making communications work
for everyone

Understanding **online false information** in the UK

Economist discussion paper series, Issue number 2

Published 27 January 2021

Ofcom economics discussion paper series in communications regulations

The discussion paper series

Ofcom is committed to encouraging debate on all aspects of media and communications regulation, and to creating rigorous evidence to support its decision-making. One of the ways we do this is through publishing discussion papers, extending across economics and other disciplines. The research aims to advance our knowledge and to generate a wider debate on the themes covered.

Acknowledgements

We are extremely grateful to Professor Victoria Nash (Oxford Internet Institute & University of Oxford) and Will Moy (Full Fact) for helpful discussions, feedback and comments. We would like to thank NewsGuard and SimilarWeb for providing access to their data and for responding to our queries.

Disclaimer

Discussion papers contribute to the work of Ofcom by providing rigorous research and encouraging debate in areas of our remit. Discussion papers are one source that Ofcom may refer to, and use to inform its views, in discharging its statutory functions. However, they do not necessarily represent the concluded position of Ofcom on particular matters.

Contents

Section

1. Overview 3

2. Context for our research 7

3. Analysis of false information online 18

Annex

A1. Network traffic analysis 37

A2. Topic modelling 40

1. Overview

Internet, social media, mobile apps, and other digital communication technologies have revolutionised people's personal and working lives, generating significant benefits. They make information more readily available for people and in ways that were not possible before. However, there is growing concern about people's exposure to false or misleading information online, and the harm that this might cause to individuals and society. For example, the Reuters Institute recently reported that 63% of people in the UK are concerned about what is real and what is fake on the internet when it comes to news.[1]

As part of our duty to promote media literacy, we undertake research into how people access and use information online.[2] This paper contributes to the growing body of research on how to assess the availability and potential impact of false information online. We explore data sources and research techniques that can be used to investigate false information and present insights from this analysis.

Our analysis uses NewsGuard's assessment of websites that "repeatedly publish false information".[3] This enables us to classify a large volume of websites using common criteria. We rely on third party data as it is not practicable at this stage for us to reach a view ourselves as to what constitutes false information online. However, we recognise that different definitions and methodologies can result in different classifications. The findings presented in this paper should therefore be interpreted as reflecting NewsGuard's classification of false information, and not as an endorsement by Ofcom of NewsGuard's underlying methodology or the resulting classifications.

We believe that our research provides a useful contribution to the debate on this issue, and that regardless of definitions and classifications used, the insights gained from our analysis can inform future research.

What we have found – in brief

A range of factors allow false information to spread in new ways online. Lower production and distribution costs have supported a proliferation of media outlets and other providers of information. Online platforms enable people to easily access this range of information. Where online platforms give more prominence to content that captures people's attention, unscrupulous and politically motivated actors can take advantage of this to spread false information. Advances in online advertising, such as the automated allocation of online advertising space, have also inadvertently facilitated the monetisation of false information. In response to this, many online platforms have started dedicating resources to preventing or limiting the spread of false information.

[1] Reuters Institute for the Study of Journalism, 2020. *Digital News Report*, p19.
[2] Ofcom. *About media literacy*.
[3] As described in Section 3, NewsGuard is increasingly being used by professional journalists, technology companies and the advertising industry.

New research approaches can help us to understand false information online. The scale of content generated and shared online has increased to such an extent that new approaches might be needed to analyse this issue effectively. This paper explores how new insights can be gained by using machine learning techniques to analyse a combination of large datasets capturing: (i) a classification of websites that "repeatedly publish false information" according to NewsGuard ('false information websites'); (ii) UK user traffic to individual websites by SimilarWeb; and (iii) content available on false information websites.

Our research approach provides insights about false information in the UK which can inform future research. Between September 2018 and August 2020, the 177 false information websites in our sample attracted on average 14 million visits a month in the UK – a small but still significant volume of online traffic. A majority of these visits were concentrated within a few larger websites, with a long tail of smaller websites each receiving limited traffic. A significant fraction of traffic to the false information websites we analyse originates from people finding their way there via online platforms (such as search engines or social media).

Using machine learning techniques, we find that articles on the English language false information websites in our sample cover topics such as health, government, politics and conflict. While not all of these articles necessarily constitute false information, our analysis indicates that these websites report on topics that have the potential to harm individuals and society if they are falsely reported on.

Our analysis also suggests that levels of access to false information websites can vary by demographic group. For example, when compared to their share of the total UK population, younger age groups account for a disproportionate share of visits to the false information websites we analyse. Our analysis also indicates that male audiences account for a higher share of traffic to these websites than female audiences. Websites that receive more traffic from female audiences are more likely to cover topics related to health compared to those with predominantly male audiences.

Further research using different methodologies and data can complement our approach. Our analysis has some limitations. It uses a website-level classification from NewsGuard, and we recognise that a different classification could yield different insights. As our assessment is at the website rather than article level, our approach can only indicate the availability of content that might constitute false information. Richer datasets featuring article-level classification and traffic data would provide a more detailed understanding of false information consumed by UK users. Our analysis is also limited to textual data on websites. Analysing the consumption and topics of false information on other media (such as video) or in different settings (such as social media) could broaden our understanding.

Ofcom's related responsibilities, publications and activities

Ofcom's responsibilities

As part of our media literacy role, we seek to understand how people consume information across different media, including false information. We have a statutory duty to promote media literacy, which enables people to have the skills, knowledge and understanding to make full use of the opportunities presented by both traditional and new communications services. Media literacy also helps people to manage content and communications, and to protect themselves and their families from the potential risks associated with using these services. As part of this, we have a duty to undertake research into relevant matters.[4] Improving media literacy is important in countering the spread and consumption of false information and the harms that this can generate.

In its full response to the Online Harms White Paper consultation, Government confirmed it would appoint Ofcom as the Online Harms regulator, subject to parliamentary approval through the passage of Online Harms legislation.[5] Alongside a requirement for companies in scope to address 'disinformation' and 'misinformation' that poses a reasonably foreseeable risk of significant harm to individuals (e.g. relating to public health), the Government response also says that "the legislation will also introduce additional provisions targeted at building understanding and driving action to tackle disinformation and misinformation". Ofcom will continue to work closely with Government on the detail of the regime and will engage with stakeholders in due course in relation to our role as the future Online Harms regulator. We will also continue engaging with the Department for Digital, Culture, Media & Sport (DCMS) on their proposed media literacy strategy and wider work on tackling misinformation and disinformation online.

Related Ofcom work

This paper contributes to our research on the potential for harm online and media literacy of online users. It complements an existing body of Ofcom publications and activities related to false information and associated topics, including the work that Ofcom undertakes on these issues from the consumer perspective.

- Our Making Sense of Media programme aims to improve the online skills, knowledge and understanding of UK adults and children. We do this through providing robust research, collaboration with and coordination of relevant stakeholders and their activities.[6] As part of the programme we regularly publish bulletins summarising media literacy activities and initiatives – especially in relation to misinformation and disinformation – by a range of organisations in the UK and overseas.[7] As part of this programme we recently commissioned a rapid evidence assessment on the impact of media literacy initiatives in online misinformation. The objective is to understand what techniques and channels can effectively be used to mitigate online misinformation. The assessment will be published in the first half of 2021. As part of the same programme, we also conducted primary qualitative analysis aimed at understanding how people encounter and how they engage with both information and misinformation. This analysis, which took place between December 2020 and January 2021, will be published in the coming months.
- Ofcom has been responding to the Covid-19 outbreak. Since April 2020, we have published research findings on a regular basis which show how people are receiving and acting on information during the current pandemic, and their attitudes towards misinformation. We

[4] Ofcom. *About media literacy*.
[5] Department for Digital, Culture, Media & Sport (DCMS) and Home Office, 2020. *Online Harms White Paper: Full government response to the consultation*.
[6] Ofcom, 2020. *Making Sense of Media*.
[7] Ofcom, 2020. *Making Sense of Media bulletins*.

have also published a compilation of resources that help people to check the truthfulness of Covid-19-related information.[8]

- Our annual News Consumption Survey contributes to an understanding of news and information consumption across the UK and our Adults Media Literacy research provides evidence on media use, attitudes and understanding among UK adults. These surveys also analyse UK consumers' use of online news and media.[9]
- In December 2019 Ofcom submitted to the European Regulators Group for Audio-visual Media Services (ERGA – a network of EU regulators and an advisory body to the EU) the results of qualitative analysis of online platforms' compliance with their self-regulatory Code of Practice on Disinformation ('CoP') during the 2019 UK parliamentary election.[10]

Scope and structure of the remainder of this paper

Section 2 discusses definitions of false information and the possible harms that this can generate. We also describe key factors that have facilitated the spread of false information online, while recognising that these developments have also generated significant benefits for individuals and society.

Section 3 presents our findings on content and other emerging characteristics of a sample of false information websites accessed by UK users. We also discuss potential avenues for further research which could build on, and complement, our approach.

[8] Ofcom, 2020. *Cutting through the Covid-19 confusion*; Ofcom, 2020. *Covid-19 news and information: consumption and attitudes.*

[9] Ofcom. *News Consumption in the UK*; Ofcom, 2019. *Adult's media literacy*.

[10] ERGA, 2019. *ERGA report on disinformation: assessment of the implementation of the code of practice*.

2. Context for our research

Developments in online and digital technology have generated significant benefits and opportunities for people and society. However, they have also created opportunities for false information to spread in new ways, and more easily than before. This is why the spread of false information online and its potential to cause harm has been a growing source of concern among researchers, policymakers, journalists and the public.[11]

In this section, we provide some background on false information as context for our research and findings in Section 3. This is based on a review of academic literature and other relevant evidence. We discuss categories and definitions of false information, the possible harms this can generate, and some of the key factors that facilitate its spread online.

Categories of false information and their potential harms

Categories of false and misleading information

People seek information for a variety of reasons, including to make everyday decisions, to make sense of the situations they face, to interpret facts and events, to get an update on current events, and for entertainment.[12] Sometimes, however, information can be misleading, inaccurate, exaggerated, or completely false.[13]

There is an ongoing discussion on the terminology used to describe false information, but there is some consensus around the distinction between 'disinformation' and 'misinformation'.[14,15]

'Disinformation' is typically defined as false or misleading information that is produced with the intention to mislead or is spread deliberately when it is known to be misleading. The exact definition, however, can vary. For example:

- the Government Online Harms White Paper defines disinformation as "information which is created or disseminated with the deliberate intent to mislead";[16]
- the Department for Digital, Culture, Media & Sport (DCMS) Fake News Committee defines disinformation as "the deliberate creation and sharing of false and/or manipulated information";[17]

[11] UNESCO, 2018. *Journalism 'Fake News' & Disinformation*, p14.
[12] R. Savolainen, 2010. *Everyday Life Information Seeking*.
[13] S. Vosoughi, D. Roy, S. Aral, 2018. *The spread of true and false news online*.
[14] For a summary see: European Commission, 2020. *Study for the "Assessment of the implementation of the Code of Practice on Disinformation"*, Final Report, pp 14 to 18.
[15] Wardle and Derekshan identify a further category of harmful information, mal-information, which they say occurs "when genuine information is shared to cause harm, often by moving information designed to stay private into the public sphere". C. Wardle and H. Derekshan, 2017. *Information Disorder: Toward an interdisciplinary framework for research and policy making*, p5.
[16] HM Government, 2019. *Online harms White paper*, p22.
[17] DCMS, 2019. *Disinformation and 'fake news': final report*, p10.

- the European Commission Communication on tackling Online Disinformation uses the term disinformation to describe "verifiably false or misleading information created, presented and disseminated for economic gain or to intentionally deceive the public";[18] and
- the Italian communications regulatory authority – AGCOM – defines disinformation as "false information likely to be understood as true".[19]

'Misinformation' instead captures false information that is not created and/or distributed with the express intention of being misleading. For example:[20]

- both the UK Government's Online Harms White Paper and DCMS Fake News Committee define misinformation as "the inadvertent sharing of false information";[21] and
- AGCOM defines misinformation as "untruthful or inaccurate information content not created with malicious intent".[22]

The distinction between disinformation and misinformation is therefore the motive for producing or distributing it. In practice, such motivation can be difficult or impossible for a third party to identify with certainty and might have little bearing on the eventual impact that false information might have.

There is, however, no consensus on what constitutes false information or how to identify it in practice.[23] For example, there is a question of whether content that is 'spun' to suit a particular agenda or features false balance (for example when opposing viewpoints are presented as being more equal than the evidence allows) might be regarded as false information. [24] More fundamentally, reaching an agreement on whether an alleged fact is true or false can be challenging, particularly when trying to assess a large volume of material in tight timescales.

Potential harms caused by false information

What makes false information particularly concerning is not simply that it is false, but rather that people cannot always easily distinguish between information which is or isn't trustworthy. False and misleading information can be harmful if it prompts people to make choices and decisions that they would not have made, or enter mental states they would not have reached, if they had not been exposed to that information. [25]

[18] European Commission. *Tackling disinformation online*.
[19] AGCOM, 2018. *Rapporto Tecnico, Le strategie di disinformazione e la filiera dei contenuti fake*, p6.
[20] The European Commission Communication on tackling Online Disinformation does not explicitly define misinformation.
[21] HM Government, 2019. *Online harms White paper*, p23; DCMS, 2019. *Disinformation and 'fake news': final report*, p10.
[22] Ibid. footnote 19, p5.
[23] For a summary see European Commission, 2020. *Study for the Assessment of the implementation of the Code of Practice on Disinformation*, Final Report, pp 14 to 15.
[24] Different researchers have used a variety of characteristics to indirectly identify false information at scale, including the source of the news (whether it is reputable or not), the content (factually incorrect or distorted views) or the spreading method (e.g. if automated 'bots' social networks are used). Many of these questions and assessments unavoidably contain some element of subjective judgment, and therefore can create scope for legitimate disagreement.
[25] This is particularly true for what is sometimes termed 'junk science', i.e. arguments presented to the public as having been scientifically verified to a high standard, but which are in fact unreliable, ill-founded or misleading. These claims are often about human health, environmental safety or the benefits or risks associated with certain products. Junk science not only leads to people being misinformed about important issues; it can also undermine legitimate scientific evidence.

There are increasing concerns around the harms that false information online can cause to people and society. Prominent topics for false information include health and political issues, but in principle it can involve any topic. Harms that people might experience may be different depending on their personal circumstances and demographics (such as their age and education). Some examples of potential harms from false information include, but are not limited to:

- encouraging people to make decisions that could damage their health or that of others; [26]
- prompting people to make damaging economic or financial decisions;
- undermining respect and tolerance towards other people or even driving discrimination or hate;
- harming people's mental state or health (for example by causing anxiety or stress);
- damaging trust or undermining participation in social or democratic institutions and processes (such as elections); [27]
- undermining public confidence and trust in news and information sources; and
- otherwise generating confusion, uncertainty or doubt about historical, current or future events or trends, leading to damaging decisions or actions.

People in the UK recognise the potential harm associated with false information. For example, in 2020 the Reuters Institute found that 63% of UK adults were worried about what is real and what is fake on the internet when it comes to news.[28,29]At the same time there is also falling trust in some news sources. The same survey showed that the average level of trust in the news in general in the UK went down by 12 percentage points, from 40% in 2019 to 28% in 2020.[30] In 2020, only 39% of people in the UK said they trusted the news media they themselves use – in 2019 that figure was 46%. Trust in news from non-traditional media outlets, such as social media and search engines, was even lower. Consistent with this, the Ofcom 2020 News Consumption Survey showed that in 2020 only 35% of the people who regularly used them for news thought social media platforms were a trustworthy news source.[31]

[26] C. M. Pulido, L. Ruiz-Eugenio,G. Redondo-Sama,B. Villarejo-Carballido, 2020. *A new application of social impact in social media for overcoming news in health*. See also Ofcom, 2019. *Online markets failures and harms*, p32: "Disinformation specifically can cause harm by distorting views, preferences and decision-making. It can affect nearly any topic, but notable examples include the political sphere, science and healthcare (for example, anti-vaccine groups".

[27] For example, the European Commission explains that disinformation "may cause public harm", intended as "threats to democratic political and policymaking processes as well as public goods such as the protection of citizens' health, the environment or security". See Communication from the Commission to the European Parliament, The Council, The European Economic and social Committee of the regions, *Tackling online disinformation: a European Approach*, page 4.

[28] The Reuters Institute, 2020. *Digital News Report 2020*, p19.

[29] International research also shows concern about false information. For example, research conducted in Germany shows that 66% of German internet users surveyed have encountered politically motivated disinformation online, while 86% are concerned that online disinformation could manipulate election results, and 83% agree that political disinformation threatens democracy. They also found that 81% of internet users surveyed had encountered coronavirus disinformation online. See the Media Authority of North Rhine-Westphalia, 2020. *Information related behaviour for elections and political disinformation - Key research findings of the most recent Forsa study*.

[30] The Reuters Institute, 2020. *Digital News Report 2020*, p61.

[31] Ofcom, 2020. *News Consumption in the UK: 2020*, p72.

False information and coronavirus (Covid-19)

False information has been a significant concern during the coronavirus (Covid-19) pandemic. In response, the Government and major digital platforms have made efforts to tackle false information relating to Covid-19.[32] Ofcom is also providing a range of information and evidence about how people are getting news and information on the pandemic.

