



UvA-DARE (Digital Academic Repository)

Online news user journeys: The role of social media, news websites, and topics

Vermeer, S.; Trilling, D.; Kruikemeier, S.; de Vreese, C.

DOI

[10.1080/21670811.2020.1767509](https://doi.org/10.1080/21670811.2020.1767509)

Publication date

2020

Document Version

Final published version

Published in

Digital Journalism

License

CC BY-NC-ND

[Link to publication](#)

Citation for published version (APA):

Vermeer, S., Trilling, D., Kruikemeier, S., & de Vreese, C. (2020). Online news user journeys: The role of social media, news websites, and topics. *Digital Journalism*, 8(9), 1114-1141. <https://doi.org/10.1080/21670811.2020.1767509>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Online News User Journeys: The Role of Social Media, News Websites, and Topics

Susan Vermeer , Damian Trilling , Sanne Kruike-meier  and
Claes de Vreese 

Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

The complexity and diversity of today's media landscape provides many challenges for scholars studying online news consumption. Yet it is unclear how news consumers navigate online. Moving forward, we used a custom-built browser plug-in—passively tracking Dutch online news consumers 24/7—to examine how context (website) and content (news topic) features affect patterns of online news consumption. This resulted in a data set containing more than one million Web pages, from 175 websites (news websites, search engines, social media), collected over 8 months in 2017/18. We used automated content analysis to retrieve news topics, and estimated Markov chains to detect consumption patterns. Our findings indicate that news consumers often directly visit their favorite (typically mainstream) news outlet, and continue browsing within that outlet. We also found a strong preference for entertainment news over any other topic. Although social media often offer entertainment news, they are not necessarily the starting point to such news.

KEYWORDS

Online news consumption; social media; political news; entertainment news; Markov chains; supervised machine learning

Introduction

Being informed about important and relevant public issues is essential for a functioning democracy. News media play a crucial role in providing people with diverse, multi-faceted perspectives on political and public issues (Eveland and Schmitt 2015). Recently, the very nature of news consumption has changed drastically as people increasingly use the Internet as their primary news source (Newman, Levy, and Nielsen 2019). The Internet offers an enormous amount of available information sources and channels, as well as greater opportunities for interaction and co-creation among news consumers. As a consequence, people increasingly find and access online news not only directly via news media organizations' own websites, but also via a great range of other pathways, including search engines and social media (Nielsen and Schröder

CONTACT Susan Vermeer  s.a.m.vermeer@uva.nl

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

2014). People actively combine these different websites into complex patterns of media use. More importantly, patterns of news consumption also vary depending on the topic people are informed about (Kleppe and Otte 2017; Nielsen and Schröder 2014). Scholars are concerned that people may predominantly consume entertainment news while neglecting political information online (see e.g. Pearson and Knobloch-Westerwick 2018).

The shift toward online news consumption has not only had a fundamental impact on consumption patterns, the multitude of different pathways and topics also represents many challenges for researchers studying news use today. Previously, most research methods analyzing online news consumption focus on recalled user behavior by survey, diary, or interview methods (e.g. Taneja et al. 2012; Tewksbury 2003). Self-reports may, however, be less accurate as some people tend to over-report news usage, because it is socially desirable or it is hard to recall content they consumed (Prior 2009). Recently, more accurate avenues for studying online news consumption have opened up. Such approaches include analyses of log files (i.e. visit records) provided by website owners (e.g. Menchen-Trevino and Karr 2012) or passive tracking through third-party panels (e.g. Glenski, Pennycuff, and Weninger 2017). However, these approaches do not always provide insights into the actual individually encountered information and the access to information via different media (for an overview see e.g. Haim and Nienierza 2019).

To overcome this limitation, we deployed a custom-built browser plug-in (collecting Web activities 24/7 of a group of Dutch respondents), to (1) passively track online media use, and (2) analyze the way news consumers find and act upon a wide range of online news outlets, search engines as well as social media, and news topics, and (3) retrieve a more detailed picture of online news consumption. We focus on browsing behavior on desktop and laptop computers. By means of this innovative research method, we aim to better understand how people navigate today's media landscape; also, whether it makes selection of particular news topics (e.g. politics, business, entertainment) more likely. More specifically, we focus on the following question: To what extent do *context* (i.e. the type of website) and *content* (i.e. the news topic) features affect patterns of online news consumption? In the remainder of this article, we consider the *type of website* in which the news item was encountered (news website, search engine, social media) as a context feature, whereas content features refer to the *actual topic* of encountered news items (e.g. politics, entertainment).

To make sense of the patterns of news use online, we applied Markov models to passively tracked user data. Markov chains allow us to examine the sequence of news-related pages a user is visiting. We already know from television-viewing behavior, for example, that it is at least partly shaped by lead-in and lead-out effects: watching a given television show can make it more likely to watch the show prior and subsequently on the same channel (Wonneberger, Schoenbach, and van Meurs 2012). Previous work by Möller et al. (2019) has illustrated that Web users can also be distinguished based on certain modes of news use online (e.g. using search engines to access news online). Despite this, there are several gaps in our knowledge about the complex patterns of media use, particularly the interrelation between news websites and topics. Using a Web mining approach, this study fills this gap with an in-depth

understanding of the sequentiality of online news consumption, by focusing on the following: (1) online user behaviour, (2) context features (i.e. the type of website), as well as (3) content features (i.e. the news topic).

This study contributes to the literature in two important ways. First, our findings improve our understanding of gatekeeping processes in an online news environment. As people are less restricted by traditional gatekeepers, news users have to find their way through a complex world of information as they navigate online (Pearson and Kosicki 2017)—a challenge they increasingly solve by relying on algorithmic agents (Thurman et al. 2019). In the current study, we examine whether people use social media and/or alternative media outlets to access more news online. Second, this study offers an empirical confirmation of how content and context features of online media interact in affecting how people consume news.

Online News Consumption

Gatekeepers in an Online News Environment

The Internet has brought various changes to the way people consume news, as it (1) allows people to have increased control over the news they select; (2) provides a wide variety of sources to keep informed about public issues; and, (3) changes the gatekeeping process so that it is less controlled by traditional media (see Majó-Vázquez, Cardenal, and González-Bailón 2017). According to gatekeeping theory, gatekeepers “facilitate or constrain the diffusion of information as they decide which messages to allow past the gates” (Shoemaker and Vos 2009, 21). For decades, traditional gatekeepers (e.g. journalists, print and broadcast media organizations) exercised control over what information reached society and how social reality was framed (Shoemaker and Vos 2009). In recent years, however, gatekeeping tasks are increasingly carried out by non-journalistic actors and platforms (Wallace 2018). For example, individuals contribute to news coverage on social media (e.g. uploading media) and alternative news outlets (e.g. BuzzFeed, The Huffington Post) compete successfully against traditional news outlets for power over the public agenda (Fletcher and Park 2017). Additionally, news users increasingly rely on algorithmic agents (Thurman et al. 2019): Google’s algorithm compiles our search results, news websites recommend us certain news content, and Facebook’s news feed algorithm compiles our daily dose of news. In this way, algorithmic selection can increase personalization, individualization, and customization (Shin and Park 2019). Welbers and Opgenhaffen (2018) indeed found that news organizations’ Facebook pages can have a strong influence on the diffusion of news items on Facebook. All in all, algorithmically driven technology is drastically revolutionizing society and becoming an integral part of everyday life.

As a result, news is mostly free to access and people consume individual news items from a number of different outlets, accessed via a variety of pathways (Pearson and Kosicki 2017). Hence, the Internet has resulted in a complex labyrinth of information, in which the options for news consumption are almost infinite (Pearson and Knobloch-Westerwick 2018).

