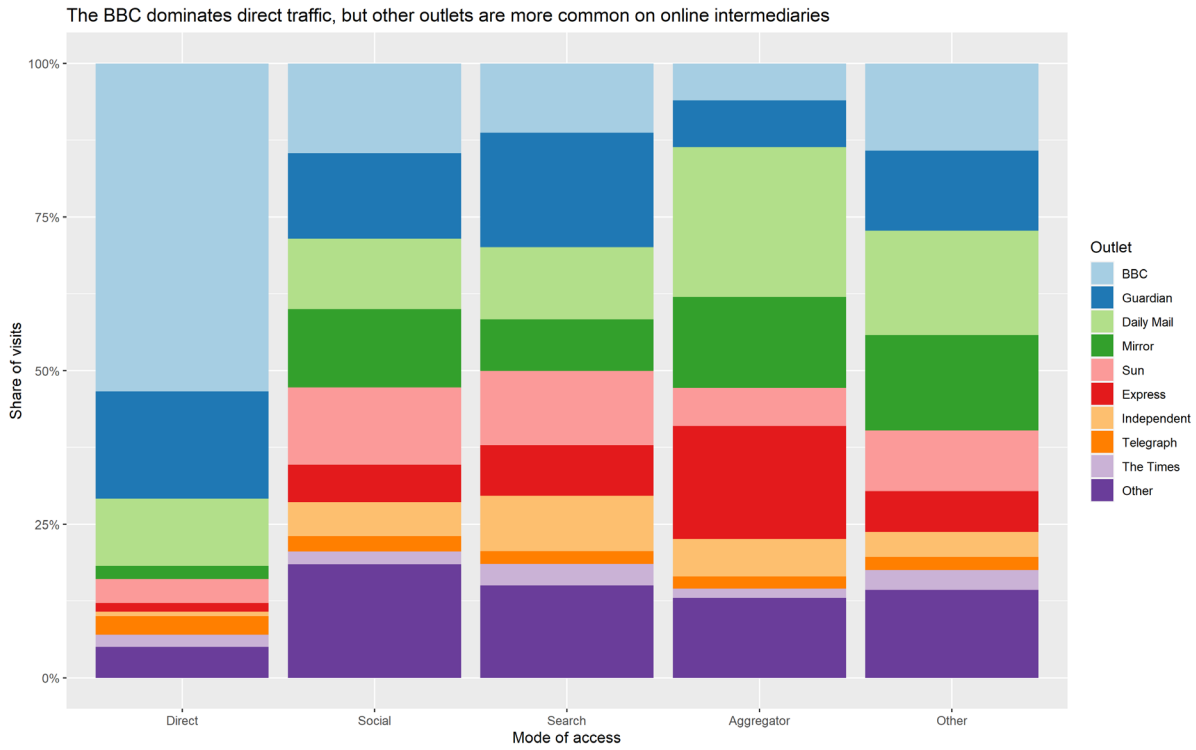
combination of editorially driven curation of news content and recommender systems on their services and in this respect, they may be more similar to a direct news source than social media and search engines.

Table 4: Regression tables for model in equation *(1)*

	Topic diversity		Outlet diversity	
Access modes	Estimates	Confidence Interval (95%)	Estimates	Confidence Interval (95%)
(Intercept)	0.66	[0.65 ; 0.68]	0.11	[0.10 ; 0.13]
Social	-0.20	[-0.25 ; -0.15]	0.50	[0.45 ; 0.55]
Aggregator	-0.10	[-0.24 ; 0.03]	0.68	[0.53 ; 0.83]
Search	-0.47	[-0.50 ; -0.43]	0.69	[0.65 ; 0.73]
Other	-0.21	[-0.23 ; -0.19]	0.28	[0.26 ; 0.31]
Observations	3,755		3,755	
R^2 / R^2 adjusted	0.181 / 0.180		0.301 / 0.301	

Diving deeper into the results, the distribution of outlets by access mode (Figure 4) indicates that lower outlet diversity in direct access is driven, in part, by the fact that the BBC takes up a very large share of articles accessed directly. Our finding here is in line with previous research (Fletcher, Kalogeropoulos, & Nielsen, 2023). OIs appear to be sending users to a wider range of news outlets but in doing so are not increasing the diversity of the topics that they access.