In late March 2020 we began a weekly online survey of around 2,000 people, asking them about their consumption habits relating to news and information about Covid-19. This work furthers the understanding around the access, consumption and critical engagement with news at this time, recognising that habits may intensify or change due to the nature of the crisis. We have been publishing findings from the survey on our website on a regular basis, as part of our Making Sense of Media programme.[33]

During the first week of Ofcom's survey (fieldwork between 27-29 March 2020) 46% of the respondents said that, in the week just before the survey, they had come across information or news about Covid-19 that they thought was false or misleading.[34] Of these, two-thirds (66%) said that they were seeing this sort of information at least once a day. Four in ten people surveyed said they found it hard to know what was true or false about the virus.

By week thirty-three of our survey (fieldwork between 6 and 8 November), 37% of respondents said that, in the week just before the survey, they had come across information or news about Covid-19 that they thought was false or misleading. The results showed a gradual decrease from a peak of 50% in weeks three and five. We saw notable variation in results by age. Sixteen to twenty-four-year-olds were the most likely to report coming across false or misleading information (42%) compared with 28% of over-65-year-olds. Over a third (37%) of male respondents reported seeing false information compared to 29% of female respondents.

Week thirty-three of the survey also found that half of respondents (51%) who use social media saw news or information that had warnings or notices from the platform attached, saying that the information may be untrustworthy or untrue. Of those who came across these warnings, over half (53%) said they clicked through to view the content that was flagged in this way.

Four in ten respondents (40%) said they find it hard to know what is true and what is false about the coronavirus, which is the same proportion as in week one. More female (45%) than male (36%) respondents said this. Those aged 16-24 (45%) were the most likely age group to agree with the statement. Less deprived households (35%) were less likely to say this than other socio-economic households.

Given the increased concern about misinformation during this time, we provide information about fact-checking and debunking websites and tools as part of our Making Sense of Media programme.[35]

False information on Covid-19 and 5G

[32] See, for example, UK Government, 2020. *Social media giants agree package of measures with UK Government to tackle vaccine disinformation*.

[33] Ofcom, 2020. *Covid-19 news and information: consumption and attitudes*.

One example of false information relates to claims that 5G is connected to the spread of coronavirus. Despite there being no credible scientific basis for these claims, in some parts of the UK mobile phone masts have been vandalised because of these incorrect claims. Engineers from mobile phone companies have also been harassed as they carried out their work.[36] In April 2020 Ofcom produced statements to address the myths around 5G and coronavirus, highlighting that there is no scientific evidence of any link between them.[37]

Developments in the digital era affecting news and information sources

Digitalisation has changed how news and information are created, stored and disseminated.[38] This has delivered vast benefits for individuals, organisations and society – for example, by making information more readily available for people and in ways that were not possible before. However, it has also contributed to the spread of false information.[39]

Developments that have been particularly relevant for the spread of false information online include lower distribution costs, the advent of online intermediaries, audience dispersion and new monetisation mechanisms through online advertising. We describe each of these in turn.

Reduced costs and increased number of media outlets

Changes in digital technology have had a sustained impact on the production, distribution and monetisation of news and information, particularly in recent years.[40] Distribution costs have decreased, especially for online media outlets (providers of news articles, interviews, videos and other content) and other emerging providers of information (such as prominent social media users including ‘bloggers’ and ‘influencers’). Their cost of news publishing is significantly lower compared to traditional media, as, for example, printing or physical distribution is no longer needed.[41]

Over time, these reduced costs have facilitated the entry of many different types of outlets and sources, ranging from edited online-only newspapers to non-traditional content providers, such as podcasters, vloggers, bloggers and social media influencers. Consequently, the number of news media outlets online has grown hugely in recent years.[42] However, lower costs have also facilitated the entry of media outlets and other providers of information with low standards or malintent.

[34] Ofcom, 2020. Half UK adults exposed to false claims about coronavirus.

[35] Ofcom, 2020. Cutting through the Covid-19 confusion.

[36] Ofcom, 2020. Ofcom update on 5G vandalism.

[37] Ofcom, 2020. Clearing up the myths around 5G and the Coronavirus.

[38] J. Pavlik, 2010. *The impact of technology on Journalism*, Journalism Studies.

[39] P. Napoli, 2019. *Social media and the public interests - Media regulation in the disinformation age*, Columbia university press.

[40] UNESCO, 2017. Protecting Journalism Sources in the Digital Age.

[41] Ahmad, 2019. The decline of conventional news media and challenges of immersing in new technology.

[42] J. Chan, D. F. Stone, 20129. *Media proliferation and partisan selective exposure*, Public Choice.

The advent of online intermediaries and their use of algorithms

Search engines and news aggregators

The internet has enormously increased the choice of media and content available to audiences. The potential number of sources of information far exceeds what a reader could access in the pre-internet era. This means that the ability to effectively search and find information has become of critical importance to internet users.[43]

Search engines (such as Google and Microsoft Bing), and news aggregators (such as Google News) have filled this gap by offering important aggregation and intermediation services to audiences. They offer media outlets and other information providers the possibility of a wider audience and enable audiences to access and navigate news and information from a very large number of sources.[44] They help audiences to find and explore news, information and other content using algorithm-driven rankings. These algorithms rank content based on a combination of criteria, which can include: popularity with the wider public; relevance to the user's query; and a user's previously revealed preferences and behaviour.

Social media platforms

Social media platforms also act as intermediaries for news and information. They have become a prominent source of news and information in many countries.[45,46] The Ofcom News Consumption Survey for 2020 shows that 45% of UK adults consume news via social media, either directly on the platform or after being redirected to a website from a social media service. Facebook is the leading source of this in the UK, as it is used by 76% of those who consume news via social media.[47]

Social media platforms provide access to a wide range of content (such as pictures, videos, news and information) in a single place. Like search engines, the curation of content that appears on many social media platforms (public social media such as Facebook, Twitter, and Instagram) does not involve direct human judgment, but instead is the result of 'news feed algorithms'.[48] Similar to the ranking algorithms that search engines use, these algorithms are intended to classify and filter information to reflect what users may want to read. This process can take account of a user's pre-

[43] Y. Chen, G. Jeon, U. Kim, 2010. *A day without a search engine: an experimental study of line and offline search*.

[44] For example, compared the pre-digital era, people can more easily read different articles from a range of publishers instead of buying a bundled package of articles in a single print edition. See European Commission – JRC, 2018. *The digital transformation of news media and the rise of disinformation fake news*, *Digital economy working paper 2018-02*, p6.

[45] Facebook, for example, was launched in 2004 and reached more than one billion users in 2012. With over 2.7 billion monthly active users as of the second quarter of 2020, Facebook is the biggest social network worldwide. Similarly, Twitter was launched in 2006 and in 2012 more than 100 million users posted 340 million tweets daily. As of the first quarter of 2019, Twitter averaged 330 million monthly active users.

[46] Reuters Institute, *Digital News Report*, Reuters Institute and Oxford University Press, 2019. European Commission – JRC, *The digital transformation of news media and the rise of disinformation fake news.* Digital economy working paper 2018-02.

[47] Ofcom, 2020. *News Consumption in the UK*, p39.

[48] As opposed to 'private social media'/ messaging-type services such as WhatsApp, WeChat, and Instagram's Direct Messages, 'public social media' do not always allow users to choose who can share content and information with them. Users may therefore be communicated with or see content from 'connections of connections', paying advertisers or other platform participants based on the algorithms used by that social media service, in addition to their direct connections. Additionally, 'private social media' can nonetheless be an important vector for the spread of false information even if algorithms are less important for them.

identified preferences, with a view to maximising traffic, audience attention and advertising revenue.

Audience dispersion and the rise of programmatic advertising

The emergence of digital technologies has facilitated the development of new business models and new forms of consumer behaviour. In the media industry, for example, online media outlets can generate revenues in different ways to traditional media outlets.

Audience dispersion

Traditionally only a limited number of media outlets – such as TV channels, radio stations and print newspapers – could attract audiences and sell their attention to advertisers.[49] Lower costs and easier distribution have enabled more news outlets to emerge online, which has driven audiences to become more dispersed. This means that the audiences now divide their attention across a larger number of media outlets than in the past.[50] As a consequence, audiences spend a smaller proportion of their time with traditional media.[51]

Online advertising

As more online news outlets have emerged, and audiences have become more dispersed, new solutions have been developed that allow the automated allocation of online advertising across huge volumes of advertising spaces.[52]

Data-driven methods such as programmatic advertising are a key development, with advertising traded with this method accounting for 80% of all digital display advertising in 2017.[53] Programmatic advertising is the automated buying and selling of online advertising space. The automation makes transactions more efficient and effective as it provides a low-cost solution to target online audiences.[54]

Programmatic advertising allows virtually all media outlets with some audience to obtain some advertising revenues, even those with small audiences, and including those distributing potentially false and harmful information.[55]

The spread of false information online

These developments have delivered significant benefits to individuals and wider society. For example, search engines help people to quickly and easily access and explore very large amounts of

[49] Stigler Center, 2019. *Stigler Committee on digital platforms – Final report*, p158.
[50] P Napoli, 2019. *Social media and the public interests - Media regulation in the disinformation age, Columbia university press*.
[51] The Cairncross review, 2019. *A sustainable future for journalism*.
[52] Competition and Markets Authority, 2020. *Online platforms and digital advertising.*
[53] PwC & IAB, 2017. *Adspend Study*.
[54] Such solutions distribute adverts and advertising revenues across a network of online outlets. This can be done in accordance with subject matter or audience criteria specified by the advertisers. For instance, an advertiser might wish for their advertising to appear within or alongside content that meets certain keyword criteria. It could also require the advertising to be directed to individuals who have demonstrated certain online behaviours (e.g. searching or viewing a webpage for hiking equipment) or to those who are likely to meet certain age, gender and geographic criteria.
[55] Competition and Markets Authority, 2019. *Online platforms and digital advertising*, market study interim report.

information. Such developments also influence the way people interact with each other and encourage valued and meaningful interactions. They can allow anyone to reach potentially large audiences with their messages and content. However, they have also facilitated the spread of false information among online users. We now discuss several key factors which can contribute to this.

Users can be vulnerable to the spread of false information online

The developments described above provide an opportunity for false information to spread easily and rapidly in an online setting – especially among those audiences less inclined or able to judge the reliability of news and information sources.[56]

Search engines, news aggregators and social media help people to find content and information. Some research finds that many of the people using search engines and social media for news are typically exposed to more diverse content, compared to those who do not use these services.[57]

However, there are also concerns that personalised services based on algorithms can indirectly reflect users' cognitive biases, which in turn has the potential to facilitate the spread of false information.

Some studies find that people have the tendency to prefer information which confirms their pre-existing or like-minded views rather than those which challenge them – this is known as confirmation bias.[58,59] This can inadvertently be reflected in the results provided by search engines, news aggregators and social media, in so far as algorithms take account of users' own preferences.[60]

Moreover, people tend to be more likely to trust and share information consumed by their social connections, and show some tendency to more easily associate or bond with people similar to them.[61] Researchers have also argued that less diverse personal networks result in limited "social worlds in a way that has powerful implications for the information they receive, the attitudes they form, and the interactions they experience".[62]

Such aspects of social psychology can be particularly relevant to social media's news-feed algorithms, which can also incorporate data from a user's social connections. In such cases, higher priority can be given to content posted, shared or viewed by the user's virtual connections and networks.[63] This means that the information and content seen by users may be driven by the choices, biases and behaviour of other users (such as their network of contacts).

[56] Some evidence indicates that less educated people, younger people, or people who spend less time on social media are less likely to recognise false information. See H Allcott and M Gentzkow, 2017. *Social Media and Fake News in the 2016 Election*.

[57] R. Fletcher, R. K. Nielsen, 2018. *Automated Serendipity*; E. Bakshy, S. Messing, L. A. Adamic, 2015. *Exposure to ideologically diverse news and opinion on Facebook*.

[58] A. Hunt, and M. Gentzkow. 2017, Social Media and Fake News in the 2016 Election, Journal of Economic Perspectives.

[59] D. Flynn, B. Nyhan, and J. Reifler, 2017. The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics. Advances in Political Psychology.

[60] C. Thornhill, Q. Meeus, J. Peperkam, B. Berendt, *A digital nudge to consumer confirmation Bias*; see also, for example, Google, 2018. *A reintroduction to Google's featured snippet?*.

[61] See, for example, T.W. Chan, J.H. Goldthorpe, 2007. Social Status and Newspaper Readership, American Journal of Sociology; E. Bakshy, B. Karrer, and L. Adamic, 2009. Social influence and the diffusion of user-created content. Conference: Proceedings 10th ACM Conference on Electronic Commerce (EC2009), Stanford, California, USA, July 6--10, 2009.

[62] See, for example, M. McPhereson, L. Smith-Lovin, J. M Cook, 2011. *Birds of a Feather: Homophily in Social Networks*.

[63] Facebook, *News Feed today*.

The combined effect is that some users, despite the wide availability of diverse content, might engage online through community-like structures that focus on a very limited set of sources of information (selective exposure), despite efforts from platforms to prevent this.[64]

Recent research finds that the above dynamics (individual cognitive biases, algorithms, and social psychology) can play a pivotal role in the spread of false information.[65] They conclude that when users focus on specific narratives and join polarised groups they end up reinforcing their worldview, even when based on false information, and dismissing opposing information.[66] Such a phenomenon, in turn, can facilitate the spread of false information and has been found to be particularly significant in online settings.[67,68]

Some suppliers deliberately spread false information

For-profit disinformation

The rise of online advertising allows virtually all media outlets, no matter how small their audience, to attract some advertising revenues. While this has many benefits, including facilitating media plurality, it does mean that some media outlets or information providers with low standards or malicious intent can seek to monetise false information deliberately designed to attract the attention of audiences.[69]

Many online platforms dedicate resources to preventing or limiting the spread of false information.[70]

- Twitter has tried to put a spotlight on credible information by making it easier to find on the platform. An example of this is a tab which makes it easier to find the latest accurate information on Covid-19. The tab includes curated pages highlighting the latest news such as public service announcements.[71]
- Facebook has invested in a fact-checking programme to tackle the spread of false information on its platform. This is in collaboration with a range of certified independent third-party fact-checking organisations. The focus of the programme is "identifying and addressing viral misinformation, particularly clear hoaxes that have no basis in fact".[72]

[64] A. L. Schmidt et al., 2017. *Anatomy of news consumption on Facebook*.
[65] Ibid. footnote 64.
[66] W. Quattrociocchi , A. Scala A, C.R. Sunstein, 2016. *Echo chambers on Facebook*, Social Science Research Network.
[67] A. Bessi A, et al., 2015. *Trend of narratives in the age of misinformation*; A. Bessi A, et al., 2015. *Viral misinformation: The role of homophily and polarization*; M. Del Vicario, et al., 2016. *The spreading of misinformation online*; D. Mocanu, L. Rossi, Q. Zhang, M. Karsai, W. Quattrociocchi, 2015. *Collective attention in the age of (mis) information*; A. Bessi, et al., 2015. *Science vs conspiracy: Collective narratives in the age of misinformation*.
[68] A. L. Schmidt et al., 2017. *Anatomy of news consumption on Facebook*.
[69] C. Ryan et al., 2020. *Monetizing disinformation in the attention economy: The case of genetically modified organisms (GMOs)*, European Management Journal.
[70] DCMS, 2020. *Social media giants agree package of measures with UK government to take vaccine disinformation*.
[71] Twitter, *Know the facts*.
[72] Facebook, *Fact-Checking on Facebook*.

- Google announced in April 2020 that it was investing $6.5 million in funding fact-checkers and non-profit organisations fighting misinformation around the world, with an immediate focus on coronavirus.[73]

Despite these recent initiatives, some information providers succeed in exploiting how search and newsfeed algorithms operate to attract attention to content featuring false information. For example, where social media news feed algorithms give greater prominence to content which captures people's attention, this can be exploited by false information providers using strategies to maximise the viral potential of content.[74]

Some research also suggests that the spread of false information online appears to be facilitated by social media. Recent empirical literature concludes that editorial control of news distribution decreased following the emergence of social media, and that false information spreads faster on social media. [75,76]

Politically motivated disinformation

The use of false information for political motives is not new. However, it has been argued that online platforms can offer a cheap, efficient, and easy-to-access vehicle for influencing larger numbers of people.[77] There is a growing concern that some individuals or organisations with political, rather than commercial, objectives may take advantage of the opportunities offered by online platforms (especially social media) to use false information to improperly influence people's views and opinions.[78,79,80,81]

Like providers who seek to generate revenues, these actors can design their content to capture people's attention and enhance its viral potential. This can be achieved in several ways, such as the use of 'bots' (or 'social bots'). These are social media accounts that operate according to programmed instructions, although human users may control them some of the time. They communicate autonomously on social media, often with the task of influencing the course of discussion and/or the opinions of their readers.[82]

User generated content can facilitate the unintentional spread of false information

Digital technology enables people to share and re-share 'user-generated content' on social media, potentially reaching very wide audiences. In some cases, this content can include false information,

[73] Google News Initiative, *Covid-19: $6.5 million to help fight coronavirus misinformation*.

[74] Facebook, News Feed today.