With the rise of online news consumption, various scholars have proposed alternative frameworks and metaphors to supplement gatekeeping theory. Bruns (2005)

introduced the concept of *gatewatching*: “the observation of the output gates of news publications and other sources, in order to identify important material as it becomes available” (Bruns 2005, 17). In other words, *gatewatchers*, for example on social media, monitor the content coming through the gates from journalism and share it with specific audiences.

To make sense of the patterns of news use online, Thorson and Wells (2016) developed a framework for mapping information exposure. Using the concept of curated flows, they outline how news flows are curated by various actors, such as social networks and algorithms. Although every user’s experience is unique—given the complexity of the information environment today—it is important to discover trends and patterns within algorithmically curated media environments (Thorson and Wells 2016).

To retrieve a more detailed understanding of these trends and patterns, Pearson and Kosicki (2017) sought to supply a way-finding framework. They describe the online news process as a news users’ journey from the moment they arrive on the Internet through the necessary choices they undertake to arrive at the news content they want. One key assumption of the way-finding framework is that news consumption is increasingly passive (Pearson and Kosicki 2017), as people are simply wandering or browsing through the Internet and can be incidentally exposed to news (e.g. via social media, see Gil de Zúñiga, Weeks, and Ardèvol-Abreu 2017). Additionally, news users encounter and consume news items from multiple news outlets, not because they have not found one they are satisfied with, instead they desire and appreciate the ability to gather information from a variety of news outlets depending on the topic and occasion (Pearson and Kosicki 2017).

A High-Choice Media Environment: Cause for Concern?

The availability of a wide variety of news outlets has raised concerns about the democratic implications of such a high-choice media environment. According to Prior (2007), people’s preferences have become more important predictors of news outlets and content they expose themselves to. “The greater media choice there is, the more selective people have to be; and the more selective people have to be, the more important their preferences become” (Strömbäck, Falasca, and Kruike-meier 2018, 2). It has never been as easy for people highly interested in politics to decide when, where, and through which outlet(s) they consume political news. For people less interested in politics, however, it has become easier to decide—either intentionally or incidentally—whether they want to expose themselves to political news. In this way, increased media choice enables some to easily avoid (political) news content all together (Prior 2007). Another concern is that commercialized news media increasingly seek to grab attention with entertainment news to capture the attention of news readers (Tandoc 2014). Hence, the growing offer of (and interest in) entertainment news seems to precede the offer of (and interest in) political news (de Waal and Schoenbach 2010).

Although such concerns exist, theoretical and empirical assumptions about online news consumption and its implications often ignore the combined influence of two important components which are of decisive influence online: context features and

content features (Orellana-Rodriguez and Keane 2018). We argue that these two main categories of features should be explored in understanding patterns of online news consumption. Context features can relate to *how* the news is accessed (e.g. search engines, social media). Content features concern *what* is written in a news item; for instance the topic of the news item (e.g. politics, business, entertainment). In the following, we discuss the existing literature before drawing hypotheses to test to what extent context features and content features interact in today's media ecosystem.

Context Features

First, the context in which online news is consumed is of great importance as our media system is becoming increasingly hybrid (Chadwick 2013). Previous studies elaborated on the notion of context in many different ways, ranging from different popularity cues (Haim, Kümpel, and Brosius 2018) to different accompanying user comments (Anderson et al. 2014). In a recent study, Orellana-Rodriguez and Keane (2018) capture the context surrounding the posting of a news-tweet. In this way, context features relate to the conditions in which the tweet is posted, such as time and day of publication, location of the author, or geographical focus of the tweet. In line with recent work by Haim and Nienierza (2019), we address context as the way in which users encounter news. More specifically, we consider the type of website in which the news was encountered (e.g. news website, search engine, social media) as a context feature.

Using a gatekeeping model, a news user arrives at a news websites' home page and subsequently selects a news item of interest from those available. The gatekeeper, in this case the journalist, editor, or recommender system (Möller et al. 2018), selects the news items a news consumer can choose from online (i.e. the headlines on the news website). On the contrary, Pearson and Kosicki (2017) suggest that a news user can also read news via a search engine or social media platform, and browse through stories based on their interests, searched keywords, or network. For instance, by searching and subsequently selecting a certain news item in a search engine, the news user can be redirected to a news website. In this way, the user is directed straight to the news article, and does not visit the website's home page. Thus, rather than choosing from a collection of news articles made by a journalist (the gatekeeper on the news website), a news user selects from a compilation of news items from a variety of sources pulled together by a search engine.

To understand this process, Newman, Levy, and Nielsen (2019) examined how citizens encounter online news. Overall, online news consumers are most likely to access news by going directly to a news website or app (29%), and to a somewhat lesser extent through search engines (25%) and social media (23%; Newman, Levy, and Nielsen 2019). However, they found very significant country differences (e.g. due to market size, competition, and regulation). The results indicate that online news consumers in countries like the United States and Canada use many different routes to online news content. Interestingly, online news consumers in countries, such as Finland, Norway, Sweden, and the Netherlands, were far more likely to access news by going directly to a news website or app, and to a lesser extent through search

engines and social media. This resonates with results by Kleppe and Otte (2017), who explored online behavior in the Netherlands, and found that 59% of the news items that were read on news websites were accessed via a news website's home page, indicating that this is still the main driver of traffic to news items. Particularly, traffic to mainstream news websites from, for example, search engines accounts for a very small fraction of their total traffic. Recent work exploring users' pathways to online news indicate that—despite the rise of search engines and social media—direct access to a news website (i.e. the homepage) remains the most important gateway to online news. All in all, there is evidence that most online news consumption results from individuals visiting a news website's homepage (Flaxman, Goel, and Rao 2016).

Yet, most previous studies, as the Digital News Report (Newman, Levy, and Nielsen 2019), relied on respondents' self-reported measures in retrieving survey, diary, or interview data (see also Tewksbury, Weaver, and Maddex 2001), while various scholars asserted that media use is often overestimated (Prior 2009). The shift toward online news consumption has not only had a fundamental impact on consumption patterns, the multitude of different pathways and topics also represents many challenges for researchers studying news use today. Since users encounter news in many different ways, it is more challenging for respondents to reliably self-assess exposure to news online. Hence, we expect that news consumers' actual online behavior may show a more accurate picture. Based on previous work, we expect the following:

H1: *News websites are more likely to serve as an entry-point to access news online compared to search engines or social media.*

Another critical issue facing online news consumption is identifying which news experiences tend to inspire follow-up actions, such as searching for additional information (Nielsen and Schröder 2014). Online information is more immediate, as information can be consumed at any time and in an interactive way; for example by finding more information via hyperlinks (Strömbäck, Falasca, and Kruikemeier 2018).

As news users increasingly shift their attention to online news and traditional circulation declines (Franklin 2008), it is increasingly important to understand what websites news users access throughout their journey. Following the uses and gratifications tradition, news users actively select among news outlets based on their ability to gratify their needs for information, entertainment, social interaction, and self-expression (Diddi and LaRose 2006; Lee 2013; Ruggiero 2000). Different news consumption *motivations* lead to different news consumption *choices*. News users seeking social interaction are more likely to use Facebook and Twitter, whereas opinion validation is motivational to consume news from national newspapers and broadcasters (Lee 2013).

Recent work has demonstrated that news content differs considerably between tabloid, broadsheet, regional, and financial newspapers (Boukes and Vliegenthart 2020). Broadsheets and tabloids, for example, often not have a divergent news agenda, different journalistic values and styles, and a specific target audience. Besides a diverging identity, Richardson and Stanyer (2011) found that tabloid and broadsheet online newspapers differ in the way that news users use the interactive opportunities (e.g. message boards, polls, chat rooms) to express themselves.