Figure 4: Shares of outlets across access modes

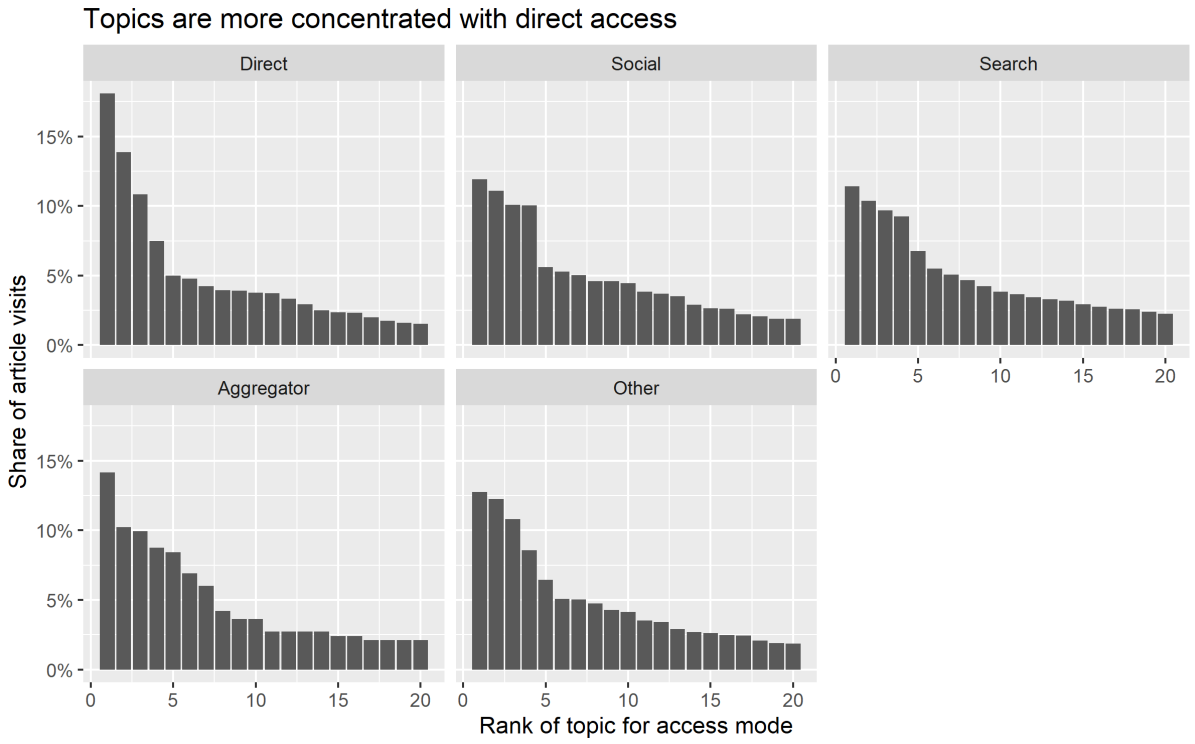


One possible explanation for this pattern is that social media platforms are using algorithmic curation to send users to articles from topics that more likely to drive engagement rather than the wider set of topics they might encounter on an outlet's website. This would be consistent with the high level of ideological segregation in news browsing observed on Facebook.

A key insight, which helps explain how these patterns can arise, is that diversity in the aggregate does not translate into diversity at the individual level. If everyone reads only about their favourite topic, and everyone's favourite topic is different, then we would observe a large diversity of topics in the aggregate, but no diversity at the individual level.