[75] H. Allcott and G. Matthew, 2017. *Social media and fake news in the 2016 election, Journal of Economic Perspectives*.

[76] S. Vosoughi, D. Roy, S. Aral, 2018. *The Spread of true and false information online*. This empirical research found that it takes true stories about six times as long to reach the same number of people as it does for false stories.

[77] R. Waltzman, 2017. *The Weaponization of Information: The Need for Cognitive Security, RAND Corporation*.

[78] The Policy institute centre for the study of media, communication & power, Kings College London, *Weaponising News*.

[79] J. Willemo, 2019. *Trends and developments in the malicious use of social media*, NATO Strategic Communications Centre of Excellence.

[80] D. Martin, J. Shapiro, 2019. *Trends in Online foreign influence efforts*.

[81] S. Bradshaw, P.N. Howard, 2018. *Challenging Truth and Trust: A Global Inventory of Organised Social Medial Manipulation, University of Oxford*.

[82] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, 2016*. The rise of social bots*.

even when the user who shared it has neither a profit motive nor the intention to improperly manipulate the political (or other) views of other users.

As discussed above, in some cases false information may be particularly prone to going viral if it exploits cognitive biases, for example by confirming pre-existing views or by being given particular credence due to being shared by close social connections.[83]

[83] See footnotes 64 to 68.

3. Analysis of false information online

This section presents the results of our quantitative analysis of false information online. Our key objectives are:

- to shed some light on the nature and potential harm that false information can generate by exploring large datasets and new research techniques; and
- to apply these techniques to the UK context and extract some emerging findings on the nature of false information that could inform future research.

This work is therefore intended to be an early step towards a better understanding of how we could assess the prevalence and impact of false information online.

In this section:

- we describe the different ways false information can spread online and identify which parts of the false information ecosystem are the subject of our analysis – namely websites that NewsGuard identifies as repeatedly publishing false information (for the sake of brevity 'false information websites');
- we describe our data sources, how they are used in our analysis, and their limitations;
- we analyse the scale of traffic to false information websites and compare this with traffic to websites that according to NewsGuard meet minimum standards of credibility and transparency (for the sake of brevity 'trustworthy information websites' henceforth);
- we identify the topics and the content on which false information websites focus. This provides an indication of the main areas and aspects of people's lives that might be affected by exposure to false information on these websites;
- we assess some demographics of users of false information websites and how users reach these websites. These are compared to equivalent figures for trustworthy information websites; and
- we conclude by highlighting what this analysis contributes, as well some of its limitations and possible avenues for further research.

Our analysis covers a sub-set of the online information ecosystem

People access information, and false information, from a wide and varied set of online sources, including websites, social media (such as Facebook), video sharing platforms (YouTube), aggregators (Yahoo!) and mobile applications (Apple News).[84,85]

As illustrated in Figure 1, there is a variety of content forms that feed through different media sources. Different media sources tend to focus on certain forms of content more than others. For

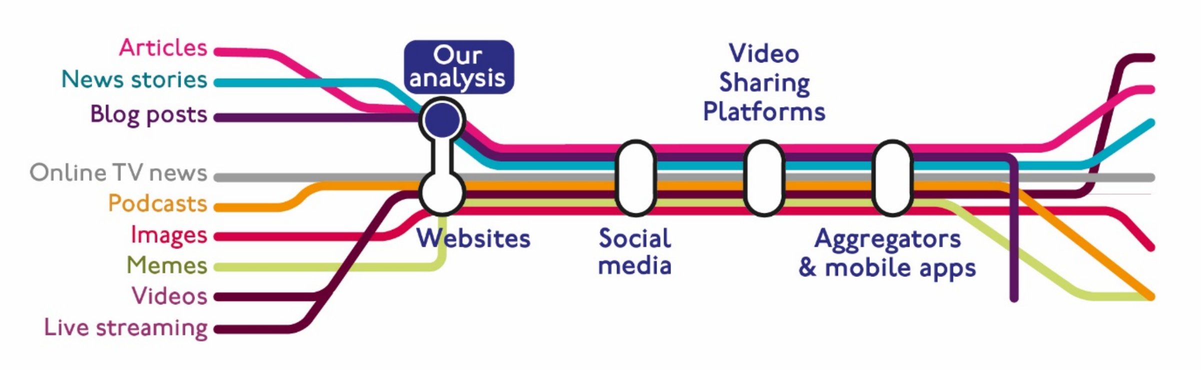
[84] University of Pennsylvania, Annenberg school for communication workshop, 2017. *Understanding and addressing the disinformation ecosystem*.

[85] A mobile application is a software application developed specifically for use on small, wireless computing devices (such as smartphones and tablets), rather than desktop or laptop computers.

example, video sharing platforms tend to feature visual content (such as videos, live streaming and TV news) while websites are focused more on written content (such as news stories, articles and blogs).

Nevertheless, it is worth noting that the same information can 'travel' across different sources in various forms. For example, a piece of information contained in an article published on a website can be shared on social media in a post or selected by aggregators as part of the information to be distributed to their users.

Figure 1: The online information ecosystem – sources, content forms and focus of our analysis



Source: Ofcom.

Our analysis focuses on one part of the online information ecosystem where people might be exposed to false information – articles, news stories and blog posts on websites, as highlighted in blue in the figure above. Our assessment therefore captures only some of the instances where people may consume false information but not others (for example through videos, images and podcasts). Similarly, it does not include potentially misleading advertisements or marketing, even where this was posted on websites that are the subject of our analysis. However, our news consumption survey suggests that websites play an important role in online news consumption by UK users, including being the most common way of accessing news online.[86]

Our analysis combines a range of data sources

We use NewsGuard's classification of websites that repeatedly publish false information

The scale of content online means it would not be feasible for us to individually and manually classify a sufficiently large volume of articles or websites.

To perform analysis at scale we therefore rely on third-party data from the prominent news rating organisation NewsGuard, which makes classifications at the website level based on human judgement. NewsGuard provides 'trust ratings' to help consumers assess the quality of the

[86] Ofcom, 2020. News Consumption in the UK, p45 and p53.

information they access online. It rates over 6,000 websites, which account for 95% of online engagement across the US, UK, Germany, France, and Italy.[87,88]

NewsGuard is used by several other organisations to identify false information. For example:

- in an effort to combat false information, Microsoft has partnered with NewsGuard and is providing its ratings through an integration in the Microsoft Edge mobile browser and a free plug-in for the Edge web browser; [89]
- AGCOM recently used NewsGuard intelligence to perform analysis included in its *Report on online disinformation – special issue on coronavirus*.[90] NewsGuard is also a member of the AGCOM panel on digital platforms and big data;[91] and
- in 2020, the advertising agency IPG Mediabrands partnered with NewsGuard globally to "help brands keep their advertisements off of misinformation".[92,93]

NewsGuard rates websites based on nine criteria of credibility and transparency. A website is given a green rating if it "generally adheres to basic standards of credibility and transparency", while red ratings are given to those that do not. For the purpose of our analysis, we use NewsGuard's trust rating as follows (as of September 2020):

- 3,291 websites are green-rated by NewsGuard and are not found to repeatedly publish false content (trustworthy information websites);
- 1,083 websites are red-rated by NewsGuard and are found to repeatedly publish false content (false information websites); and
- the red-rated websites which NewsGuard does not identify as repeatedly publishing false content are excluded from our analysis.[94]

NewsGuard's classification relates to websites rather than specific articles, so not all content on false information websites is necessarily false information. The opposite is also true: articles containing false information may occasionally appear on trustworthy information websites. In addition, NewsGuard's classification does not assess whether the information it identifies as false was uploaded with the intent to mislead. We therefore do not establish in our analysis whether websites deliberately spread false information (which could be defined as disinformation).

[87] NewsGuard, *We help you decide which news sources to trust — with ratings from humans, not algorithms*.

[88] In November 2018, a study measured the impact of news consumers having access to NewsGuard's ratings and labels. This was based on a panel of 25,000 users in the US who installed and used the NewsGuard extension during their daily online activities for one week. At the end of the week, panellists were sent a survey to provide feedback on their experience. Of the 706 panellist who provided such feedback, the overall finding was that they would be less likely to read or share news from red-rated websites and would be more likely to read or share news from green-rated websites.

[89] Microsoft is one of the signatories of the EU Code of Practice on Disinformation, a self-regulatory code of practice that aims to address the spread of online disinformation and fake news. European Commission, 2018. *Code of practice on disinformation*.

[90] AGCOM, 2020. *Report on online disinformation, Special issue on coronavirus*.

[91] AGCOM, 2020. *Soggetti aderenti al tavolo piattaforme digitali e big data*.

[92] This partnership was launched by the London-based IPG Mediabrands UK in 2019.

[93] NewsGuard, 2020. *NewsGuard and IPG Mediabrands expand partnership to help advertisers in the U.S. and Europe advertise on trustworthy news sites while avoiding misinformation*.

[94] This is to exclude from our analysis websites that fail other journalistic standards (e.g. transparency with regards to ownership or financing).

We recognise that what constitutes false information might change depending on the exact definition used, and this will often involve an element of subjective judgment. While our use of NewsGuard's classification enables us to perform our analysis, different classifications could prove to be equally or more valid. The findings in this paper should therefore be interpreted as reflecting NewsGuard's classification of false information and not as an endorsement by Ofcom of the underlying methodology or the resulting classifications.

We provide more detail on NewsGuard and its rating process in Annex 1.

We use SimilarWeb as our source of website traffic data

We use web analytics provider SimilarWeb to measure the level of UK user traffic to individual websites. There are a variety of such market intelligence tools, each using a unique methodology and approach to collect, process, and estimate website traffic data.

We have selected SimilarWeb as it has better coverage of UK traffic to the false information websites identified by NewsGuard. It is possible that results based on our analysis of SimilarWeb's data deviate from that of other providers that may take different approaches to measurement.

As explained in Annex 1, SimilarWeb analyses hundreds of sources to construct its statistics. These include a panel of users who allow SimilarWeb to monitor their internet activity as well as direct observations such as websites' own traffic statistics.

Despite this wide coverage, SimilarWeb does not hold statistics for some of the smaller websites in the NewsGuard sample, as these do not attract enough UK visits to be included in SimilarWeb's panel. Our analysis of website traffic is therefore limited to a subset of the websites classified by NewsGuard, covering 2,093 (out of a total of 3,291) trustworthy information websites and 177 (out of 1,083) false information websites from September 2018 to August 2020.[95] It is therefore possible that some of our findings do not reflect smaller websites that are excluded from our analysis.

In addition, the audience and the traffic metrics from SimilarWeb track adult users only, and therefore do not account for the engagement of anyone aged under 18 years old.

We scrape articles from false information websites

Starting from the list of the 177 false information websites for which SimilarWeb holds UK traffic data, we developed a web-scraping approach that allows us to download all of the articles (their title, body, and URL) contained on the homepage of these websites from August 2018 to August 2020. While we do not scrape the entirety of the websites, the articles available on the homepages should give us a representative sample of articles which would have featured prominently on these websites during the period analysed.

As we use these articles for the topic modelling analysis described below, we filter out some of the websites in SimilarWeb's list either because they do not contain text-based news (such as video websites) or because their articles are not in English. We also limit our analysis to websites which both have their content stored on the Internet Archive and which permit automated scraping of

[95] We were only able to retrieve 24 months of historical data with SimilarWeb, which is why our web traffic analysis omits the month of August 2018.

their content.[96,97] This further decreases the websites we included in our topic modelling analysis from 177 to 97.

See Annex 2 for more detail on our approach to data collection for the topic modelling analysis.

User engagement with false information websites in the UK

We assess the scale of user engagement with false information websites in our sample by measuring the number of visits to these websites by UK users.[98] This gives an indication of the level of potential exposure of UK audiences to false information as they visited this subset of websites. However, the number of actual visits to articles containing false information on these websites is likely to be lower, as not all their articles are necessarily false.

For the period between September 2018 and August 2020, we have assessed the total monthly traffic from UK audiences on mobile and desktop devices to the 2,093 trustworthy information websites and the 177 false information websites in our sample. Trustworthy information websites attract around two billion visits every month, while false information websites attract around 14 million visits every month (approximately a 140:1 ratio). This difference in total traffic in part reflects the fact that most websites in our sample are classified as trustworthy information websites.

That said, trustworthy information websites also have larger traffic levels per website than false information websites in our sample, as shown in Figure 2 and Figure 3 below.[99] On average a trustworthy information website attracts around 960,000 visits monthly, across mobile and desktop, while the average false information website attracts only around 80,000 visits (a ratio of 12:1). For both trustworthy and false information websites, mobile devices account for the largest share of traffic.

Monthly traffic in March 2020 was 17% higher than the 2020 average for trustworthy information websites. For false information websites, monthly traffic was 21% higher in April 2020 than the 2020 average. This significant increase in traffic may reflect the early stages of the Covid-19 pandemic. This is consistent with the analysis reported in our Online Nations report, which found that UK internet users spent on average 18% more time online in April 2020 than in January 2020.[100]

[96] For more information see: *The Internet Archive.*

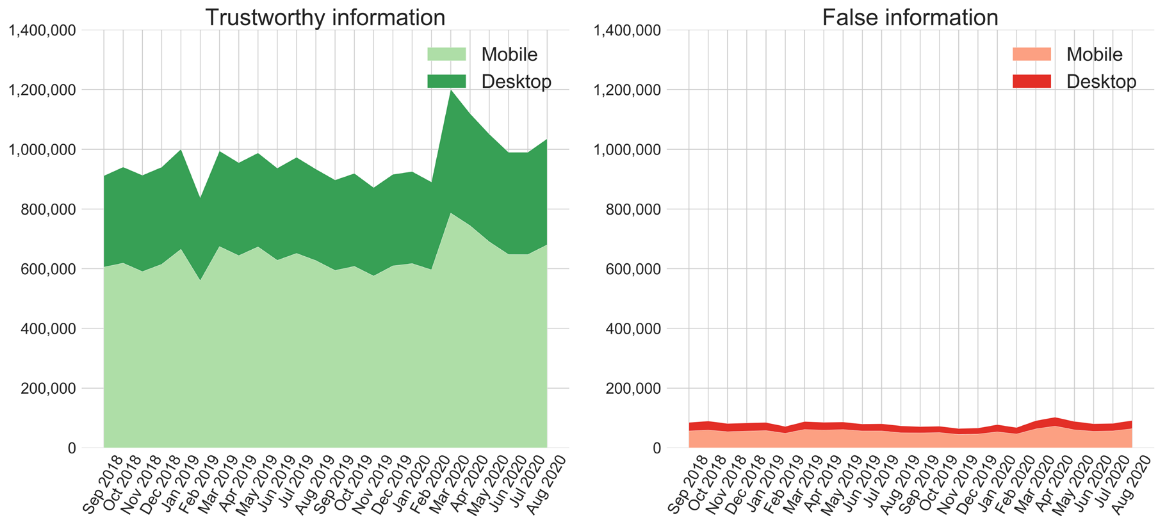
[97] We also do not scrape any content that is behind a paywall.

[98] SimilarWeb calculates a visit for a website if a visitor accesses one or more pages. Subsequent page views are included in the same visit until the user is inactive for more than 30 minutes. If a user becomes active again after 30 minutes, that counts as a new visit. A new session will also start at midnight.

[99] The dip in traffic observed in February 2019 is partly explained by the shorter length (i.e. the lower number of days) for that month.

[100] Ofcom, 2020. *Online Nation*.

Figure 2: Average monthly visits to information websites (by device), September 2018 to August 2020

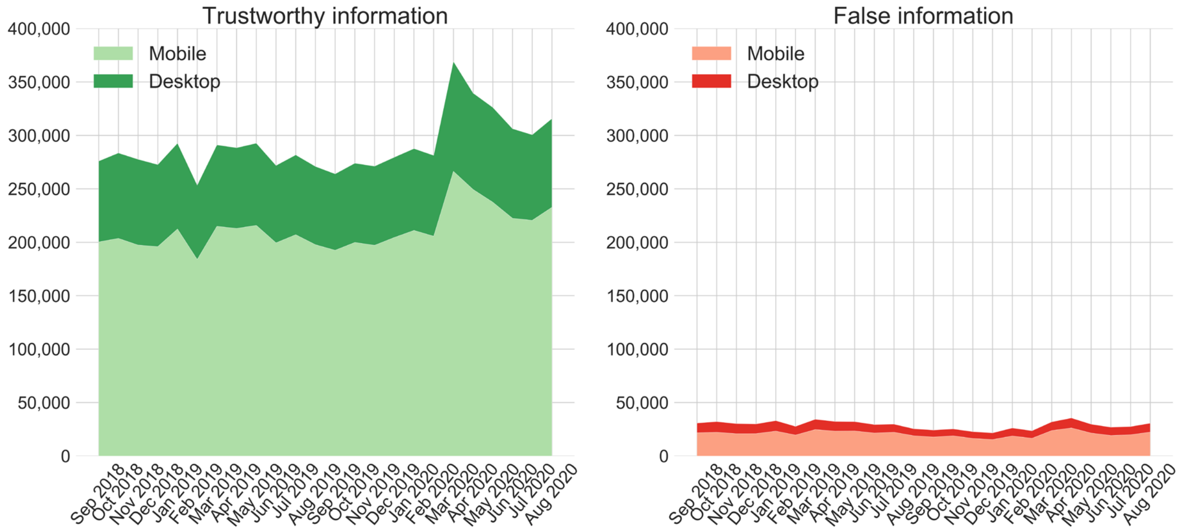


Source: Ofcom analysis of NewsGuard and SimilarWeb data.