Although previous work has focused on what type of websites have been accessed (as filling out surveys can be very labour-intensive; Newman, Levy, and Nielsen 2019),

it remains unclear how likely it is for people to change from one particular website to another website (e.g. changing from a search engine to a specific news outlet). By deploying a custom-built system, we aim to understand how people combine different websites into complex patterns of media use, and how social media and search engines affect this process. This allows us to depict a comprehensive picture of citizens' media use and news consumption rather than focusing on one specific outlet at a time. By exploring this process as a *news user journey*, which according to (Pearson and Kosicki 2017, 1090) can be defined as: "guiding the news user from the moment they arrive on the Internet through the necessary choices they undertake to arrive at the content they want", allows us formulate the following research question:

RQ1: *How do news consumers combine different outlets (e.g., news websites, search engines, social media) while browsing online?*

Content Features

Beyond *how* news is accessed online, the online news experience also varies depending on *what* type of news is accessed (i.e. news topics; Nielsen and Schröder 2014). In a high-choice media environment, politics constantly competes with entertainment (Prior 2005); as many people generally prefer the latter (Prior 2007). Entertainment news—as opposed to political news—refers to news that lacks a policy component; rather it includes subject matters such as crime, disaster, celebrities, and sports (Baum 2002). Tewksbury (2003) conducted a survey in the U.S. to examine the topics people select at online news websites. The results indeed indicate that online news users choose to read about politics and public issues less frequently, compared to news about sports, arts, and entertainment. Since both entertainment and political news are available on numerous websites, people's content preferences become key to understanding online news consumption (Prior 2005). For example, politically interested people are able to access more information and increase their political knowledge, and those who prefer non-political content can easily escape the news, pick up less political information, and opt for more entertaining options (Sunstein 2007).

Gatekeepers (e.g. news organizations, journalists) have a strong impact on citizens' awareness of political and public issues, as they decide which events are covered and how these events are described in news messages (Chong and Druckman 2007; Welbers et al. 2018). In our current high-choice media environment, self-selection of news become more prominent. In other words, citizens are less guided in their news consumption by traditional gatekeepers. Self-selection of news has been often discussed by previous work and shows that self-selection is particularly important for selecting specific topics (Bennett and Iyengar 2008; Iyengar and Hahn 2009; Pearson and Knobloch-Westerwick 2018). This strand of research notes that politics constantly competes with entertainment news (Prior 2005). More specifically, as a vast amount and almost endless stream of news is provided online, it allows people to consume news content that matches their individual interests and needs, which in turn might increase opportunities to select mainly entertainment content and avoid political news; thereby leading to a decrease in political news consumption, which may have detrimental effects on political knowledge and engagement.

However, conversely, people might still inadvertently consume news and information on the Internet when they are not actively seeking it. More self-selection does not always lead to more control (i.e. power over users' news consumption). People can be exposed to information while reading news on other topics or navigating online for non-political purposes (i.e. "people inadvertently consume news and information on the Internet when they are not actively seeking it", Tewksbury, Weaver, and Maddex 2001, p. 2608), so-called 'incidental exposure'. In the same way that people receive general news and information either intentionally and incidentally (Beaudoin 2008), it is likely that when people use the Internet they will unintentionally be exposed to news and information about politics. Online news websites offer hyperlinks to related stories and enlightening readers about the interconnection of many news events. Accordingly, when people start a Web session, it is possible they may encounter news items or political information without the motivation to become informed (Kim, Chen, and Gil De Zúñiga 2013). Tewksbury, Weaver, and Maddex (2001), using survey data, already indicated that the more frequently people went online, the more likely they were to report incidental news exposure. In turn, incidental exposure is positively associated with both offline and online political participation (Kim, Chen, and Gil De Zúñiga 2013), as well as knowledge about counter-attitudinal political information (Lu and Lee 2019).

It is yet unclear how likely it is for people to change—either intentionally or incidentally—from one particular topic to another topic (e.g. changing from entertainment news to political news). Therefore, we formulate the following research question:

RQ2: *How do news consumers combine various news topics (e.g., politics, business, entertainment news) while browsing online?*

Five the Interplay of Context and Content Features

Besides examining context and content features separately, we also look into their interplay. Research indicates that certain topics in the news are more likely to be accessed through one pathway over another (Pew Research Center 2017). Then, another question that arises is whether particular pathways to news are able to support exposure to particular news topics, such as political news (Pearson and Kosicki 2017). Focusing on both context and content features allows to understand how different types of news websites and topics set the boundaries for a media environment (see e.g. Elenbaas et al. 2014).

Exposure to information and news stories increasingly occurs through online social networks (Bakshy, Messing, and Adamic 2015). According to a recent study of Newman, Levy, and Nielsen (2019), Facebook and Twitter are used by many users to discover news, which is eventually discussed via messaging apps such as WhatsApp. Such networks have the potential to expose users to more diverse viewpoints, and at the same time have the potential to limit exposure to attitude-challenging information. The content people encounter on social media, for example, depends on their ties (e.g. friends, acquaintances) and what information those ties share (Bakshy, Messing, and Adamic 2015). As Trilling, Tolochko, and Burscher (2017) indicate, topics influence how news is shared on social media. Particularly, news about politics and

foreign affairs is poorly shared, whilst entertainment news is more likely to be shared and engaged with on social media. According to Bright (2016), this can be attributed to people's desire to enhance their social status, as the nature of the shared content reflects on the individual's identity (Trilling, Tolochko, and Burscher 2017).

Pew Research Center (2015) found that 73% of Facebook news consumers regularly encounter entertainment news in their news feeds. Political news, on the other hand, is seen by just over half (55%). Relatedly, Bakshy, Messing, and Adamic (2015) analyzed news feeds of approximately ten million active American Facebook users over a six-month period. They found that merely 13% of all news items posted on Facebook contained hard news (i.e. politics, foreign affairs). Taken together, we propose that users are more often exposed to entertainment news than to political news on social media. Using a browser plug-in, we are able to capture individually encountered content through algorithmic gatekeepers (in this case Facebook). We hypothesize the following:

H2a: *Social media are more likely to expose users to entertainment news than to political news.*

Bakshy, Messing, and Adamic (2015) found that Facebook users merely clicked on 7% of hard news links available in their feeds. Instead, political news is more likely to be discovered by going directly to a news website: about half of the news covering politics occurred through an individual going directly to a news website (46%, see Pew Research Center 2017). Entertainment news, on the other hand, is much more likely to be accessed through social media compared to news websites and search engines. Everything considered, it can be argued that entertainment news emerges as a social online news experience (Pew Research Center 2017). We hypothesize:

H2b: *Entertainment news is more likely to be accessed via social media than via news websites or search engines.*

Method

Data Collection

As part of a larger project, we recruited respondents via the LISS (Longitudinal Internet Studies for the Social Sciences) panel of CentERdata, which is based on a true probability sample of the Dutch population (Scherpenzeel 2011). Panel members were included in our sample if they (1) are aged eighteen years and older, and (2) agreed to install a browser plug-in on their personal computer. We collected data from our respondents in two ways: having them fill out an online survey and tracking their online media use (for more information see Möller et al. 2019).

Panel members installed a Google Chrome or Mozilla Firefox plug-in *Robin*, a custom-built system that registers respondents' Web behavior, and have their browsing behavior being monitored in the period between 1 July 2017 and 15 March 2018. Our sample is not necessarily representative of all Web behavior (e.g. smartphone browsing, browsing using work computers are not considered). We were unable to include consumption patterns on mobile devices due to limited resources as well as technical constraints. We focus on browsing behavior on desktop or laptop computers.

Nonetheless, in 2017, computers (52%) were the main access point for online news in the Netherlands compared to smartphones (49%) or tablets (26%; Newman et al. 2017).