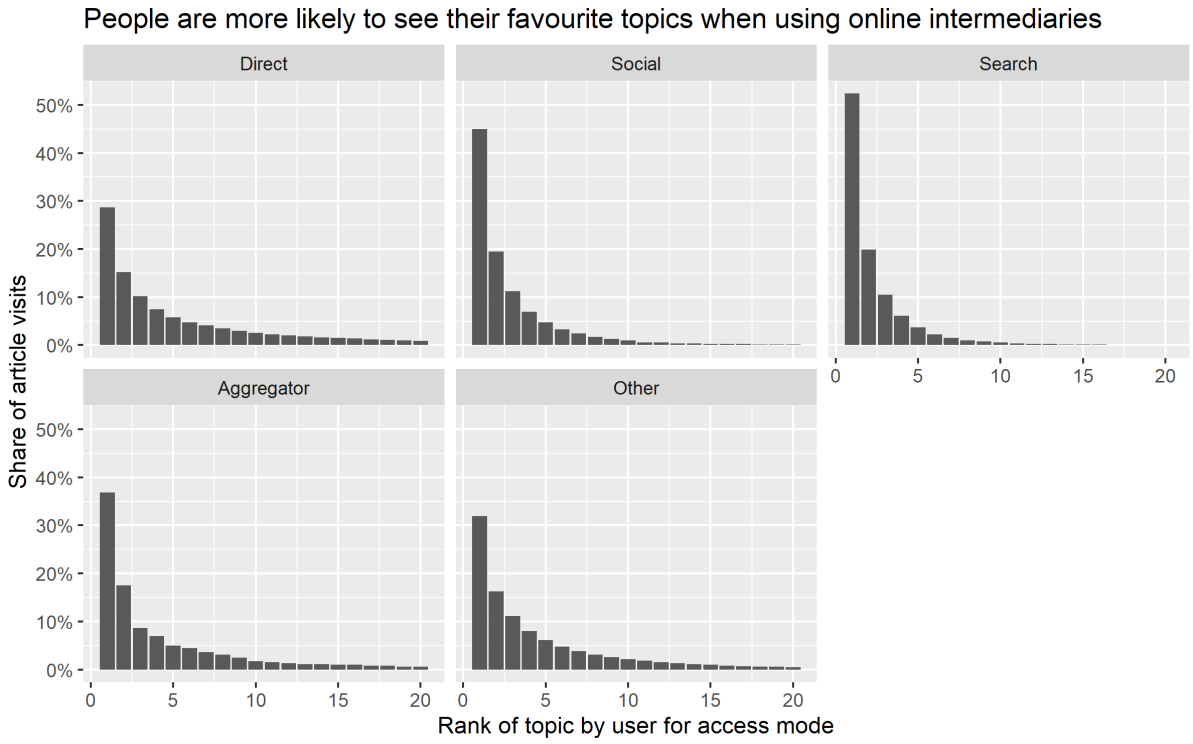
We illustrate this effect below. For each access mode, we count the news sessions for each topic, and rank the most viewed topics from left to right. Note that the rankings between access modes can be different. This gives us a sense of the distribution of topic popularity across the whole sample. Figure 5 shows those topic distributions. The most popular topic among directly accessed news sessions makes up approximately 18% of all directly accessed news sessions. For social media, the most popular topic garners approximately 12% of all news sessions started via social media. Calculated on this aggregated basis, it would appear as if news accessed through social media were more diverse.

Figure 5: Topic distributions for all news articles across access modes



We next count the news sessions for each topic and again rank them from left to right, but we do this *separately for each individual* and for each access mode. The favourite topic of one person can be different from the favourite topic of the next person, and OIs can tailor news recommendations to each person's tastes. We then aggregate these ranked topics (e.g., favourite, second most favourite, etc.) over all individuals and again produce the distributions of topics in Figure 6. Among directly accessed news sessions, close to 30% are on a person's favourite topic. However, among news sessions accessed through social media 45%, and among news sessions accessed through news aggregators more than 50% are on a person's favourite news topic. Thus, while OIs select news from a large pool of topics compared to direct access, they do so to tailor their news recommendation to every individual's taste, resulting in lower diversity at the individual level.

Figure 6: Individual topic distributions across access modes



Robustness of the results

To verify the robustness of our results, we considered several modifications to our baseline methodology. In earlier testing, we also considered using alternative sentence embedding models and different specifications for identifying the access mode of user sessions. We decided to use sentence transformers because of their state-of-the-art performance in the academic literature and because they did not require us to specify additional parameters to generate the embeddings. The different specifications for identifying access modes made no substantive difference to our regression results, which is consistent to what we found in our 2022 Discussion Document,[33] so we decided to proceed only with the base scenario.

We also tested whether our results still hold if we used an alternative method to quantify content diversity by taking the mean of the pairwise distances between the sentence embeddings of all the headlines that each user accessed as in Möller et al. (2020). We chose this modification because in early testing we found that our results were most sensitive to the hyperparameters of UMAP (dimensionality reduction) and HDBSCAN (clustering) in our topic modelling. Since the distance between sentence embeddings does not involve dimensionality reduction or clustering, we can avoid this source of instability altogether.

The results of this alternative method are in Table 5 below. The results echo the main finding that diversity is lower if news articles are accessed through social media and search engines.