In Figure 3 we plot unique visits to all of the information websites in our sample. Unique visits remove a device's second and subsequent visits to the same website in each month but include all visits to different websites in our sample made by the same device. So, for example, should a device visit two websites in a given month (no matter how many times), this will be counted as two unique visits.

Figure 3 shows that the trustworthy information websites in our sample on average attract significantly more unique visits, at least 250,000 per month per website compared to the 30,000 for false information websites. Therefore, our sample suggests that, on average, trustworthy information websites have significantly more users accessing their articles than false information websites in the same sample.

Figure 3: Average unique monthly visits to information websites (by device), September 2018 to August 2020



Source: Ofcom analysis of NewsGuard and SimilarWeb data.

However, there is substantial variation in the scale of traffic to individual false information websites. A majority of traffic from UK audiences into false information websites in our sample is accounted for by five websites. There is a long tail of small providers with relatively few monthly visits.

Table 1 shows the total and average monthly visits to the top five, the top ten and all other false information websites in our sample (the tail accounting for 167 websites). Table 1 shows that the top five websites alone accounted for about 54% of the total UK traffic to false information websites in our sample, and the top ten accounted for 66%. The ten largest false information websites on average attract 0.9 million visits per month each, which is similar to the monthly average traffic to trustworthy information websites.

This is in contrast with the long tail of 167 false information websites which fall outside of this top ten. With a total traffic of 4.8 million visits, these smaller players accounted for only 34% of total UK traffic to false information websites in our sample. On average this represents 29,000 visits per month per website.

Table 1: Monthly visits to false information websites in the UK (million), September 2018 to August 2020

	Total	Share	Average
Top five websites	7.5m	54%	1,505,000
Top ten websites	9.2m	66%	917,000
Others (167 websites)	4.8m	34%	29,000
All	14.0m	100%	79,000

Source: Ofcom analysis of NewsGuard and SimilarWeb data.

Such concentration is seen in other contexts online. For example, as part of our Online Nation research report we found that only seven websites with video sharing capabilities attracted more than ten million visits in a month (September 2019).[101] Many other websites with video sharing capabilities where visited by a more limited number of users – many by only a few thousand users.

Topics and content characteristics of false information websites

As discussed in the previous section, false information can generate harm to individuals and society. This is especially true for certain aspects of people's lives such as health and politics. It is therefore valuable to understand what topics are covered by websites which publish false information.

We use machine learning techniques to analyse a large body of content

We have applied a machine learning technique called topic modelling to a large sample of articles from NewsGuard's list of websites that repeatedly publish false information.

[101] Ofcom, 2020. *Online Nation*.

As NewsGuard's classification is at the website level, we cannot limit our analysis to articles that constitute false information.[102] This means our analysis captures both trustworthy and false information articles published on these websites. We also do not capture any false information that may have been be published on trustworthy websites in our sample. The analysis therefore gives an indication as to whether false information websites capture topics that may result in people taking harmful decisions, but it does not provide conclusive evidence of this.

We have used a topic modelling algorithm to analyse the words contained in a very large collection of English language articles, news stories, and blogs that were published on false information websites. The topic modelling algorithm we have implemented is called Latent Dirichlet Allocation (LDA), a machine learning technique that obtains a thematic summary of a collection of documents at a scale that would be impractical for a human to review.[103]

We have applied this technique on a sample of 406,267 articles from 97 websites which NewsGuard identified as repeatedly publishing false information as of September 2020. We selected articles which were published between August 2018 to August 2020.

Given that our sample is made up of articles, news stories, and blogs in the English language, their content will in principle reflect the interests of different English-speaking audiences around the world. However, we focus on the subset of false information websites that SimilarWeb holds UK traffic data for to try to better isolate news that are more likely to have been consumed by UK users.

Our topic modelling does not reflect that different individual articles may have differences in traffic or other measures of audience engagement. This is because our data only includes traffic data at the level of websites rather than individual articles. This means, for example, that an article that was read by thousands of people has the same weight as an article that was read by only ten people. As a result, our topic analysis can be characterised as looking at the 'supply' of false information by websites and does not attempt to capture how many users it reached.

We find that false information websites address a range of topics

Based on the methodology and the assumptions described in Annex 2, we find that content provided by the 97 false information websites in our sample can be grouped into 15 topics. Each of these topics is composed of a cluster of words identified by the algorithm.[104] To make it easier to interpret, we have manually assigned topic labels to each of these 15 groups based on an inspection of both their constituent words and a sample of the articles assigned to each topic. These are set out in Table 2 below.[105]

[102] Limiting our analysis to articles that are solely false information would be impractical, as it would require a manual, qualitative review of a every article in our sample to establish its veracity.

[103] The same methodology has also been applied by AGCOM to study the topics of false information websites in Italy. AGCOM, 2018. *News vs Fake in the information system*.

[104] LDA defines a topic as a probability distribution over words. Here we present the top 10 words that are most likely to fall within each topic.

[105] For more information on how we arrive at these labels and a more detailed description of the topics see Annex 2.

Table 2: Topic labels and relevant words

Topic label	Key words
Government/Control	Government, Power, Nation, Control, Right, Public, Freedom, Force, Order, Democracy
Health/Research	Study, Vaccine, Health, Research, Child, Patient, Cancer, Disease, Treatment, Risk, Drug
Politics/US	Trump, President, Democrat, Abortion, Republican, Biden, Election, Bill, Support, American
Lifestyle/Religion	Life, Child, School, Church, Feel, Love, Family, Woman, Christian, Parent, Mind
Media/Equality	Media, Black, News, Facebook, Tweet, White, Woman, Conservative, Racist, Hate
Economy/Finance	Company, Market, Money, Economy, Price, Business, Bank, Increase, Rate, Financial
Military/Conflict	Iran, Military, Israel, Force, Attack, Syria, Palestinian, Turkey, Iraq, Terrorist
Police/Protests	Police, Officer, Protest, Kill, Fire, Shoot, Family, Arrest, Local, Attack, Incident, Violence
Pandemic	Coronavirus, Covid, Health, Virus, Case, Death, Government, Pandemic, Test, Hospital
Crime	Case, Court, Child, Charge, Claim, Arrest, Federal, Judge, Epstein, Woman, Prison, Crime
Politics/Allegations	Trump, Investigation, President, Information, Evidence, Campaign, Intelligence, Email
Science/Technology	Water, Earth, Space, Energy, Scientist, Climate, Technology, Climate change, Research
Health/Nutrition	Food, Help, Health, Diet, Product, Skin, Water, Healthy, Body, Plant, Natural, Eat
Politics/Europe	Russia, United Kingdom, British, Europe, Putin, France, Germany, Brexit, Parliament
US Foreign Policy	China, India, Trade, Trump, Hong Kong, International, Tariff, Sanction, Deal, North Korea

Source: Ofcom analysis.

By inspecting the topics and related key words, we observe that the articles from the websites in the sample cover certain topics.

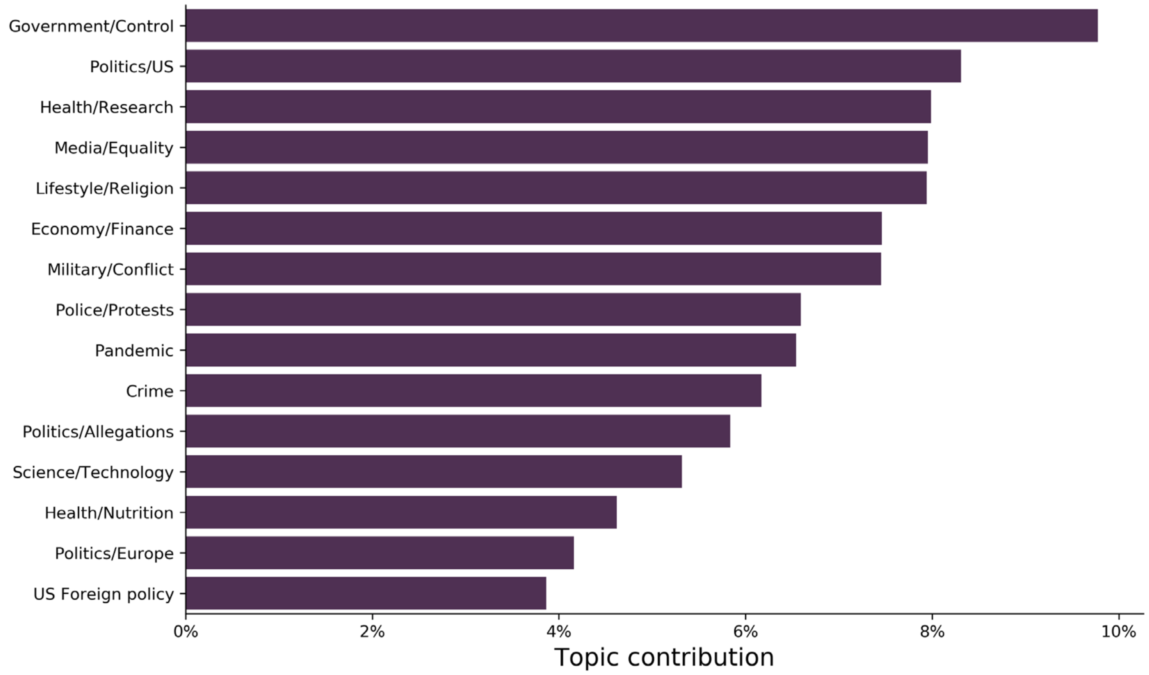
- Topics directly relating to health. For example, in the 'Health/Research' topic, our model assigns a high probability to words such as 'disease', 'cancer', 'risk', in conjunction with words such as 'vaccine', 'child', 'help'. By reviewing a sample of articles on this topic, we find that these words at times are used to provoke negative feelings. For example, several of the articles about vaccines focus only on their possible harmful effects (rather than any benefits).
- Topics associated with government, European and US politics. From a review of a sample of articles that capture these topics, we find examples of articles which point to controversy (e.g. allegations on politicians and public figures in the 'Politics/Allegations' topic) and conspiracy theories (e.g. the government conspiring to control its citizen in the 'Government/Control' topic). Several of the articles belonging to the latter topic adopt an 'us versus them' mentality, where the authors position themselves and their readers as 'heroic' figures trying to fight an evil system.[106]

[106] J. Roozenbeek, and S. van der Linden, 2019. T*he fake news game: actively inoculating against the risk of misinformation*. Journal of Risk Research, 22(5), pp.570-580; and J. Atkinson, 2005. *Metaspin: Demonisation of media manipulation*. Political Science, 57(2), pp.17-27.

- Topics related to conflict (including protests, war and international disputes), which can cause people to be worried our apprehensive.

In Figure 4 below we present the frequency of the topics identified in Table 2 in the collection of all articles in our dataset.[107] From this we can see that the topics concerning government, politics, and health are the most prevalent in our sample of articles, with about 30% of all the articles in our sample being assigned these topic labels.

Figure 4: Topic prevalence for all the articles in our sample



Source: Ofcom analysis.

If the articles on these topics were spreading false information, they could generate harm to individuals and to society, including by: inducing people to make decisions that could damage their health; exacerbating societal divisions; and damaging people's trust in media outlets and democratic institutions.

For example, when reviewing a sample of articles assigned to the 'Pandemic' topic, we found some articles that present theories which suggest that 5G is connected to the spread of the coronavirus. There is no credible scientific basis for these types of claims, yet as we discuss in Section 2, they have resulted in mobile phone masts being vandalised and engineers from mobile phone companies being harassed in the UK. This has led Ofcom to speak out publicly against these theories, due to potential knock-on effects for emergency and volunteering services essential during the pandemic.[108]

These harmful impacts could be aggravated where the content is particularly effective in attracting people's attention. For example, the fact that we see evidence of articles which seek to inspire negative feelings points to a potential for these articles to be particularly engaging. This is consistent

[107] This figure aggregates the topic distributions of each individual article, and therefore represents the prevalence of each topic in the collection of articles.

[108] Ofcom, 2020. *Clearing up the myths around 5G and the coronavirus*.

with existing research finding evidence of a negative slant in media coverage, driven by the fact that people tend to exhibit a preference for negative news.[109] This is rooted both in psychology (people tend to weight negative information more heavily than positive information) and rationality (negative information is perceived as more useful than positive information).[110]

Articles relating to pandemics, police and protests spiked in recent months

To assess the performance of our model, we check whether our topic model responds in predictable ways to events over time. Figure 5 plots the share of articles for two topics over the time period covered by our sample (August 2018 to August 2020). It illustrates how these topics from our model respond in predicable ways to external events. For example:

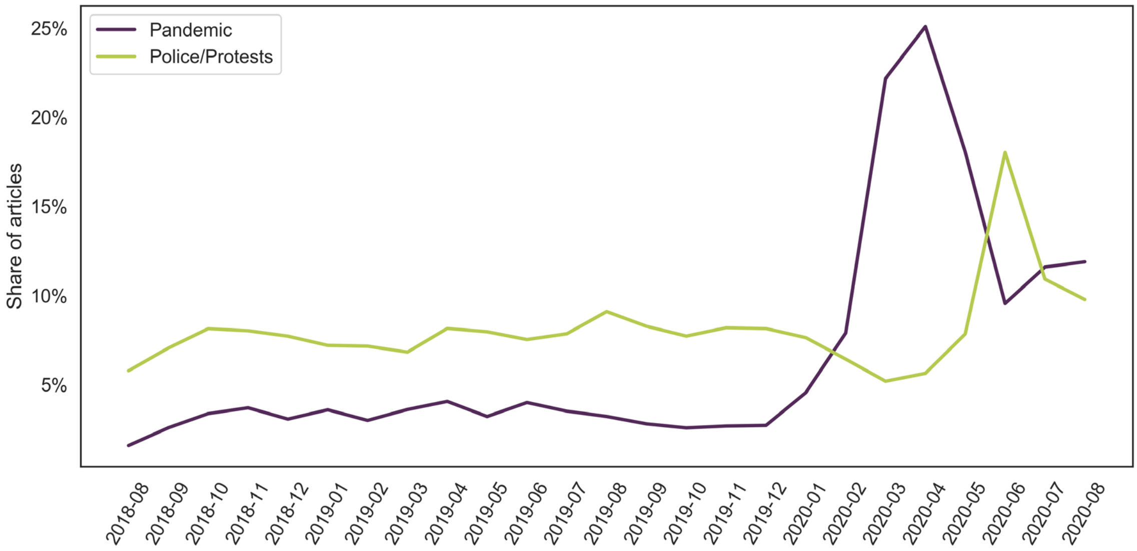
- the share of articles about the pandemic begins to increase around January 2020 and peaks in March, when Covid-19 was declared a pandemic by the World Health Organization; and
- the share of articles about police/protests peaks around June 2020, after the death of George Floyd at the end of May 2020 sparked several anti-racism protests in many countries.

This figure also illustrates that the pandemic topic captures a small fraction of articles from the start of the sample, well before the outbreak of the Covid-19 pandemic. This is because these articles use similar vocabulary to the one used by the topic we labelled pandemic. For example, we find evidence of articles discussing other viruses (such as Ebola, measles, or the common flu) prior to the Covid-19 outbreak.

[109] See, for example: K. Leetaru, 2011. *Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space*; M. Heinz, and J. Swinnen, 2015. *Media slant in economic news: A factor 20*. Economics Letters, 132, pp.18-20; G. Koren, and N. Klein, 1991. *Bias against negative studies in newspaper reports of medical research*. Jama, 266(13), pp.1824-1826; M. O'Connell, 1999. *Is Irish public opinion towards crime distorted by media bias?*. European Journal of Communication, 14(2), pp.191-212; J. Ditton, and J. Duffy, 1983. *Bias in the newspaper reporting of crime news.* J. Brit. *Criminology*, 23, p.159; N. Kalaitzandonakes, L.A Marks, and S.S.Vickner, 2004. *Media coverage of biotech foods and influence on consumer choice*. American Journal of Agricultural Economics, 86(5), pp.1238-1246.

[110] M. Trussler, and S. Soroka, 2014. *Consumer demand for cynical and negative news frames*, The International Journal of Press/Politics, 19(3), pp.36; S. Soroka, P. Fournier, and L. Nir, 2019. *Cross-national evidence of a negativity bias in psychophysiological reactions to news*. Proceedings of the National Academy of Sciences, 116(38), pp.18888-18892; J.J. McCluskey, and J.F. Swinnen, 2007. *Rational ignorance and negative news in the information market*.

Figure 5: Share of articles about 'Pandemic' and 'Police/Protests' over time



Source: Ofcom analysis.