The plug-in software was compatible with MacOS, Linux, and Windows. Whenever respondents accessed one of the white-listed websites (which includes an exhaustive list of news websites, as well as social media and search engines), the plug-in transmitted all Web traffic (HTTP and HTTPS) to our servers. In this way, the system not only collects the URL request (web page address), but also the date, time, session ID and referer URL (i.e. the address of the Web page that linked to the resource being requested) for each recorded visit.¹ To guarantee respondents' privacy as much as possible, we filtered the raw data to exclude sensitive information. We stored the data in an Elasticsearch database on a server that is not directly available for the researchers. Instead, *Robout*, a Python library, is made available on another secured server to complement *Robin*. We conducted the analyses using *Robout* and a Elasticsearch database on the second server so no sensitive data would leave the environment (more information, see Möller et al. 2019).

To be able to analyze respondents' online activities, the raw data needs to be captured and transformed into relevant and meaningful information. During the first stage of the pre-processing phase, Web behavior is extracted for 175 news websites (e.g. www.telegraaf.nl, www.volkskrant.nl). In our study, a news website refers to a web page whose primary offering is news content. Besides, we included search engines (e.g. www.google.nl, www.yahoo.com) and social media (e.g. www.facebook.com, www.twitter.com) to gain a better understanding of online news patterns. Second, the raw data is cleaned by removing entries irrelevant for our goals (e.g. advertisements). In the third stage of the pre-processing phase, we determined what we consider a Web session. We implemented the following algorithm: (1) group the tracking data by user ID and time, (2) create a new session if there is a 30-minute gap between records in the data (using the same methodology as Athey and Mobius 2012), and (3) assign the same session ID to records that are connected.

Respondents

Figure 1 illustrates an overview of the data collection and pre-processing phase. At the end of the process, we have a set of 354 respondents who constitute the user base that we use for further analyses: 48.5% were male, mean age was 47.2 ($SD=19.2$), and 15.7% had a low level of education (e.g. primary school), 38.3% had a medium level of education (e.g. college), and 44.6% had a high level of education (e.g. university). Compared to the Dutch population as a whole, the sample is slightly older.

Measures

We focus on the following measures:

- *Entry-point*: The first recorded visit of each Web session has a referrer (e.g., search engine, social media or any other website);

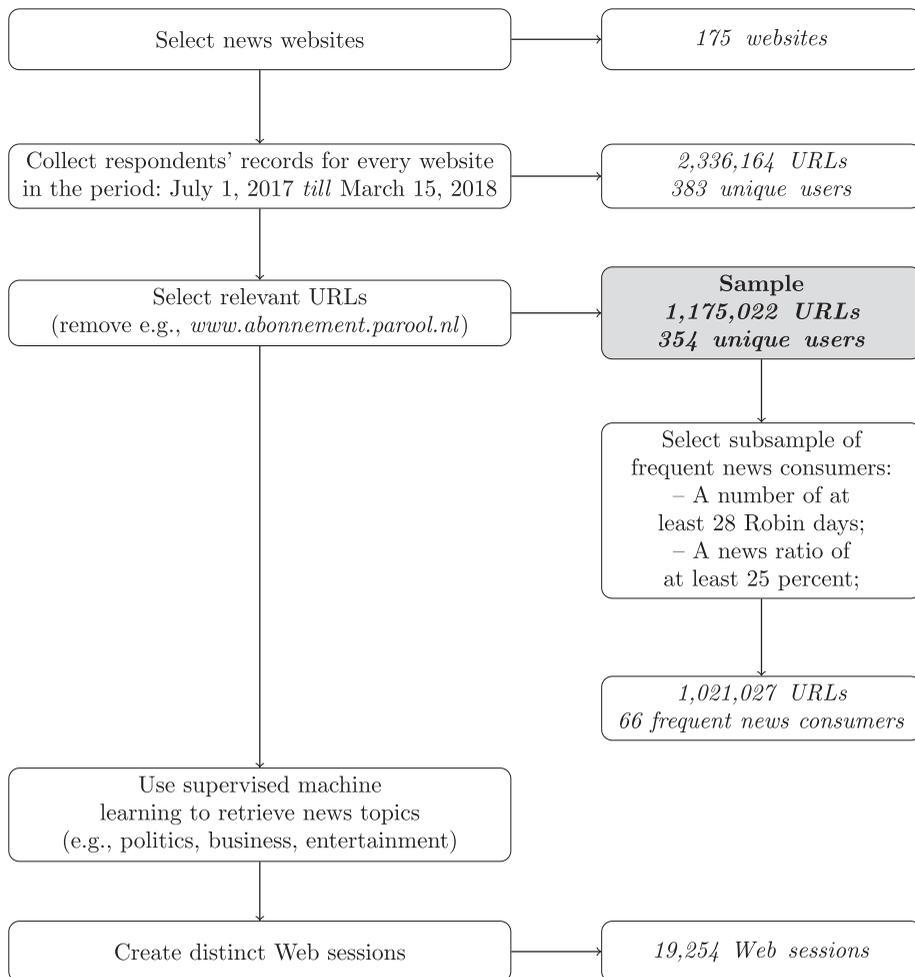


Figure 1. Phases of data collection and pre-processing.

- *Context*: (1) Tabloids, (2) Broadsheets, (3) Online-only outlets, (4) International outlets, (5) Broadcasters, (6) Facebook, (7) Twitter, (8) Search engines, (9) Other news outlets, and (10) Other websites;
- *Content*: (1) Politics, (2) Business, (3) Entertainment, and (4) Other.

Analytical Strategy

To obtain further insights into the frequency of visits to particular Web pages we explored which websites (*context* features) were accessed. Next, we used automated content analysis to retrieve the topic of each news item (*content* features). Finally, to detect usage patterns (in terms of *context* and *content* features), we modeled Web sessions using Markov chains.

Automated Content Analysis

Based on a supervised machine learning method, we determined what topic is covered in each news item. In order to examine news topics, we extracted the raw text (HTML) for every accessed URL. Then, we used the package *jusText* (Pomikálek 2011) to strip HTML markup or JavaScript code and remove non-textual content, such as navigation links, headers and footers. This resulted in the main content, containing full sentences, of every accessed news item. Eventually, we stored the data in an Elasticsearch database. In order to effectively train our classifier, we could merely use Dutch-language news items (i.e. English, German, French and Frisian websites were excluded when examining the *content* of news websites; $n_{\text{websites}} = 18$). To train the classifier, we used a coding scheme developed by Shoemaker and Cohen (2005) to guide the coding process. We distinguished four news categories: (1) Politics (e.g. *internal politics*, *international politics*, and *military and defense*); (2) Business (e.g. *economy*, and *education*); (3) Entertainment (e.g. *sports*, *culture*, and *human interest*); and (4) Other (e.g. *science and technology*, *environment*, and *religion and beliefs*). If an item would suitably be annotated as being relevant to more than one topic, we annotated the most dominant topic present when merely reading the first five sentences of a news item.

Two human coders independently coded approximately 500 news items. The assignment of news items to these four categories reached a Cohen's kappa score of .88. On this basis, one coder analyzed an additional 3200 news items in a step-wise approach.

Next, we used the Python *scikit-learn machine learning* library (Pedregosa et al. 2011) to train and test the Passive-Aggressive (PA) algorithm, which is known to perform well in various text classification tasks (Crammer et al. 2006), including Dutch-language news items (Burscher, Vliegthart, and De Vreese 2015). We applied a random sampling procedure to split the data set into a training set (80%), on which we trained the classifier, and a test set (20%), on which we evaluated the classifier (Burscher, Vliegthart, and De Vreese 2015). Eventually, we measured its ability to accurately classify unseen labeled, which computed the following scores: *precision* of .82, *recall* of .83, and *accuracy* of .82. Finally, we used the classifier to predict the topic for all news items outside the training set.

Markov Chains

Furthermore, we use Markov chains (absorbing Markov chains, in particular), as described by Vermeer and Trilling (2020) to answer RQ1 and RQ2.² Markov chains are mathematical systems that change from one state (i.e. a situation) to another state. All possible states are listed in a so-called *state space*. A Markov chain provides the likelihood of transitioning from one state to any other state (see Gilks, Richardson, and Spiegelhalter 1996).