[33] Discussion Document, 2022.

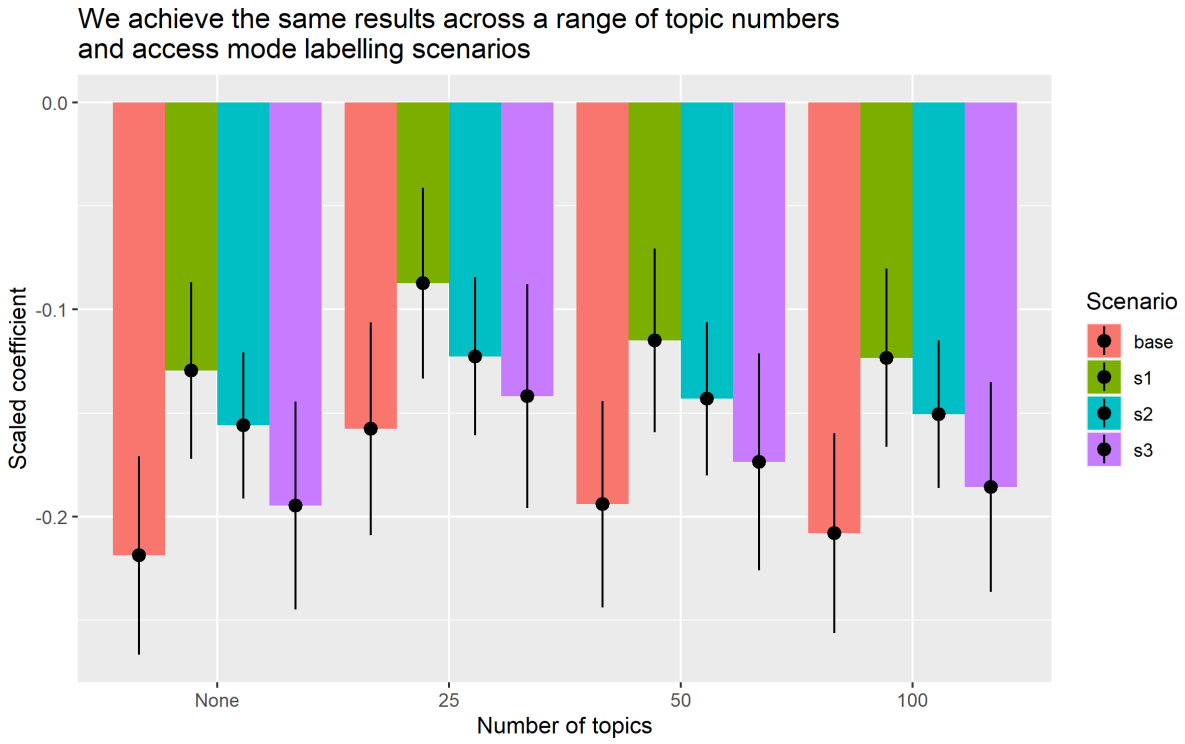
Table 5: Robustness of results

	Pairwise distances	
Access modes	Estimates	Confidence Interval (95%)
(Intercept)	1.29	[1.28 ; 1.30]
Social	-0.15	[-0.19 ; -0.11]
Aggregator	-0.18	[-0.28 ; -0.08]
Search	-0.27	[-0.29 ; -0.24]
Other	-0.16	[-0.18 ; -0.14]
Observations	5,186	
R^2 / R^2 adjusted	0.090 / 0.089	

We also re-estimated our baseline regression model using a) a variety of different algorithms for identifying the access mode[34] and b) a range of different constraints for the number of topics. For the sake of simplicity, we will only present the estimated coefficients for the social media share of news sessions, but we also found statistically significant results for search as in the baseline model. Since the number of topics systematically impacts the mean entropy, we also scaled the entropy values to range between 0 and 1. The results are presented in Figure 7. The estimated coefficient on social media is significant and negative in all cases.

[34] These alternative access mode classifications differ from the benchmark classification by varying the maximum time that we permit a news session to last (one hour in the benchmark classification) or the maximum number of steps which we allow a news article to be away from a homepage visit (five in the benchmark classification). See also Data section for a description of the benchmark classification.

Figure 7: Scaled coefficients on social media for robustness checks



Overall, these checks confirm that our results are robust to a variety of alternative specifications for topic modelling and alternative approaches to measuring the diversity of news consumption.