The largest false information websites in the UK have a greater focus on European politics

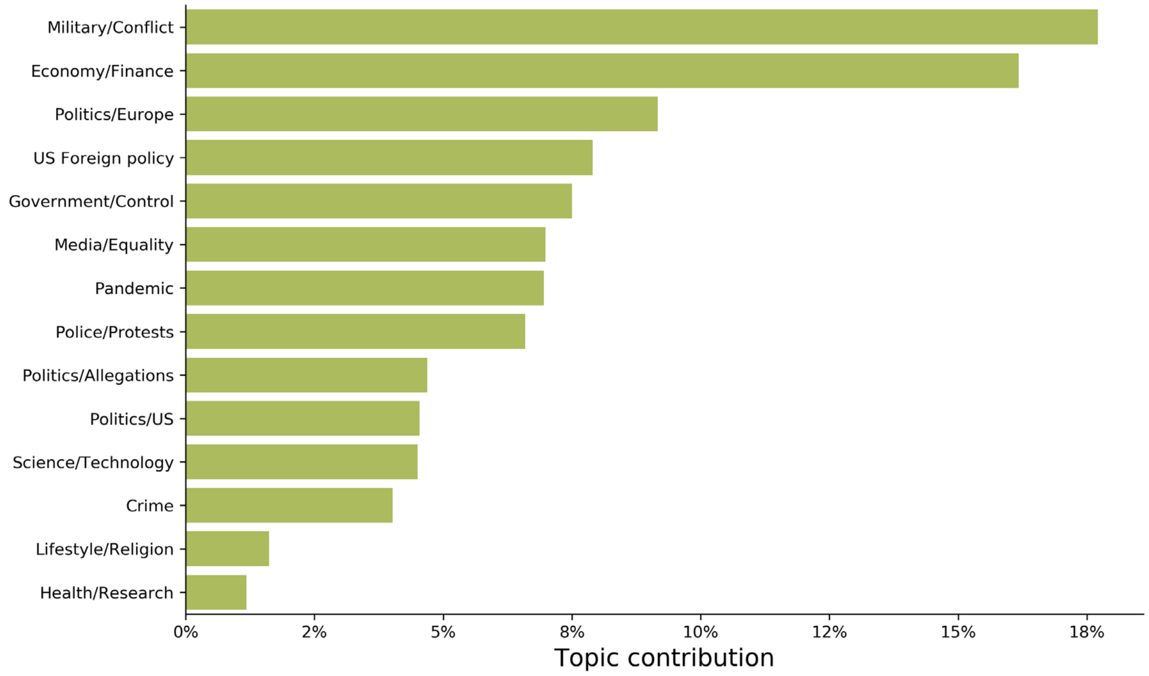
Given that the top five websites accounted for about 54% of total UK traffic to the false information websites in our sample (see Table 1), we also analyse which topics these websites publish on.

Figure 6 illustrates the topic distribution of these top five websites. To allow us to compare the distribution of topics available on these websites to those in our overall sample, the figure re-uses the topics generated from the modelling of all websites in our sample.[111]

Compared to the topic distribution from the model trained on all of the articles in our sample (see Figure 4), we notice some differences in the topic proportions of the top five false information websites in the UK. Namely, the topic 'Politics/US' makes up a smaller share of articles, whereas the topic 'Politics/Europe' accounts for a larger share.

[111] In other words, the figure is not based on a topic model run only on the articles scraped from the top 5 false information websites.

Figure 6: Topic prevalence in the sample of articles from the top five false information websites by UK traffic



Source: Ofcom analysis.

Modes of transmission and the role of online platforms

In Figure 7 we illustrate the different routes to both trustworthy information websites and false information websites for traffic that originates from desktop devices.

This shows that direct traffic is the leading source of traffic to the trustworthy information websites and false information websites we analyse.[112] Specifically, 52% of UK users' visits to false information websites in our dataset are made via direct access.[113] Similarly, 46% of traffic to trustworthy information websites in our sample originates from direct traffic. This is consistent with a CMA market study which analysed the sources of website traffic for online publishers and found that 55% of desktop traffic is through direct access.[114]

Search engines and social media represent the other two major sources of traffic to both types of websites in our sample.[115] As explained in Section 2, people increasingly access information online through an intermediary. This is reflected in Figure 7.

- Thirty-eight per cent of visits to trustworthy information websites were made via a search engine, while only 23% of visits to false information websites were made via a search engine.

[112] These are visits where users access a website by typing directly, or auto-completing, the URL in the navigation bar of their browser or by clicking a 'favourite' tab.

[113] We note that it is plausible the websites in our sample are initially found through search and/or social media but visits in the months following those initial visits may then be through direct access.

[114] Competition and Markets Authority, 2020. *Online platforms and digital advertising*, Table 5.7.

[115] Our analysis does not capture those users who just read the headline or snippet from an article – on either a search engine or social media – without actually clicking on the link. Similarly, we cannot identify those users who do not actually read and engage with the content after clicking on a link.

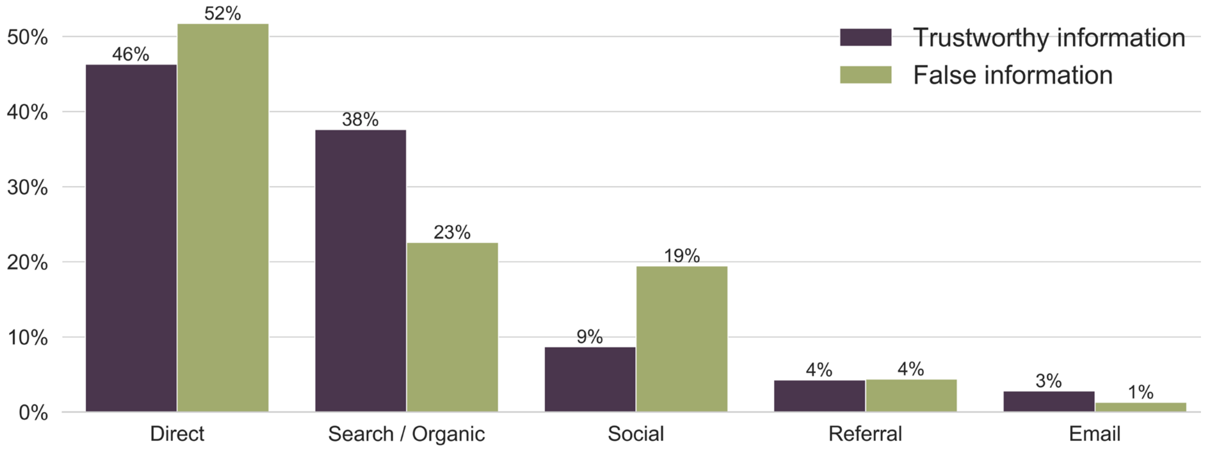
That is, in relative terms, search engines contribute more to traffic towards trustworthy information websites than they do to traffic towards false information websites.

- Nine per cent of visits to trustworthy information websites originate from users following a link on a social media platform, while 19% of visits to false information websites come from social media. That is, social media's relative contribution to total traffic towards false information websites is more than double its relative contribution to trustworthy information websites' traffic. This is consistent with other Ofcom research, which finds that social media represent an important access point for news and information.[116]

Finally, Figure 7 illustrates that access via email referrals are a very small proportion of total visits for both trustworthy information websites and false information websites in our dataset.

The false information proportions reported in Figure 7 are weighted towards the top 10 websites due to their significantly higher traffic volumes in our sample. Moreover, SimilarWeb does not record traffic for the smaller false information websites in the NewsGuard sample. It is therefore possible that the contribution of different sources of traffic split differently if we were to study a richer dataset of smaller false information websites only.

Figure 7: Sources of traffic to information websites (desktop only), September 2018 to August 2020



Source: Ofcom analysis of NewsGuard and SimilarWeb data.

Comparison of demographic characteristics of audiences

We have explored some of the characteristics of the audiences that engage with false information websites in our sample. Our analysis is restricted by the limited availability of demographic information for users in the dataset we use. Nonetheless, the data provides useful insights on the demographic of users who are more likely to engage with – and potentially be harmed by – online false information.[117]

[116] Ofcom, 2020. *News Consumption in the UK: 2020*.

[117] Note the potential for harm for a given engagement with or exposure to false information may differ across demographics.

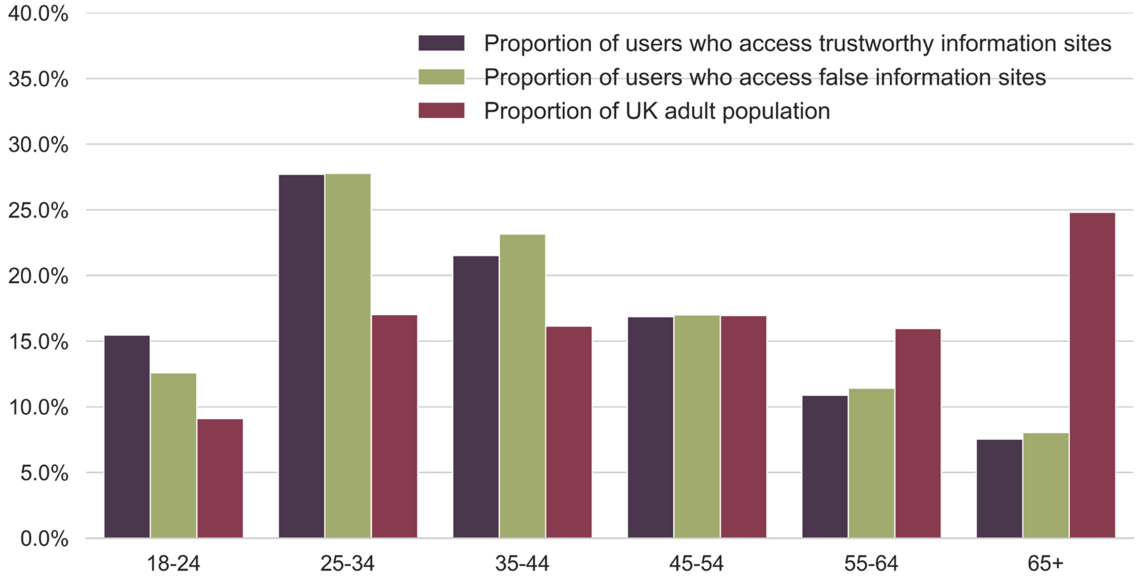
Analysis by age

Figure 8 illustrates the proportion of users that access trustworthy information websites and false information websites for different age groups, against the proportion of the adult population for each age group in the UK.[118]

The figure shows that among those UK users who access false information websites in our sample, 13% of users are 18-24 years old. Among those UK users who access the trustworthy information websites in our sample, a slightly higher share, 15%, are 18-24 years old. Conversely, some of the older age groups (e.g. 35-44) contribute, in relative terms, slightly more to traffic towards false information websites than to trustworthy information websites.

In addition, users in the 18-24, 25-34 and 35-44 age ranges account for more than a proportionate share of traffic to both types of websites, when compared to their share of the total UK population. The 25-34 olds contribute most to traffic to both the trustworthy information websites and false information websites we analyse. The oldest age groups (55-64 and 65+) account for a smaller share of traffic to both sets of websites.

Figure 8: Breakdown of traffic to information websites by age group (both desktop and mobile devices), September 2018 to August 2020



Source: Ofcom analysis of NewsGuard and SimilarWeb data. Population data from the ONS.

As our data only captures user engagement with websites, the findings above might reflect the fact that younger audiences have different preferences in how they consume information and use online services. They still may be exposed to false information, but on websites that are not in our sample, or as they consume information via other media sources online.

The same principle applies to the oldest age groups – the fact that their share of consumption of online information is smaller than their share of the total population might also reflect that they obtain information via a different medium (for example print newspapers rather than websites).

[118] ONS, 2020. *Estimates of the population for the UK, England, Wales, Scotland, and Northern Ireland*.

Consistent with this, as part of Ofcom's 2020 News Consumption Survey, we found that those aged 65+ are more likely to use offline sources such as TV, radio and printed newspapers for news.[119]

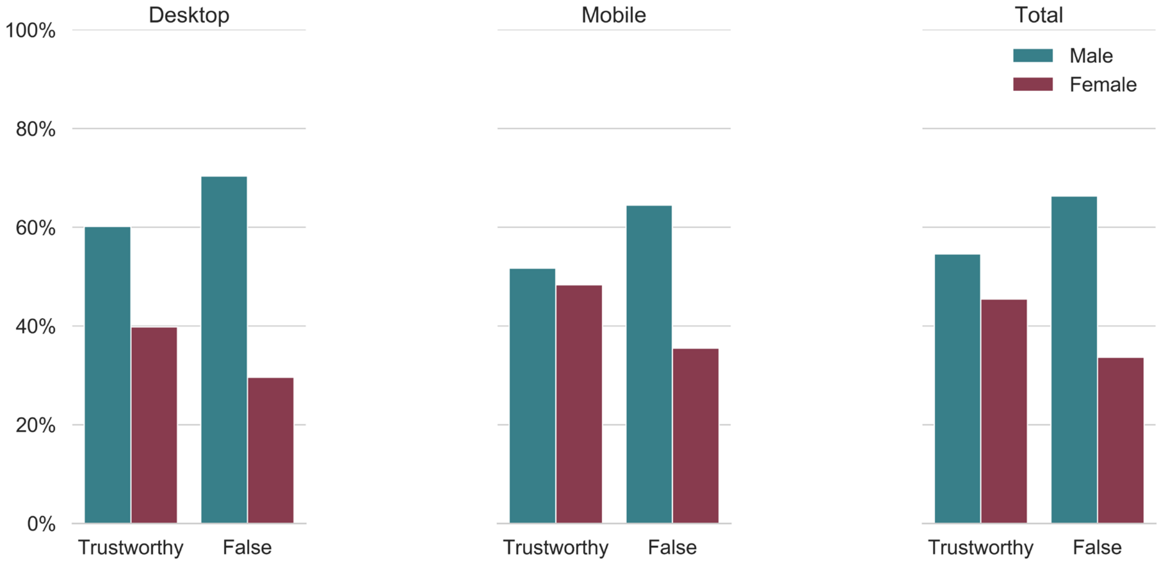
Analysis by gender

Figure 9 illustrates the proportion of visits accounted for by each gender for both trustworthy and false information websites.

This shows that males account for a greater proportion of users than females across all types of information websites in our sample. This is consistent with findings from the ONS Internet Access Survey which show that men are more likely to consume news online than women.[120]

In addition, Figure 9 shows that the gap is larger for false information websites. In the case of false information websites in our sample, the share of male users is about 70%. This is significantly higher than for the trustworthy information websites we analyse, where the proportion of male users is about 55%.

Figure 9: Breakdown of traffic to information websites by gender (by device), September 2018 to August 2020



Source: Ofcom analysis of NewsGuard and SimilarWeb data.

From our topic modelling analysis, we also found some indication of differences by gender in the types of websites visited. Specifically, websites that are mostly visited by female audiences are more likely to cover topics related to health. This is shown in Figure 10, where we plot the topic distribution of websites whose audiences are made up of more female users than male users. More detail on our demographic analysis is provided in Annex 2.[121] The ONS Internet Access Survey also

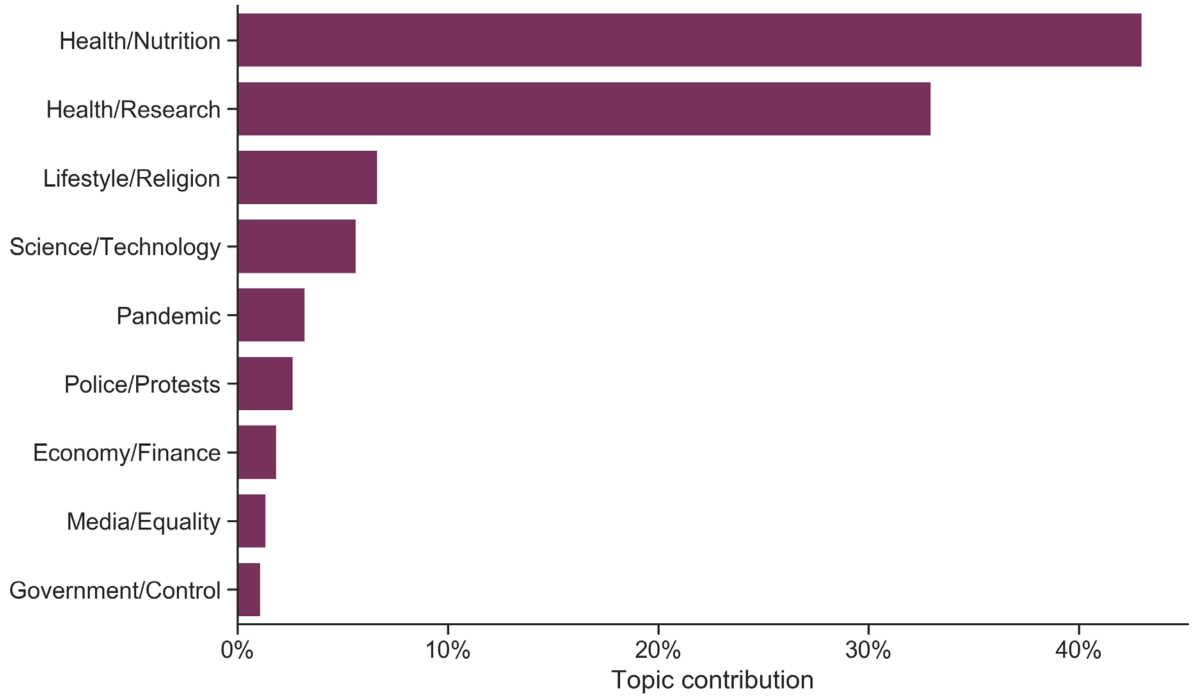
[119] Ofcom, 2020. News Consumption in the UK: 2020, p7.