Figure 2 illustrates an exemplary Markov process with two states (A and B) in the state space. In total, there are four possible transitions, as a state can transition back into itself. Each number represents the probability of the Markov process changing from one state to another state, with the direction indicated by the arrow (Gilks, Richardson, and Spiegelhalter 1996). For example, if the Markov process is in state A,

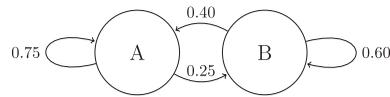


Figure 2. Example of a two-state *Markov Chain* process.

then the probability it changes to state B is 0.25, whereas the probability it remains in state A is 0.75.

In the context of this study we use two types of Markov chains:

- *Context*: The probability of users changing from one website to another website, using the state space described in the *Measures* section;
- *Content*: The probability of users changing from one news topic to another news topic, using the state space described in the *Measures* section.

The begin state is the first web page (or first state) of every Web session. Furthermore, we added a final ‘absorbing’ state (i.e. End of session) to every Web session, representing the exit point (i.e. a state once entered, cannot be left).

Robustness Check

To verify the robustness of our findings, we selected a subsample of panel members that we call *frequent news consumers* ($N=66$): 64.6% were male, mean age was 49.7 ($SD=15.1$), and 18.5% had a low level of education, 41.5% had a medium level of education, and 40% had a high level of education. We used the following measures to select frequent news consumers:

- *Robin days*: The respondent has at least 28 days of Web activity (i.e., the number of days Robin recorded Web activity between the first and last record) during the period under study (i.e. 258 days);
- *News ratio*: The number of news consuming days (i.e., the number of days of recorded visits on news websites); set against *Robin days* (in %) is at least 25%.

The main results hold.

Results

After the pre-processing phase of the collected data, we had 1,175,022 relevant URLs providing us insight in the Web behavior of 354 news consumers.

Context Features

First, we indicated whether respondents used a referrer, for example social media, as an entry-point to access news online. [Figure 3](#) presents the first five steps of a Web session, in which step 0 is the entry-point, and 1–5 are follow-up actions. Interestingly, at the entry-level, search engines and Facebook seem to serve as important gateways for tabloids and/or other news outlets. Access to tabloids increased from 15.4% at

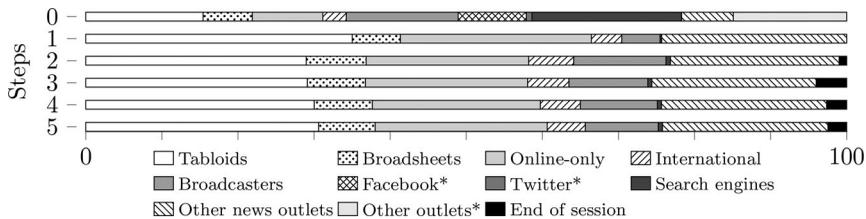


Figure 3. Steps 0 – 5 (0 = entry-point).

Note. *Facebook, Twitter and Other websites are merely taken into account as an entry-point.

step 0 to 35.0% at step 1, and access to other news outlets increased from 6.8% to 24.3%.

To examine H1, we examined the entry-point to a Web session. The observed frequencies indicate that online news consumers were most likely to get news by going directly to news websites ($n=4,611$), followed by search engines ($n=1,627$), other websites ($n=1,234$), and social media ($n=804$). To examine whether news websites are more likely to serve as an entry-point to access news online compared to search engines or social media, we used multilevel multinomial logit models (i.e. including a random intercept). In this way, we consider the nested structure of respondents' Web sessions. The results indicate that news consumers are more likely to use news websites as an entry-point to access news, compared to social media ($b = -.98, p < .001$), search engines ($b = -.28, p = .03$), and other websites ($b = -.55, p < .001$; $N_{\text{obs}} = 8,276, N_{\text{groups}} = 331$; $\text{MLE} = -7,665.82$). Accordingly, H1—news websites are more likely to serve as an entry-point to access news online compared to search engines or social media—is supported by the data. This means that users have a strong tendency to directly visit the news website's home page when they want to keep up with the news.

To retrieve a closer examination of the data, we estimated Markov chains. This allowed us to capture sequential dependence by modeling users' navigational behaviors. For every existing Web session, we created a pattern of states (e.g. *Facebook* → *tabloid* → *tabloid* → *end of Web session*). To explore the pattern of accessed news outlets, we constructed transition matrices (for every respondent; $N = 354$) representing transition probabilities between all possible states (i.e. news websites). Based on the transition probability matrix, we define the transition paths at an aggregate-level.

Figure 4 represents any possible transition in terms of context. In this way, we argue that the current website is expected to be a good clue to grasp the next website. We assigned weights to the edges to represent the probability of users changing from one news site to another news site. In other words, the thicker the line, the higher the transition probability.

Three patterns are particularly clear in Figure 4. First, the results indicate that search engines are an essential entry-point for any type of news outlet, particularly for other news outlets (20.47), as well as for broadcasters (13.43), online-only outlets (9.30), broadsheets (8.61), and international outlets (8.04). This means that users use search engines to access all types of news websites. Second, as highlighted in grey, we found

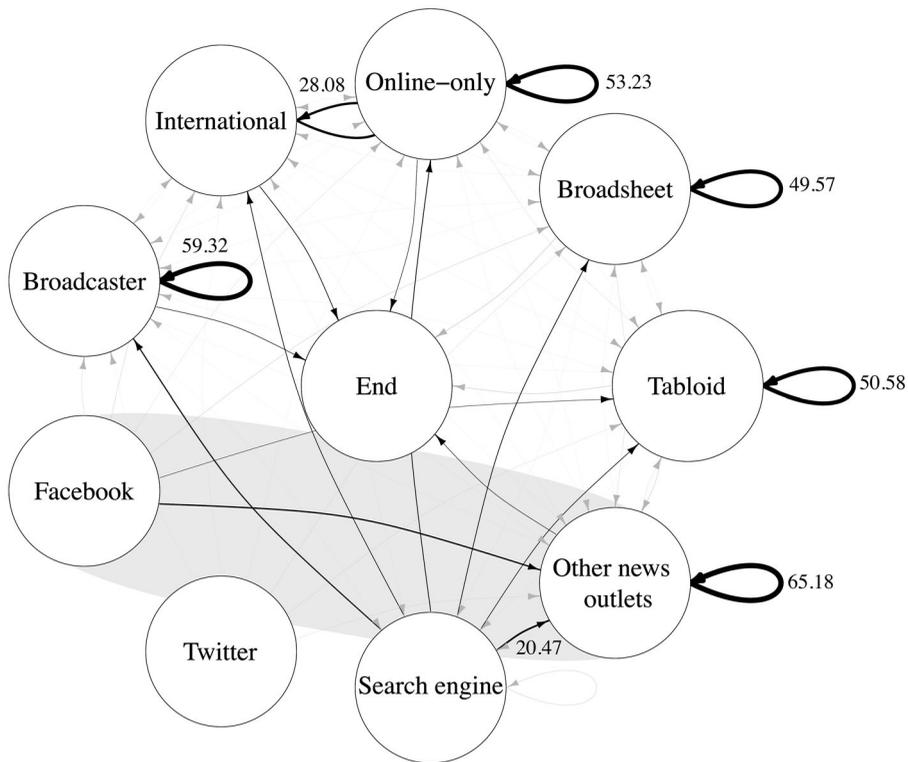


Figure 4. The probability of users changing from one website to another website.

Note. (i) The edges present the mean transition probabilities - the thicker the line, the higher the transition probability; (ii) As Facebook, Twitter and Other websites are merely taken into account as an entry-point they do not have any incoming edges, and it is impossible to transition from these outlets to the end of a Web session; (iii) Mean transition probabilities higher than twenty percent are presented in actual numbers next to the corresponding edge; (iv) Highlighted vertices provide highly relevant insights and are discussed in more detail in the main text.

that particularly other news outlets benefit from Facebook referrals (16.12), compared to, for example, tabloids (4.02) or broadsheets (2.41). Finally, the transition probabilities of interactions within the same outlet: broadcaster, online-only, tabloid, broadsheet, and international, were higher than interactions between different news outlets (for more details see Figure 6). This means that online news consumers have a strong tendency to continue browsing within the same news outlet.

Content Features

We also explored the transitional probabilities with the content identified. Figure 5 presents the probability for every possible transition in terms of content. We can discover two clear navigation patterns

First, after reading a news item (about any topic) users frequently return to a homepage or section page and continue browsing from there (i.e. Politics: 16.18, Business: 16.80, Entertainment: 20.48, and Other: 10.34). This means that the homepage is still important, as pages providing an overview of the news (for a specific topic) help users navigate the wealth of online news.

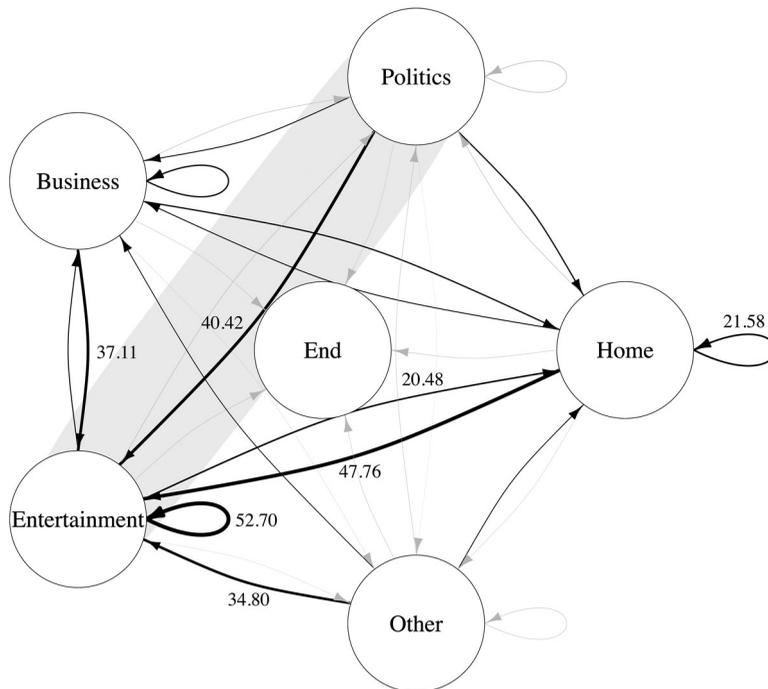


Figure 5. The probability of users changing from one news topic to another news topic.

Note. (i) The edges present the mean transition probabilities - the thicker the line, the higher the transition probability; (ii) Homepage can also refer to section pages; providing an overview of the news for a specific topic; (iii) Mean transition probabilities higher than twenty percent are presented in actual numbers next to the corresponding edge; (iv) Highlighted vertices provide highly relevant insights and are discussed in more detail in the main text.

Second, we found a relative preference for entertainment news over any other news topic. Online news consumers are very likely to transfer to entertainment news during a Web session, and also continue reading entertainment news (see Figure 5). Users are less inclined to read about political issues (as highlighted in grey). Consuming political news, on the other hand, decreases people's desire to continue reading about this topic. Instead, people rather switch to business (11.29) or somewhat less cognitively demanding news items, such as entertainment (40.42). This means that the consumption of entertainment news seems to precede the consumption of any other type of news (for more details see Figure 7).

Context and Content Features

Turning to the interaction between context and content features in online news consumption, our hypotheses predict that social media are more likely to expose users to entertainment news than to political news (H2a),³ and in turn entertainment news is more likely to be accessed via social media than via news websites or search engines (H2b). To test H2a, we examined the content of news items on Facebook. In the period July 1, 2017 till March 15, 2018, 203 respondents were exposed to 149,123 different Facebook items. 4195 Facebook items were news-related. The observing frequencies indicate that social media users were most likely to encounter entertainment

news ($n=3,329$), followed by political news ($n=407$), business news ($n=350$), and other news ($n=77$). In order to determine to what extent social media expose users to various news topics, we used multilevel multinomial logit models (i.e. including a random intercept). The results indicate that social media are more likely to expose users to entertainment news, compared to political news ($b = -2.36, p < .001$), business news ($b = -2.51, p < .001$), and other news ($b = -4.02, p < .001$; $N_{\text{obs}} = 4,195$, $N_{\text{groups}} = 59$; $\text{MLE} = -3,028.77$). This means that, when browsing on Facebook, news consumers are more likely to encounter topics such as sports, culture, fashion, and human interest, and less likely to encounter political news. Accordingly, H2a is supported by the data.

Finally, we examined whether entertainment news is accessed more often via social media than via news websites or search engines. The observing frequencies indicate that people, at the first step of a Web session, access entertainment news via news websites ($n=1,131$), followed by other websites ($n=562$), search engines ($n=513$), and social media ($n=319$). By using multilevel multinomial logit models, we found that news consumers are more likely to access entertainment news via news websites, compared to social media ($b = -.20, p < .001$), and search engines ($b = .28, p < .001$; $N_{\text{obs}} = 2,525$, $N_{\text{groups}} = 239$; $\text{MLE} = -2,535.17$). Accordingly, H2b is not supported by the data.

Discussion and Conclusion

The very nature of news consumption has changed drastically as people increasingly use the Internet as their primary news source. Hence, users face greater difficulty navigating the complex wealth of information, and strive to cope with this information overload. In turn, this could draw them toward increasingly consuming entertainment news (Pearson and Kosicki 2017). To examine such concerns, we aimed to gain a better understanding of online news consumption patterns. Our design enabled us to (1) passively record online browsing behavior, and (2) analyze the way news consumers find and act upon a wide range of online news outlets, search engines, and social media, as well as a variety of news topics (e.g. politics, entertainment).

The results indicate that news websites (compared to search engines or social media) most often serve as an entry-point to a Web session. In line with previous studies (Kleppe and Otte 2017; Newman et al. 2017), the homepage of a news website can be considered as the main driver of traffic to such outlets. Although social media are not often used as an entry-point to a news item, our results do give an indication on the role of social media in the online news consumption process. When examining the entry-point of a Web session, we found that Facebook is particularly important in generating Web traffic to other, mostly local and regional, news outlets. One could argue that users do not need Facebook to find typically mainstream news outlets, or as argued by Mukerjee, Majó-Vázquez, and González-Bailón (2018), users consume most of their mainstream news directly from their news feed. Based on our outcomes, however, we can only speculate about these dynamics. Search engines, on the other hand, seem to serve any type of news outlet. Furthermore, when news users browse through a news website, they are more likely to continue browsing within that same

outlet. The vast majority of online news consumption, thus, resembles offline reading habits, with individuals directly visiting homepages of their (probably) favorite, typically mainstream, news outlets (Flaxman, Goel, and Rao 2016).

In terms of content, we examined how people combine news topics during a Web session. While previous studies have examined people's selection of entertainment news (e.g. Prior 2005), such studies did not quantitatively track what people choose to consume, and how they combine different topics into complex patterns of media use. The results of our tracking data certainly confirm previous findings (e.g. Pearson and Knobloch-Westerwick 2018) that news consumers prefer entertainment news over any other topic (e.g. politics, business). An explanation for this result might be that, due to a commercialized news business, news organizations increasingly seek to grab attention with entertainment news (Tandoc 2014). As a consequence, the growing offer of (and interest in) entertainment news seems to precede the offer of (and interest in) political news. In this way, people seem to predominantly selecting and consuming entertainment news in an online news environment (Pearson and Knobloch-Westerwick 2018).