[120] ONS, 2020. Internet Access Survey, Table 6.

[121] As we do not have demographic information at the article level. In order to perform this analysis, we identify each website by its main demographic. We then assume that all the articles within that website are read by an audience that matches the website's main demographic. Then, we group the articles by demographic and, using the topic model trained on all the articles in our sample, we extract the topic distributions for each demographic.

points to this, as it found that women use the internet to look for health-related information more than men.

Figure 10: Topic distribution of websites with predominantly female audiences



Source: Ofcom analysis.

Concluding remarks

What this paper contributes

This paper contributes to ongoing research on how issues relating to false information online can be analysed using large datasets and new quantitative methodologies. Our analysis has provided a range of insights on how we could assess the scale and potential harm arising from false information.

Specifically, we shed some light on how much traffic some of the largest false information websites, according to NewsGuard, receive from the UK. While overall traffic to the false information websites in our sample is small relative to all online navigation, it is not insignificant. We also find that traffic to such websites is concentrated among a few large players, with a long tail of smaller websites attracting limited traffic.

We also built a large dataset of articles from English language false information websites over time and used machine learning techniques to produce insights on the topics discussed, their frequencies, and their evolution over time. This analysis suggests that the topics covered on these websites (such as health, politics, and conflict) have the potential to generate harm, if reported on falsely.

Our analysis provides data on the routes by which users access false information websites. We find that social media and search are key contributors of traffic alongside direct navigation. We also assess the demographics of visitors to these websites and find that female audiences tend to engage more with websites which report on topics related to health.

Avenues for further research

This paper explores data sources and research techniques that can be used to investigate false information online. We recognise that there are important and wider questions that our research does not address, different approaches that may be taken, as well as potential limitations to our analysis. Together these suggest possible avenues for further research.

First, richer datasets may allow for more refined insights and overcome some of the challenges we have encountered. For example, our topic modelling analyses articles that are available on false information websites. However, as NewsGuard classifies websites rather than individual articles, we could not restrict our model to individual articles which had been identified as false information. An alternative approach could be to use data from fact checkers that debunk individual articles such as Snopes and Full Fact.

Another example is the difficulties we encountered in building a dataset which could be used for our analysis.

- We have used SimilarWeb to obtain UK traffic and demographic data for the websites in our sample. However, it has proven challenging to obtain data for websites that receive smaller volumes of traffic.
- We have relied on web scraping and the Internet Archive to collect articles for our topic modelling of false information websites. This means that our sample of articles excludes websites that ban bots or when the Internet Archive does not hold data for them, which can occur especially for smaller, more niche websites.

This means that our findings may not reflect websites with smaller volumes of traffic. This can be particularly relevant for the analysis of false information websites, as traffic data was only available for 177 websites out of 1,083 as categories by NewsGuard. Datasets which capture these smaller players could therefore complement our approach.

Second, as illustrated in Figure 1, the scope of our analysis was limited to textual data and information websites, and therefore our research approach captures a subset of the supply of false information. While we analyse occasions when UK users may be exposed to false information as they navigate directly to websites, they might also come across such information in other formats (such as video) and in other locations (social media, podcasts, video sharing platforms, or news aggregators). Studies focusing on other areas of the false information ecosystem could therefore complement our research.[122]

Third, we acknowledge that there is no single, uncontested view of what constitutes false information. While we have used NewsGuard's data to identify false information websites, a useful exercise would be to repeat our approach using data from other independent rating organisations.

[122] For example, *Measuring the reach of "fake news" and online disinformation in Europe* by the Reuters institute (2018) looks at social media reach false information websites focusing on France and Italy.

Fourth, the dataset that we have built for our topic modelling analysis does not contain information on the reach of individual articles. As mentioned above, our research can therefore be characterised as assessing the 'supply side' of information online, which could be complemented with other research capturing the perspective of the reader. Such 'demand side' research could explore which articles are read, whether users can distinguish false information from truth, and how false information may impact their views and actions. One such example is the survey Ofcom commissioned as part of our Making Sense of Media Programme to investigate people's access, consumption, and engagement with news during the Covid-19 outbreak.[123]

Finally, performing a topic modelling analysis for trustworthy information websites may also be a useful complement to this analysis. This could serve as a comparator for the topic modelling run on false information websites, which would further isolate the characteristics that are unique to false information.

[123] Ofcom, 2020. *Covid-19 news and information: consumption and attitudes*.

A1. Network traffic analysis

This Annex provides further detail on the methodology and sources of data that we use for our web traffic analysis described in Section 3.

NewsGuard

In order to analyse false information at scale, we use data from the news rating organisation NewsGuard to identify false information websites.

NewsGuard is an organisation that "employs a team of trained journalists and experienced editors to review and rate news and information websites" to help the public to identify credible sources of information. To do so, it reviews information websites based on nine criteria that capture standards of credibility and transparency.[124]

Based on these criteria, each website receives a score out 100 as shown in Table A1.1.[125] Websites with a score of at least 60 points are 'green-rated' and considered credible sources of information. Websites whose score is below 60 are 'red-rated' and not considered credible sources of information. As part of NewsGuard's verification process, websites which are found not meeting the criteria needed for a 'green' rating are contacted and given a chance to comment.

Table A1.1: NewsGuard's rating criteria

Criteria	Description	Points
Credibility	It does not repeatedly publish false content, without providing prompt corrections	22
	The content providers report and present information in a fair and accurate way, referencing multiple credible or first-hand sources, and without distorting and misrepresenting information	18
	Errors are regularly corrected or clarified	12.5
	There is a clear distinction between factual and opinion statements	12.5
	The headlines used are not deceptive	10
Transparency	Discloses ownership and financing information, in a way that is accessible to the user	7.5
	Labels advertisements and paid content clearly	7.5
	Makes information about possible conflict of interests available	5

[124] NewsGuard provides ratings for more than 4,000 websites which they claim account for 95% of engagement with news websites in the UK.

[125] Platforms or websites that deal with satire or humour, or those that mainly consist of user-generated content, are labelled differently. See NewsGuard, *Rating process and criteria*.

	Provides names, with contact or biographical information, of the content creators	5

Source: NewsGuard.

We have used this rating system to classify websites as follows:

- we classify a website as a 'false information website' if it does not comply with the first credibility requirement ("Does not repeatedly publish false content"), regardless of their overall NewsGuard score;[126] and
- we classify a website as a 'trustworthy information website' if it receives a green rating. This is a more restrictive criterion as we want this category to only include websites that NewsGuard finds credible and transparent.

We thus exclude from the analysis all those websites that fail to meet NewsGuard's credibility and transparency requirements, for reasons other than posting false information. This is to isolate the false information issue from criteria related to other journalistic standards.

SimilarWeb

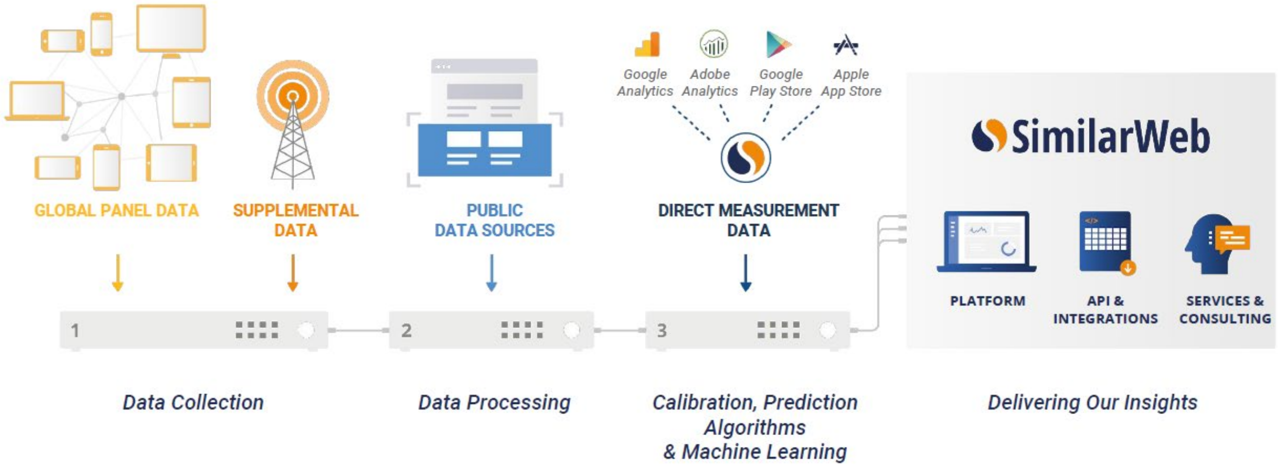
SimilarWeb is a provider of website-level audience and traffic statistics, which are used for wide-ranging purposes that span from competitive analysis to market research. [127]

SimilarWeb's statistics are constructed as shown in Figure A1.1:

- in the data collection stage, SimilarWeb gathers data from a "diverse panel of users" who allow SimilarWeb to monitor their internet activity anonymously; [128]
- after the data from the panel is processed and cleaned, it is supplemented with public data for mapping and categorisation purposes;
- SimilarWeb then calibrates this data using information from hundreds of thousands of directly measured websites and apps (i.e. direct measurement data); and
- finally, to improve the accuracy of their estimations, the data is run through prediction and machine learning algorithms to obtain their final estimates. This methodology is applied equally to all the websites available on the platform.

[126] There are no green rated websites that are classed as "repeatedly publishing false content" as of September 2020.
[127] See similarweb.com
[128] SimilarWeb ensures that its panels are representative of the population of each country they have a presence in.

Figure A1.1: SimilarWeb's data pipeline



Source: SimilarWeb.

There are a variety of market intelligence tools like SimilarWeb, each using a unique methodology and approach to collect, process, and estimate website audience and traffic data. For example, unlike other providers of market intelligence tools, SimilarWeb does not rely on cookies for counting unique visits, as they are susceptible to being manually or automatically deleted.

Different services may also have different definitions for the metrics they report, meaning that they can report different statistics on user traffic. In the case of SimilarWeb, a visit is defined as an individual period of time that users spend on a website. A visit ends either after 30 minutes of inactivity or if the user leaves the website for more than 30 minutes.

In comparison to the other market intelligence tools for which we have seen similar data, SimilarWeb appears to offer a better coverage of smaller websites in terms of both mobile and desktop traffic. This is crucial for our analysis given that the bulk of false information websites in our sample feature limited traffic. We have therefore chosen to use SimilarWeb for our analysis.

Nonetheless, SimilarWeb only provides UK traffic data for about 20% of false information websites and about 70% of the trustworthy information websites in the NewsGuard sample, which may be due to the websites not receiving sufficient UK traffic or because they are not tracked by SimilarWeb. While this is still a reasonable sample, it is not as broad as the entire NewsGuard dataset.

For the analysis in Section 3, we use SimilarWeb's statistics that cover traffic, sources of traffic and audience demographics. The statistics we report are all monthly series referring to the period from September 2018 to August 2020, and they have been constructed as follows:

- the traffic data we report is aggregated by date and type of information website;
- the data on sources of traffic and audience demographics are a weighted average for each indicator, using websites' total traffic as weights. This is to recognise that websites which receive more traffic should carry more influence in relation to these metrics.

A2. Topic modelling

This Annex describes our approach to identifying the topics of false information websites. As mentioned in Section 3, we have used a topic modelling algorithm called Latent Dirichlet Allocation (LDA) on a sample of 406,267 articles from 97 websites that "repeatedly publish false content" according to NewsGuard from August 2018 to August 2020.

In this Annex, we first describe our approach to data collection, cleaning, and pre-processing. We then discuss the modelling methodology and present some additional visualisations to the results presented in Section 3. These should be interpreted in the context of the limitations set out in Section 3.

Data

Data collection

Our starting point is the list of 1,083 websites that NewsGuard identifies as "repeatedly publish[ing] false content" as of September 2020.

We narrow down this list to the 177 websites for which SimilarWeb holds UK traffic data. The purpose of this is twofold: it allows us to focus on the websites which UK audiences are more likely to have visited, and it improves the processing times of our analysis. From this list, we remove domains that are associated with countries where English is not a first language based on the URLs' top-level domains (e.g. '.it' or '.es') and websites that do not contain news articles (e.g. video-based sites). This reduces the list from 177 to 144 websites.

After identifying the target websites for our analysis, we use the Internet Archive's 'Wayback Machine API' to request the URLs to the homepage of each of the 144 websites in our list for every calendar day between August 2018 to August 2020.[129] The API returns the link to the websites' archived snapshot that is closest to the date requested, as well as the date of the returned snapshot. Websites can request not to have their data archived by the Internet Archive, or the Internet Archive may not have archived certain websites. We have therefore excluded a further 12 websites from our target list that were not available through the Wayback Machine API.

We then develop a web scraping framework that sends a request to all the links of the websites' homepages as returned by the Wayback Machine API. Websites can opt out of being scraped via their 'robots.txt' files. Our code respects the permissions set in this file and is designed not to overload servers (e.g. by throttling requests). We also do not scrape any content that is behind a paywall.

The code then follows all the hyperlinks on the homepages and extracts the following items from every hyperlink:

- the title;
- the main body; and

[129] Wayback Machine, *Wayback Machine APIs*.

- the website’s domain name.

Even though we have filtered out websites with top-level domains associated with countries where English is not a first language, some foreign websites have ‘.com‘ or ‘.net‘ top-level domains. We therefore run a language detection tool on the body of the articles in our sample to filter out foreign language content. See Table A2.1 below for the list of the languages detected.

Table A2.1: Articles by language

Languages	Articles
English	551,794
French	10,334
German	5,609
Italian	445
Other	39

Source: Ofcom analysis.

Finally, our script incorrectly extracted some content that was structured in the same way as news articles but was a different type of content entirely (such as a website’s privacy policy page or their terms and conditions page). We have used regular expressions to filter the URLs of the articles and remove these from our sample.

We also chose to remove:

- articles without a main body;
- duplicate articles on a per-website basis, as the same article could have been in multiple pages on the same website; and
- articles with a body that is less than 500 characters long, since, based on observation, these are unlikely to be news articles.

These data cleaning tasks further reduce our sample of false information websites from 132 to 97 domains. Our final sample therefore contains 406,267 articles from 97 domains from August 2018 to August 2020.

In Table A2.2 below we provide the number of articles in our sample for every year we extracted data for. We use the date returned by the Wayback Machine API as a proxy of the articles’ publication date, since extracting the publication date via our scraping code was more prone to errors and missing values.

Table A2.2: Articles by year scraped

Year	Articles
August 2018 – December 2019	58,806

January 2019 – December 2019	208,549
January 2020 – August 2020	138,912

Source: Ofcom analysis.

We also find that we have more articles for the websites that receive more traffic. This is because these websites are more likely to have been archived by the Internet Archive. Owing to our web traffic analysis we know that traffic to false information websites is fairly concentrated around a few large players, so this distribution is to be expected.

Pre-processing

Before a topic model can be estimated, the collection of documents (also called 'corpus') needs to be cleaned of unwanted elements and pre-processed further. The final objective of this stage is to create the document-term matrix, a matrix that lists the frequency for each word in each article. Pre-processing is an important stage in topic modelling, as it has been shown to have a strong impact on the results.[130]

Most LDA studies report using a range of similar pre-processing steps, which we follow in our analysis. These are:

- forming bigrams – bigrams are words that frequently occur together and should therefore be considered a single word rather than two separate words (e.g. 'United States');
- tokenisation – this consists of transforming each article into a list of words;
- transforming all characters to lower case;
- removing punctuation and special characters such as 'URLs', hashtags, and links;
- removing stop words – these are words that frequently occur in the English vocabulary such as prepositions or articles. They are not sufficiently specific to illustrate document content, so would only increase the processing power and time needed to estimate models while adding nothing in terms of uncovering actual topics;[131]
- lemmatisation – lemmatisation serves to reduce inflected words to their root form (e.g. from writing to write);[132] and

[130] M. J. Denny, and A. Spirling, 2017. *Text pre-processing for unsupervised learning: Why it matters, when it misleads, and what to do about it*. New York University.

[131] We have supplemented the standard dictionary of stop words with a corpus-specific dictionary based on inspecting the most frequent words in our collection of documents.

[132] We prefer lemmatisation over stemming as the results of lemmatisation tend to be more reliable. This is because while stemming simply cuts-off the end of each word, lemmatisation takes into account the part of speech that each word plays in a sentence and ensures that the root word produced is part of the language.

- filtering extremes – these are terms that are either too frequent or too infrequent. We consider these words to be either too specific or too general to be used to infer the topic structure of the documents.[133]

Once pre-processing is finished, we apply the methodology described below on a corpus of 406,267 documents and 60,623 unique words.

Methodology

We have performed our analysis of topics reported on by false information websites by applying a topic modelling algorithm called Latent Dirichlet Allocation (LDA) on our sample of articles. We describe each of these techniques in more detail below.

Topic modelling

Topic modelling algorithms can be thought of as automated tools for summarising large volumes of text. More specifically, they are statistical methods that analyse the words contained in a collection of documents to discover the hidden thematic structures that run through them.[134] Topic models can help analyse large quantities of text without either knowing what might exist in the text beforehand, or any manual or qualitative methods such as manual annotation.

These techniques have been widely implemented in social sciences to identify the concepts in large collections of documents in an automated manner. For example, topic models have been applied on books to discover literary themes, on scientific journals to discover salient scientific topics, on newspapers to analyse news coverage on a specific issue, and on political speeches to analyse how styles of political communication relate to polarisation.[135, 136, 137, 138]

Topic models also have been applied to online content such as online newspapers, Tweets (i.e. public messages on Twitter.com) and blog posts. For example, Ghosh and Guha (2013), apply topic modelling to Tweets to identify topics about obesity, while Koltsova and Koltcov (2013) apply topic modelling to investigate the political agenda of LiveJournal, Russia's leading blog platform. [139,140]

[133] We set this as words that appear in less than 30 articles or in more than 40% of the articles in the overall collection of articles. We choose this threshold based on the approach taken by the literature we quote in this paper.
[134] D.M. Blei, 2012. *Probabilistic topic models*. Communications of the ACM, 55(4), pp.77-84.
[135] M.L. Jockers, and D. Mimno, 2013. *Significant themes in 19th-century literature.* Poetics, 41(6), pp.750-769.
[136] T.L. Griffiths, and M. Steyvers, 2004. *Finding scientific topics.* Proceedings of the National academy of Sciences, 101(suppl 1), pp.5228-5235.
[137] P. Di Maggio, M. Nag, and D. Blei, 2013. *Exploiting affinities between topic modelling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding*. Poetics, 41(6), pp.570-606.
[138] J. Grimmer, 2013. *Appropriators not position takers: The distorting effects of electoral incentives on congressional representation*. American Journal of Political Science, 57(3), pp.624-642
[139] D. Ghosh, and R. Guha, 2013. *What are we 'tweeting' about obesity? Mapping tweets with topic modelling and Geographic Information System*. Cartography and geographic information science, 40(2), pp.90-102.
[140] O. Koltsova, and S. Koltcov, 2013. *Mapping the public agenda with topic modelling: The case of the Russian LiveJournal*. Policy & Internet, 5(2), pp.207-227.

Latent Dirichlet Allocation

While there are various topic modelling algorithms, we chose to implement Latent Dirichlet Allocation (LDA), as it is the "simplest", "most widely used", "state-of-the-art" method for topic modelling. [141,142]

LDA is a generative probabilistic model built on two key assumptions:

- documents are probability distributions over a fixed number of topics; and
- topics are probability distributions over a fixed vocabulary.[143]

When the LDA algorithm is run on a collection of documents, the model will estimate the topics (which are probability distributions over words) and their proportions in every single document in the corpus.

Using the example from Blei's seminal paper, an article about "the use of data analysis to determine the number of genes an organism needs to survive" will exhibit, with some proportion, words about data analysis (e.g. 'data', 'computer' etc.) and words about genetics (e.g. 'gene', 'DNA' etc.). [144] The model will group these words together and the analyst can then apply labels to indicate that one topic is about data analysis while the other is about genetics.[145] Another key assumption of LDA is that documents are mixtures of topics, so, the model will also output the topic proportions for every single document. For the article in the example these could be 75% 'genetics', and 25% 'data analysis'.

Model selection

Selecting the number of topics

Before running the LDA algorithm, there are certain parameters that need to be set. In particular, the number of topics that the model should classify the words into needs to be fixed *a priori*.

Our approach to model selection pointed to the model with 15 topics being the most appropriate one for the purpose of our analysis.

As the relevant literature points out, there is no "right" number of topics: the estimated model simply needs to be useful for the purpose of the analysis.[146] This is because the objective of these

[141] D.M. Blei, A.Y. Ng, and M.I. Jordan, 2003. *Latent Dirichlet allocation.* Journal of machine Learning research, 3(Jan), pp.993-1022.

[142] C.B. Asmussen, and C. Møller, 2019. *Smart literature review: a practical topic modelling approach to exploratory literature review.* Journal of Big Data, 6(1), p.93.

[143] The distribution that is used to draw the per-document topic distributions and to allocate the words to the drawn topics is the Dirichlet distribution. The Dirichlet distribution allows for the multiplicity of topics in the documents and for the multiplicity of words in the topics. The Dirichlet distribution is a multivariate probability distribution that describes $k \geq 2$ variables $X_i, \dots, X_k$ such that each $x_i \in (0,1)$ and $\sum_{i=1}^{K} x_i = 1$. It is parametrised by a vector of positive-valued parameters $\alpha = (\alpha_1, \dots, \alpha_k)$.

[144] D.M. Blei, 2012. *Probabilistic topic models*. Communications of the ACM, 55(4), pp.77-84.

[145] The topics in LDA are 'latent' or hidden because they are visible only in the form of groups of words from the corpus. Therefore, when we talk about a topic being 'about' a subject, we are referring to those distributions over the vocabulary which place high probability on words that are related to that subject.

[146] D. Elgesem, L. Steskal, and N. Diakopoulos, 2015. *Structure and content of the discourse on climate change in the blogosphere: The big picture.* Environmental Communication, 9(2), pp.169-188.

type of models is not to correctly estimate population parameters, but to identify the right lens through which one can see the data most clearly.[147]

Generally, a lower number of topics is used to get a general overview of the corpus, while a higher number of topics is used for a more detailed analysis.[148] This is because changing the number of topics affects the granularity (level of detail) of the model. Specifically, a higher number of topics translates into a higher granularity. With a higher number of topics, each topic will represent more specific content characteristics.[149]

In general, models with a lower number of topics are overly broad and combine words from different subjects into a single topic. Conversely, models with higher number of topics offer additional identifiable topics, but these tend to be too specific.[150]

In order to select the optimal number of topics, it is conventional to estimate a set of candidate topic models varying the number of topics ('k') and using various quantitative and qualitative validation techniques to select the model that best fits the purpose of the analysis.

One common quantitative measure that is used to select the number of topics in LDA is 'perplexity'.[151] Perplexity gives an indication of the number of topics that best predicts the data, with a lower perplexity score indicating a better prediction.

One limitation of the perplexity measure is that it does not consider the interpretability of the topics. Chang et al. (2009) showed that statistical measures of topic coherence (such as perplexity) and human judgment were often not correlated.[152] This spurred the development of semantic, rather than statistical, measures of topic coherence to proxy human judgment. One such a metric of topic coherence measures the extent to which the top *N* words in each topic co-occur in the topics estimated by the model. The idea is that the higher this measure, the easier it should be for a human to interpret the topics.[153]

Given this, we use a 'cross validation' approach and train an LDA model on a portion of the data (the training set), evaluating the perplexity of the model using held-out data (the test set).[154] We also calculate the topic coherence measure described above on the training set of the documents. We repeat this procedure varying the numbers of topics, starting at $k = 5$ and finishing at $k = 80$, proceeding in steps of five.

[147] P. Di Maggio, M.Nag, and D. Blei, 2013. *Exploiting affinities between topic modelling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding*. Poetics, 41(6), pp.570-606.

[148] Ibid. footnote 147.

[149] C. Jacobi, W. Van Atteveldt, and K. Welbers, 2016. *Quantitative analysis of large amounts of journalistic texts using topic modelling.* Digital Journalism, 4(1), pp.89-106.

[150] M.S. Evans, 2014. A computational approach to qualitative analysis in large textual datasets. PloS one, 9(2), p.e87908.

[151] D.M. Blei, and J.D. Lafferty, 2007. *A correlated topic model of science.* The Annals of Applied Statistics, 1(1), pp.17-35.

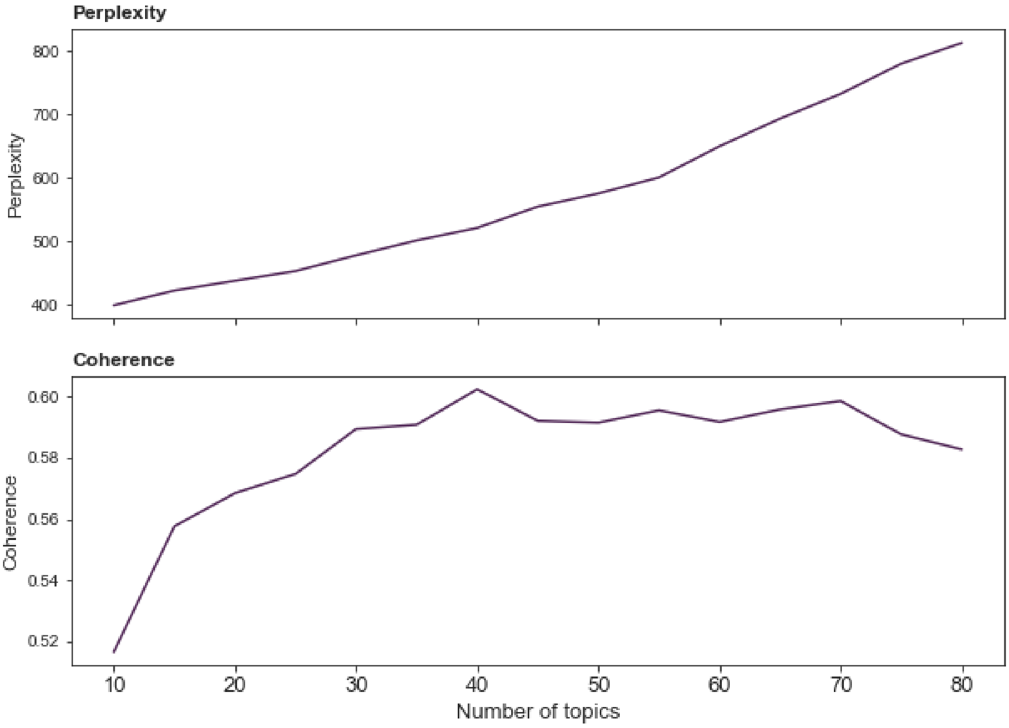
[152] J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, and D. M. Blei. *Reading tea leaves: How humans interpret topic models.* Advances in neural information processing systems, pp. 288-296. 2009.

[153] D. Mimno, H.M. Wallach, E. Talley, M. Leenders, and A. McCallum, 2011. *Optimizing semantic coherence in topic models.* In Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp. 262-272.

[154] In our analysis we also implement a k-fold cross validation procedure in which we randomly split the dataset into five groups (folds), use the first fold as a validation set and fit the model into the remaining four folds. We then aggregate the model evaluation metrics tested (perplexity/coherence) by doing a simple average over the 5 folds. We implement this method in order to make sure that the results from the analysis are reliable and not driven by how the training/test split is performed.

Figure A2.1 presents the perplexity and coherence for each model for increasing values of k. We notice that perplexity is minimised in the interval of $5 \leq k \leq 20$ (specifically at $k = 10$) while coherence is maximised where $35 \leq k \leq 45$ (specifically at $k = 40$). We therefore manually inspect the topics in these intervals, and arrive at two candidate models, namely $k = 15$ and $k = 39$, based on whether the topics were easily labelled and coherent.

Figure A2.1: Perplexity and coherence for $k = 5$ to $k = 80$



Source: Ofcom analysis.

In order to choose between the two candidate models, we compared the topics (word distributions) of both models. In inspecting the word distributions, we checked whether we were able to easily summarise each topic with a brief label (such as ‘politics’ or ‘health’), flagging as ‘miscellaneous’ or ‘junk’ topics those word distributions that did not seem to make sense. The idea is that in a good topic model the word distributions should be easily labelled, because that means that the topics are interpretable by a human.[155]

We noticed that the model with a larger number of topics was more granular but contained some miscellaneous topics. For example, for $k = 39$ there were a few topics that were coherent but not very informative, such as a topic with words such as ‘flight’, ‘travel’, ‘plane’ and ‘local’. Similarly, some topics were too specific to be useful, such as terms originating from news coverage on an individual public figure.

[155] D. Mimno, H.M. Wallach, E. Talley, M. Leenders, and A. McCallum, 2011. *Optimizing semantic coherence in topic models.* Proceedings of the conference on empirical methods in natural language processing, pp. 262-272. Association for Computational Linguistics.

While the model with 39 topics did offer some additional insight, given that the aim of our analysis is to get a general overview of the topics covered by articles published on false information websites, our judgement is that the model with 15 topics is most appropriate.

Selecting alpha and beta

The other two key parameters that need to be set in a topic model are 'alpha' and 'beta'. Alpha is the parameter controlling the concentration of the 'document-topic' distribution, while beta is the parameter controlling the concentration of the 'topic-word' distribution.

Literature suggests that alpha is of greater importance than beta and that an asymmetric alpha tends to produce better results. We therefore follow this approach and set an asymmetric alpha, leaving beta as the default value of the topic modelling library, which consists of a symmetric beta.

The intuition behind the choice of an asymmetric alpha is that we expect certain documents to cover more topics and others to cover less. For beta, we expect that certain groups of words will occur more frequently than others in every document in a given corpus, which is why we leave beta as symmetric.[156]

Labelling process

Following model selection, we have manually labelled the 15 topics in our model based on a visual inspection of their topic-word distributions and a scan of the titles of around 100 articles per topic. This inspection also allowed us to check whether our model's topic assignment matched the articles' contents.[157]

The proposed labels and word distributions for each topic identified by our model are shown in Table A2.3 and Table A2.4 below.

[156] H.M. Wallach, D.M. Mimno, and A. McCallum, 2009. *Rethinking LDA: Why priors matter*. Advances in neural information processing systems pp. 1973-1981.
[157] This is a standard procedure in these types of studies, see footnote 146.

Table A2.3: Topic labels and word distribution (Topics 1 to 8)

Topic 1: US Foreign Policy	Topic 2: Pandemic	Topic 3: Health/Research	Topic 4: Health/Nutrition	Topic 5: Science/Technology	Topic 6: Politics/US	Topic 7: Economy/Finance	Topic 8: Lifestyle/Religion
China	Coronavirus	Study	Food	Water	Trump	Company	Life
India	Covid	Vaccine	Help	Earth	President	Market	Child
Trade	Health	Health	Health	Space	Democrat	Money	School
Trump	Virus	Research	Diet	Energy	Abortion	Economy	Church
Hong Kong	Case	Child	Product	Scientist	Republican	Price	Feel
International	Death	Patient	Skin	Climate	Biden	Business	Love
Tariff	Government	Cancer	Water	Planet	Election	Bank	Family
Sanction	Pandemic	Disease	Healthy	Technology	Bill	Increase	Woman
Deal	Test	Treatment	Body	Climate change	Support	Rate	Christian
North Korea	Hospital	Risk	Plant	Research	American	Financial	Parent

Source: Ofcom analysis.

Table A2.4: Topic labels and word distribution (Topics 9 to 15)

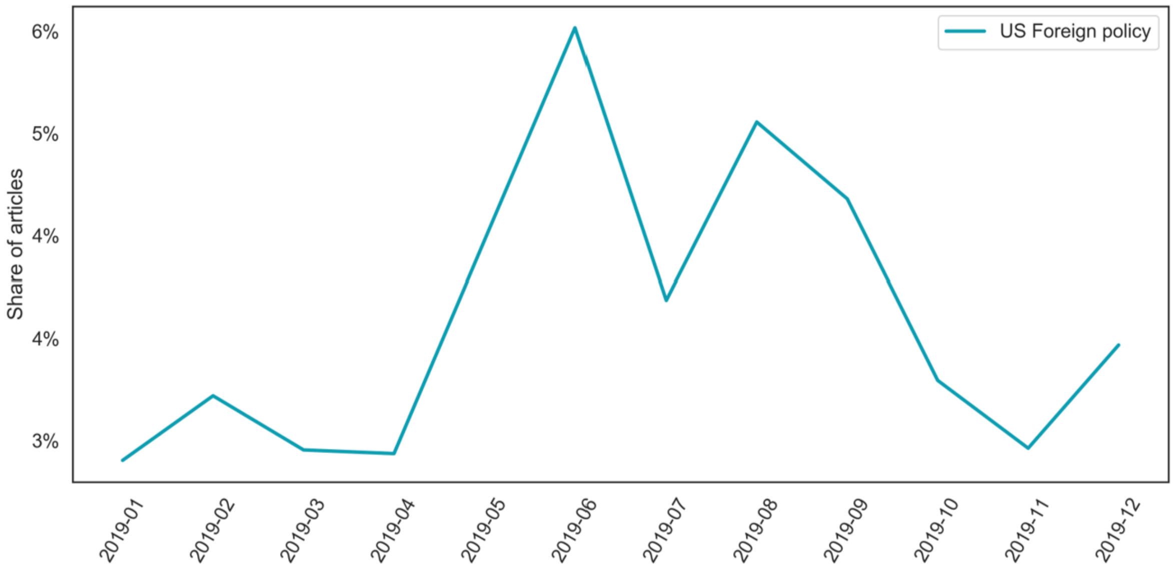
Topic 9: Politics/Europe	Topic 10: Media/Equality	Topic 11: Crime	Topic 12: Government/Control	Topic 13: Politics/Allegations	Topic 14: Military/Conflict	Topic 15: Police/Protests
Russia	Media	Case	Government	Trump	Iran	Police
United Kingdom	Black	Court	Power	Investigation	Military	Officer
British	News	Child	Nation	President	Israel	Protest
Europe	Facebook	Charge	Control	Information	Force	Kill
Putin	Tweet	Claim	Right	Evidence	Attack	Shoot
France	White	Arrest	Public	Campaign	Syria	Family
Germany	Woman	Federal	Freedom	Intelligence	Palestinian	Arrest
Brexit	Conservative	Judge	Force	Email	Turkey	Local
Parliament	Racist	Epstein	Order	Russia	Iraq	Attack
Prime minister	Hate	Woman	Democracy	Claim	Terrorist	Incident

Source: Ofcom analysis.