Finally, we looked into the interplay between content and context features, in order to determine whether certain news topics are more likely to be accessed through one pathway over another. Although users particularly encounter entertainment news on Facebook (due to e.g. friends and acquaintances), which resonates with findings by Bakshy, Messing, and Adamic (2015), entertainment news is not necessarily accessed via social media.

Theoretical Implications

Our findings help contributing to building new theories of gatekeeping processes and news use in an algorithmically curated media environments. Although news consumers are more empowered and have more control over our information flows online, journalistic gatekeepers still play an important role in online news consumption; as news consumers often rely on websites of traditional news outlets when navigating the complex wealth of online news. Our findings also indicate additional insights in the role of algorithmic curation. Using a promising approach to the observation of users within algorithmically curated media environments, we were able to explore the role of Facebook's news feed algorithm in our daily dose of news.

Furthermore, this study can help inform the debate around selective exposure (see Zuiderveen Borgesius et al. 2016). In a high-choice media environment, news consumers are more empowered and have more control over their news flows online. Our findings indeed suggest that interest in (and consumption of) entertainment news seems to precede interest in (and consumption of) political news (Prior 2007), which might be in line with a general trend indicating a growing attention for entertaining topics (de Waal and Schoenbach 2010). This could have serious consequences for society at large, as being exposed to a diverse set of news (particularly political and societal issues) is essential for a functioning democracy.

Practical Implications

Further, this study offers practical implications for news organizations (including journalists and Web developers) to understand how people navigate today's media

landscape. Ferrer-Conill and Tandoc (2018) have indicated, for example, to what extent user data can assess the performance of editorial choices. Given the complexity of online news consumption, it is important to understand the differences in news consumption processes contributed by each of the different (news) websites (Kleppe and Otte 2017). Our findings, for example, indicate that particularly other news outlets (mostly regional news outlets) benefit from Facebook referrals compared to mainstream outlets (e.g. tabloids, broadsheets).

By applying Markov models to a large amount of user behavior data, we were able to provide relevant insights for news organizations seeking to guide users to the relevant content on their website. As journalistic news production is more and more metrics-driven, users' pathways can be a relevant source of support and inspiration for editorial groups (Giromelakis et al. 2019). Utilizing Markov chains could provide relevant insights for news organizations seeking to guide users to relevant content on their website (e.g. enabling personalization, content recommendation). Such insights can, for example, help to determine whether a certain news article should be offered merely to premium members or not. As journalists, editors, and Web developers are the architects of news pathways (Pearson and Kosicki 2017), our findings can serve as a starting point to help them shape the paths users take.

Limitations and Future Research

Methodologically, we demonstrated how to use tracking data to tackle the complexity of studying news use today. Furthermore, we set out to advance a way to construct Markov chains, as these provide a better understanding of which outlets and news topics are combined during a Web session. Setting up this research design of course involved various methodological limitations (more information regarding the challenges of our browser plug-in see Bodo et al. 2017). First among these is the small number of frequent news consumers. The final number of panel members that indeed used our plug-in for an extensive period of time did not meet our expectations. Consequently, we were unable to distinguish between consumer characteristics (such as age, gender, political efficacy, interest, and ideological position) in our analyses. Future research needs to obtain a larger sample to take such features into account.

Second, respondents might have adapted their behavior to the fact that they were under observation. To account for the possibility of social desirability, we passively tracked over a relatively long period of time (over more than 8 months).

Moreover, we excluded other websites (unrelated to news) after the first entry-point of a Web session. People, however, receive general news and information either intentionally and incidentally (Beaudoin 2008); hence it is likely that when people use the Internet they will unintentionally be exposed to, for example, political news and information. Future research that aims to analyze actual news consumption should, therefore, incorporate other websites throughout a Web session to build upon the concept of incidental news exposure (Tewksbury, Weaver, and Maddex 2001). It would also be worthwhile to gain insight into particular needs and preferences regarding news content, for example by conducting in-depth interviews (see Kleppe and Otte 2017).

In the current study, the element of time has not been taken into account (i.e. minimum time on a certain web page). Future studies could clean the data in such a way that allows exploring whether users spend more *time* with political news.

Future work that leverages larger samples would also enable analyses within specific news websites. As media organizations started experimenting with the use of recommender systems to give users personalized recommendation (Möller et al. 2018), news use in algorithmically curated media environments warrants further exploration.

Finally, when studying online news consumption it is increasingly important to record media use on mobile devices (e.g. Kleppe and Otte 2017). In the current study, we were unable to include news consumption patterns on mobile devices such as smartphones and tablets (due to limited resources and technical constraints). Recent studies indicate that specifically young news consumers basically always have their smartphones with them, making it their most important device for news consumption (Newman, Levy, and Nielsen 2019). They often turn to WhatsApp and other direct messaging apps for their news (Newman, Levy, and Nielsen 2019). Future research on the role of news consumption that aims to track actual news consumption should, thus, take the practice of cross-device news consumption into account. Up until now, only few studies have experimented with monitoring news consumption on smartphones and tablets. Accordingly, for future work, it is clear that the increasing cross-device consumption of news asks for new approaches of monitoring news consumption on all possible websites.

Notwithstanding the exploratory nature of our findings, this study has provided a strong set of findings, relevant to (online) news consumption. By focusing on an online news environment, these findings not only update and advance earlier research about gatekeeping processes, but also provide a further understanding of the role of context and content features in news consumption patterns.

Notes

1. We received approval from the Ethics Review Board: 2016-PCJ-741.
2. More information on the Python module see <https://github.com/uvacw/df2markov>
3. We focused on all Facebook items that our respondents encountered throughout the period under study (not merely items that have been accessed through a referral). To obtain encountered news items on a Facebook news feed separately, we programmed an additional custom privacy-friendly parser.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The research was supported by the Research Priority Area 'Personalised Communication' of the University of Amsterdam. The work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

ORCID

Susan Vermeer  <http://orcid.org/0000-0002-9829-8057>
 Damian Trilling  <http://orcid.org/0000-0002-2586-0352>
 Sanne Kruikemeier  <http://orcid.org/0000-0002-8053-8200>
 Claes de Vreese  <http://orcid.org/0000-0002-4962-1698>