In order to understand why we arrived at the topic labels described above, it is important to understand the topics' contents. Therefore, we provide a brief overview of each topic and give a high-level explanation of what the articles we reviewed are about.

- **US Foreign Policy.** This topic is centred around the foreign policy of the United States. Specifically, many articles about this topic cover the relationship between the US and Asian countries such as China, North Korea, and India. A common subject that was covered by these articles is international trade. Several articles about this topic cover China-United States trade disputes. For example, there are several articles about the Huawei ban by the United States, and we see a spike in the frequency of articles about the US Foreign Policy topic around May 2019, when President Trump signed the executive order banning Huawei from buying vital U.S. technology (see Figure A2.2).[158] North Korea-United States relations are also widely covered by this topic, especially in the context of nuclear threats. Trade relations between the United States and India are also widely discussed.

Figure A2.2: Share of articles about the topic 'US Foreign Policy' in 2019



Source: Ofcom analysis.

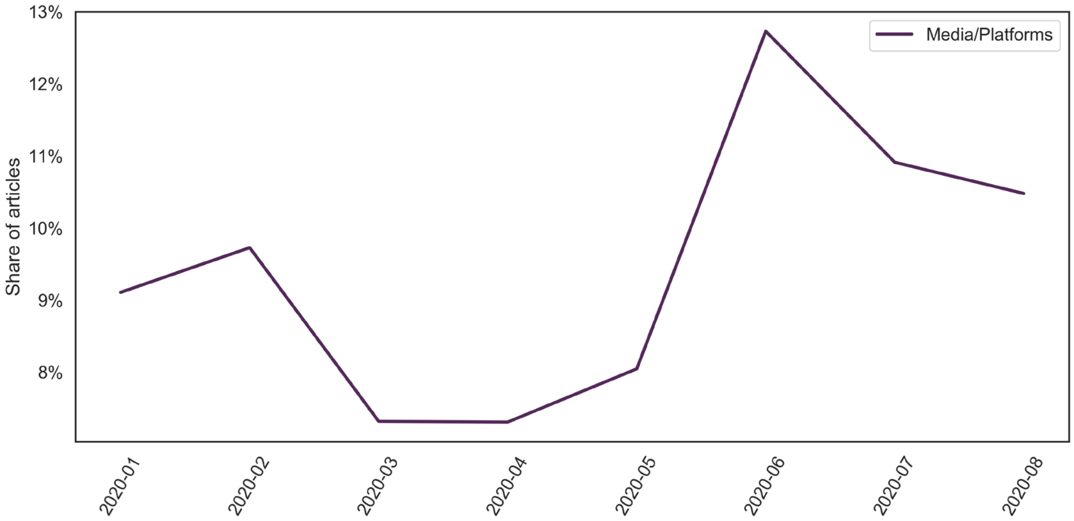
- **Pandemic.** Articles about this topic mostly cover the developments of the COVID-19 pandemic. In articles that pre-date the COVID-19 pandemic we see other viruses being discussed, such as Ebola, measles, and influenza.
- **Health/Research.** Articles about this topic tend to present information about various factors which are claimed to constitute risks to health and purport to offer advice on how to mitigate these risks. For example, articles discuss things that can lead to increased risk of diseases such as cancer, heart disease, diabetes, dementia, osteoporosis, and many others.

[158] Reuters, 2019. *Trump administration hits China's Huawei with one-two punch*. This source is not part of our dataset. We are simply quoting it to reference the event we describe in the text.

Many articles about this topic also cover drugs, vitamins, and the effect of treatments to address the above health concerns.

- **Health/Nutrition.** This topic captures articles about nutrition. Diets, recipes, and health benefits of ingredients and foods are covered by articles on this topic.
- **Science/Technology.** Articles about this topic cover a variety of issues from astronomy and space exploration, to climate change and global warming, plus advances in various technologies such as data and robotics.
- **Politics/US.** Articles about this topic mainly cover the domestic policy of the United States. For example, there is coverage of the 2020 elections, the prospective candidates for the two parties and their policies.
- **Economy/Finance.** This topic covers issues related to businesses, financial markets, and macroeconomics. For example, articles about this topic cover issues such as investment advice, stock market reports, money supply, government debt, and other similar issues.
- **Lifestyle/Religion.** Articles on this topic cover a variety of lifestyle-related themes, from spirituality and religion to relationship and family-life related advice.
- **Politics/Europe.** Articles about this topic report issues related to European politics, such as Brexit negotiations, EU-Russia relations, tensions between Ukraine and Russia and domestic and foreign policies of European countries.
- **Media/Equality.** Articles about this topic mostly cover controversies around media platforms, with a focus on social media and video platforms such as Twitter, Facebook and YouTube. For example, many articles cover alleged censorship of content by these platforms or users being banned for their behaviour, and controversies involving equality issues such as racism and sexism. We see a spike in articles about this topic in June 2020, after the death of George Floyd sparked an increased news coverage of issues involving race (see Figure A2.3).

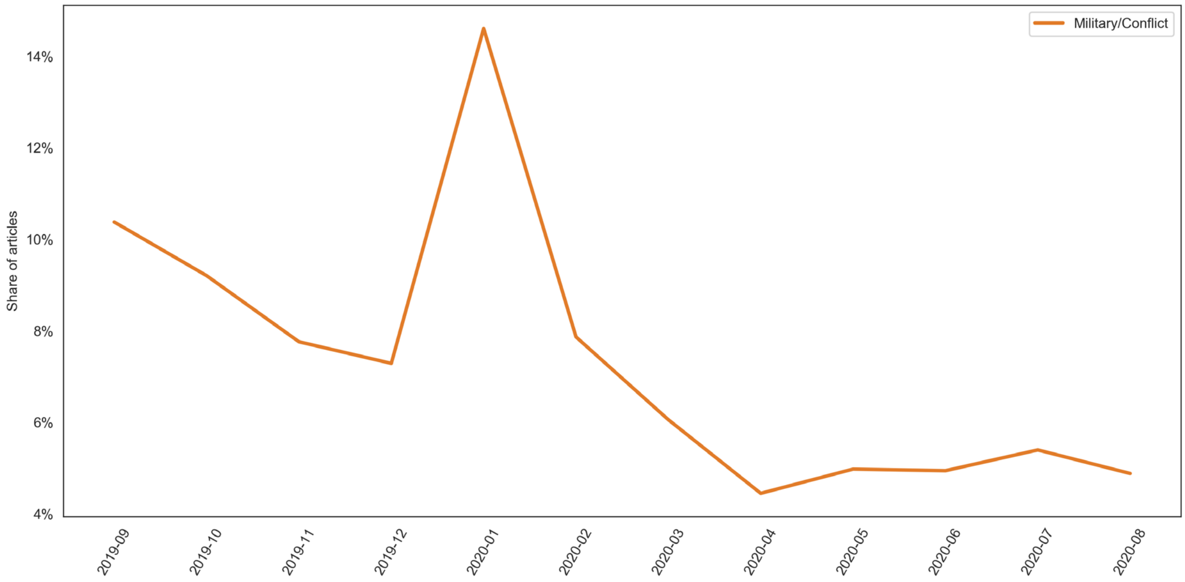
Figure A2.3: Share of articles about the topic 'Media/Platforms' in 2020



Source: Ofcom analysis.

- **Crime.** This topic captures articles that fall under the category of crime journalism, with reports on criminal cases that gained media traction over the time period, such as the Jeffrey Epstein case.
- **Government/Control.** Articles about this topic criticise governments for being anti-democratic, in several cases claiming that they are trying to control their citizens and take away their freedom, often citing conspiracy theories. For example, many articles on this topic make references to a 'deep state' (a supposed group of influential members of government agencies or the military believed to be involved in the secret manipulation or control of government policy) and a 'New World Order' (the hypothesis of a secretly emerging totalitarian world government).
- **Politics/Allegations.** Articles about this topic focus on US politics, however, unlike Topic 6 (Politics/US), the focus of this topic is on political scandals and allegations about public figures such as, for example, the Hillary Clinton email controversy and the alleged links between Russian actors and President Trump.
- **Military/Conflict.** This topic covers conflict between countries, wars, and threats of war, with particular focus on the Middle East and its relationship with the United States. For example, many articles about this topic cover tensions between Iran and the United States, which at the beginning of 2020 sparked talk of conflict on social media. This is reflected in a spike of articles about this topic in January 2020 (see Figure A2.4 below).

Figure A2.4: Share of articles about the topic ’Military/Conflict’



Source: Ofcom analysis.

- **Police/Protests**. Articles about this topic mostly cover episodes of police violence and conflicts between the police and civilians.

One approach to validating the results from the model is to check whether attention to particular topics responds in predictable ways to news events that should affect their prevalence.[159] The time trends we present above therefore support the robustness of our model, as the prevalence of several topics in our model responds as expected to news events.

Additional visualisations

Inter-topic distance map

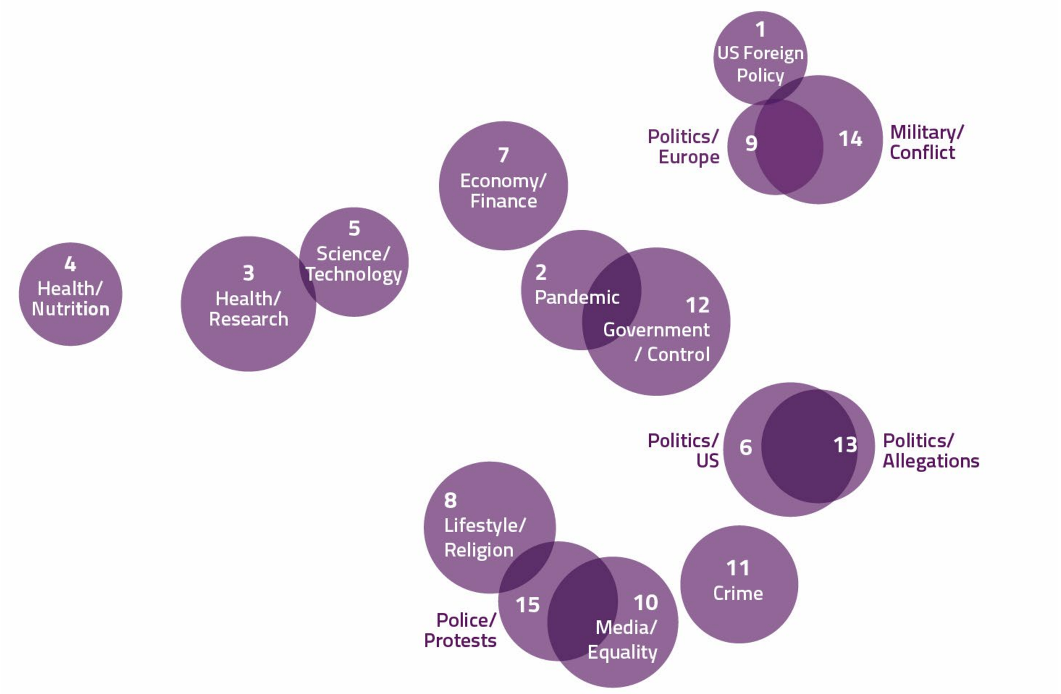
We use the topic modelling visualisation package ‘LDAvis’ to produce Figure A2.5 below.[160] In this figure, the topics are plotted as circles in a two-dimensional plane and each topic is identified by a numeric label. For example, Topic 1 corresponds to the ‘US Foreign Policy’, Topic 2 corresponds to the ‘Pandemic’ topic and so on. The words that make up each topic are presented in Table A2.3 and Table A2.4 above.

The area of the circles represents the overall prevalence of the topic in the corpus (i.e. dataset of articles), that is, the larger the circle the more prominent the topic is in the collection of documents. The distance between the centre of circles represents the similarity between topics, with topics that are semantically more similar to one another being closer together. For example, there are two topics which focus on US politics (6 and 13), which share a large overlap between them. An inspection of the words in these topics and their respective articles reveals that the former is more

[159] P.Di Maggio, M. Nag, and D.M. Blei, 2013. *Exploiting affinities between topic modelling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding.* Poetics, 41(6), pp.570-606.
[160] C. Sievert, and K. Shirley, 2014. *LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces* pp. 63-70.

about general issues in domestic US politics (e.g. coverage of the events leading up to the 2020 election) while the latter is more about political scandals and allegations about political figures (e.g. the Clinton email controversy).

Figure A2.5: Inter-topic Distance Map



Source: Ofcom analysis.

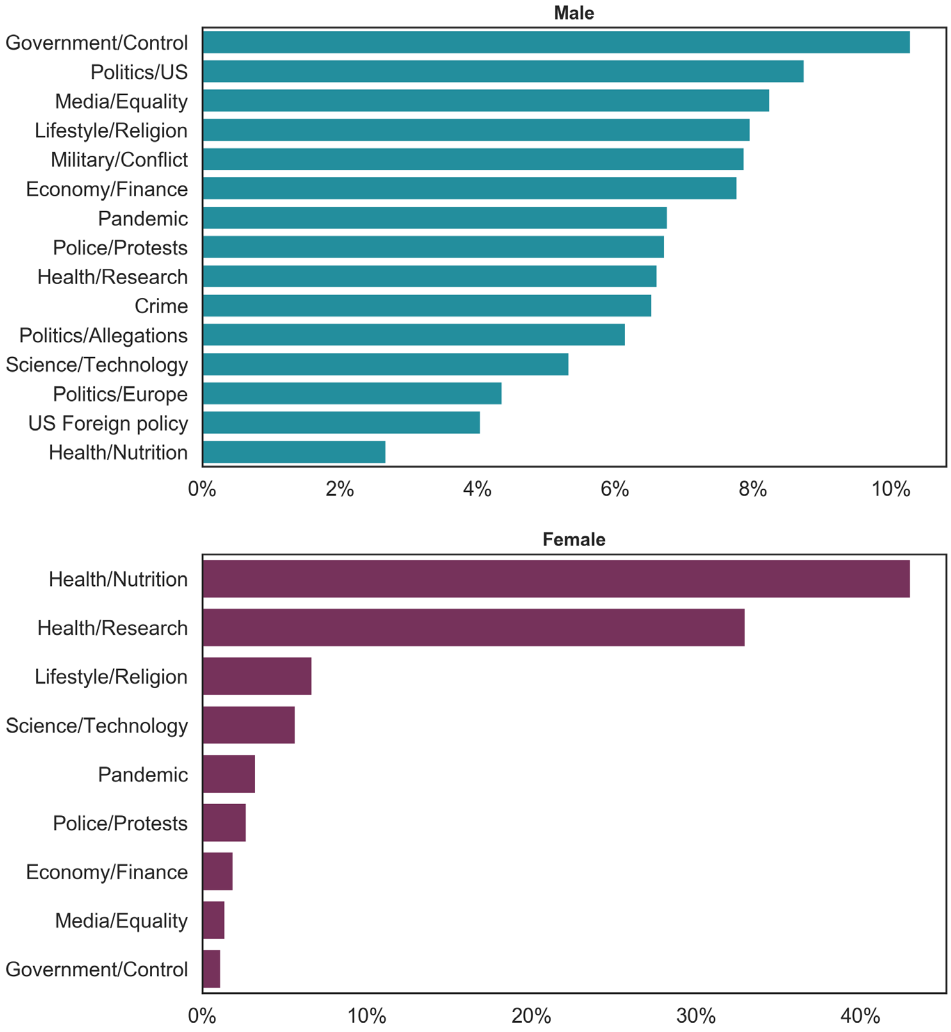
Demographic analysis

As mentioned in Section 3, we find evidence of demographic differences in the topics of the websites visited. In order to perform this analysis, we identify each website by its main demographic (using SimilarWeb's data) and assume that all the articles within that website are read by an audience that matches the website's main demographic. Then, we group the articles by demographic and extract the topic distributions for each demographic. We do so by building demographic-specific corpuses (e.g. a collection of articles on websites whose main demographic is female) and by applying the model that was trained on the whole sample to these collections of articles.

This allows us to extract the topics that are discussed in articles that we assume to be mostly read by a specific demographic (e.g. female or male readers). As mentioned in Section 3 and illustrated in Figure A2.6 below, this analysis suggests that topics about health ('Health/Research' and 'Health/Nutrition') are more common on websites most frequently visited by a female audience. Conversely, the document-topic distribution of the corpus built from websites most frequently visited by males is very similar to the overall document-topic distribution, which is expected as, on average, the share of male users of these websites is about 70%.

One limitation of this analysis is that the information is aggregated at the domain level, while the topic modelling is performed on individual articles. This means that we are effectively assuming that the audience characteristics of the readers of every single article in a website are fixed to the websites' overall audience characteristics. For example, if the gender split of the audience of a website is 80% male and 20% female, this split will apply to every single article within that website, regardless of whether the article is about politics, health or the economy.

Figure A2.6: Figure A2.6: Document-topic distributions by demographic



Source: Ofcom analysis.