References

- Anderson, A. A., D. Brossard, D. A. Scheufele, M. A. Xenos, and P. Ladwig. 2014. "The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies." *Journal of Computer-Mediated Communication* 19 (3): 373–387.
- Athey, S., and M. Mobius. 2012. *The Impact of News Aggregators on Internet News Consumption: The Case of Localization*. <https://www.gsb.stanford.edu/gsb-cmis/gsb-cmis-download-auth/406636>
- Bakshy, E., S. Messing, and L. A. Adamic. 2015. "Political Science. Exposure to Ideologically Diverse News and Opinion on Facebook." *Science (New York, N.Y.)* 348 (6239): 1130–1132.
- Baum, M. A. 2002. "Sex, Lies, and War: How Soft News Brings Foreign Policy to the Inattentive Public." *American Political Science Review* 96 (1): 91–109.
- Beaudoin, C. E. 2008. "The Internet's Impact on International Knowledge." *New Media & Society* 10 (3): 455–474.
- Bennett, W. L., and S. Iyengar. 2008. "A New Era of Minimal Effects? The Changing Foundations of Political Communication New Era of Minimal Effects." *Journal of Communication* 58 (4): 707–731.
- Bird, S., and E. Loper. 2016. The Natural Language Toolkit NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics* (pp. 63–70).
- Bodó, Balázs, Natalie Helberger, Kristina Irion, Frederik Zuiderveen Borgesius, Judith Moller, Bob van de Velde, Nadine Bol, Bram van Es, and Claes de Vreese. 2017. "Tackling the algorithmic control crisis—the technical, legal, and ethical challenges of research into algorithmic agents." *Yale JL & Tech* 19: 133.
- Boukes, M., and R. Vliegthart. 2020. "A General Pattern in the Construction of Economic Newsworthiness? Analyzing News Factors in Popular, Quality, Regional, and Financial Newspapers." *Journalism* 21 (2): 279–300.
- Bright, J. 2016. "The Social News Gap: How News Reading and News Sharing Diverge." *Journal of Communication* 66 (3): 343–365.
- Bruns, A. 2005. *Gatewatching: Collaborative Online News Production*. New York, NY: Peter Lang.
- Burscher, B., R. Vliegthart, and C. H. De Vreese. 2015. "Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?" *The Annals of the American Academy of Political and Social Science* 659 (1): 122–131.
- Chadwick, A. 2013. *The Hybrid Media System: Politics and Power*. Oxford, England: Oxford University Press.
- Chong, D., and J. N. Druckman. 2007. "Framing Theory." *Annual Review of Political Science* 10 (1): 103–126.
- Crammer, K., O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. 2006. "Online Passive-Aggressive Algorithms." *Journal of Machine Learning Research* 7: 551–585.
- de Waal, E. de., and K. Schoenbach. 2010. "News Sites' Position in the Mediascape: Uses, Evaluations and Media Displacement Effects over Time." *New Media & Society* 12 (3): 477–496.
- Diddi, A., and R. LaRose. 2006. "Getting Hooked on News: Uses and Gratifications and the Formation of News Habits among College Students in an Internet Environment." *Journal of Broadcasting & Electronic Media* 50 (2): 193–210.

- Elenbaas, M., C. de Vreese, A. Schuck, and H. Boomgaarden. 2014. "Reconciling Passive and Motivated Learning: The Saturation-Conditional Impact of Media Coverage and Motivation on Political Information." *Communication Research* 41 (4): 481–504.
- Eveland, W. P., and J. B. Schmitt. 2015. "Communication Content and Knowledge Content Matters: Integrating Manipulation and Observation in Studying News and Discussion Learning Effects." *Journal of Communication* 65 (1): 170–191.
- Ferrer-Conill, R., and E. C. Tandoc. 2018. "The Audience-Oriented Editor: Making Sense of the Audience in the Newsroom." *Digital Journalism* 6 (4): 436–453.
- Flaxman, S., S. Goel, and J. M. Rao. 2016. "Filter Bubbles, Echo Chambers, and Online News Consumption." *Public Opinion Quarterly* 80 (S1): 298–320.
- Fletcher, R., and S. Park. 2017. "The Impact of Trust in the News Media on Online News Consumption and Participation." *Digital Journalism* 5 (10): 1281–1299.
- Franklin, B. 2008. "The Future of Newspapers." *Journalism Studies* 9 (5): 630–641.
- Gil de Zúñiga, H., B. Weeks, and A. Ardèvol-Abreu. 2017. "Effects of the News-Finds-Me Perception in Communication: Social Media Use Implications for News Seeking and Learning about Politics." *Journal of Computer-Mediated Communication* 22 (3): 105–123.
- Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. New York, NY: Chapman and Hall.
- Giomelakis, D., E. Sidiropoulos, S. Gilou, and A. Veglis. 2019. "The Utilization of Web Analytics in Online Greek Journalism." *Journalism Studies* 20 (5): 609–630.
- Glenski, M., C. Pennycuff, and T. Weninger. 2017. "Consumers and Curators: Browsing and Voting Patterns on Reddit." *IEEE Transactions on Computational Social Systems* 4 (4): 196–206.
- Haim, M., A. S. Kümpel, and H.-B. Brosius. 2018. "Popularity Cues in Online Media: A Review of Conceptualizations, Operationalizations, and General Effects." *Studies in Communication — Media* 7 (2): 186–207.
- Haim, M., and A. Nienierza. 2019. "Computational Observation: Challenges and Opportunities of Automated Observation within Algorithmically Curated Media Environments Using a Browser Plug-in." *Computational Communication Research* 1 (1): 79–102.
- Iyengar, S., and K. S. Hahn. 2009. "Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use." *Journal of Communication* 59 (1): 19–39.
- Kim, Y., H. T. Chen, and H. Gil De Zúñiga. 2013. "Stumbling upon News on the Internet: Effects of Incidental News Exposure and Relative Entertainment Use on Political Engagement." *Computers in Human Behavior* 29 (6): 2607–2614.
- Kleppe, M., and M. Otte. 2017. "Analysing and Understanding News Consumption Patterns by Tracking Online User Behaviour with a Multimodal Research Design." *Digital Scholarship in the Humanities* 32 (suppl_2): ii158–170.
- Lee, A. M. 2013. "News Audiences Revisited: Theorizing the Link between Audience Motivations and News Consumption." *Journal of Broadcasting & Electronic Media* 57 (3): 300–317.
- Lu, Y., and J. K. Lee. 2019. "Stumbling upon the Other Side: Incidental Learning of Counter-Attitudinal Political Information on Facebook." *New Media & Society* 21 (1): 248–265.
- Majó-Vázquez, S., A. S. Cardenal, and S. González-Bailón. 2017. "Digital News Consumption and Copyright Intervention: Evidence from Spain before and after the 2015 "Link Tax." *Journal of Computer-Mediated Communication* 22 (5): 284–301.
- Menchen-Trevino, E., and C. Karr. 2012. "Researching Real-World Web Use with Roxy: Collecting Observational Web Data with Informed Consent." *Journal of Information Technology & Politics* 9 (3): 254–268.
- Möller, J., D. Trilling, N. Helberger, and B. van Es. 2018. "Do Not Blame It on the Algorithm: An Empirical Assessment of Multiple Recommender Systems and Their Impact on Content Diversity." *Information Communication and Society* 21 (7): 959–977.
- Möller, J., R. N. van de Velde, L. Merten, and C. Puschmann. 2019. "Explaining Online News Engagement Based on Browsing Behavior: Creatures of Habit." *Social Science Computer Review*. doi:10.1177/0894439319828012
- Mukerjee, S., S. Majó-Vázquez, and S. González-Bailón. 2018. "Networks of Audience Overlap in the Consumption of Digital News." *Journal of Communication* 68 (1): 26–50.

- Newman, N., R. Fletcher, A. Kalogeropoulos, D. A. Levy, and R. K. Nielsen. 2017. Reuters Institute Digital News Report 2017 (Tech. Rep.).
- Newman, N., D. Levy, and R. K. Nielsen. 2019. *Reuters Institute Digital News Report 2019* (Tech. Rep.).
- Nielsen, R. K., and K. C. Schrøder. 2014. "The Relative Importance of Social Media for Accessing, Finding, and Engaging with News: An Eight-Country Cross-Media Comparison." *Digital Journalism* 2 (4): 472–489.
- Orellana-Rodriguez, C., and M. T. Keane. 2018. "Attention to News and Its Dissemination on Twitter: A Survey." *Computer Science Review* 29: 74–94.
- Pearson, G. D., and S. Knobloch-Westerwick. 2018. "Perusing Pages and Skimming Screens: Exploring Differing Patterns of Selective Exposure to Hard News and Professional Sources in Online and Print News." *New Media & Society* 20 (10): 3580–3596.
- Pearson, G. D., and G. M. Kosicki. 2017. "How Way-Finding is Challenging Gatekeeping in the Digital Age." *Journalism Studies* 18 (9): 1087–1105.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.
- Pew Research Center. 2015. *The Evolving Role of News on Twitter and Facebook* (Tech. Rep.). <http://pewrsr.ch/1M8rcq2>
- Pew Research Center. 2017. *How American Encounter, Recall and Act Upon Digital News* (Tech. Rep.). <http://pewrsr.ch/2kqjA7t>
- Pomikálek, J. 2011. "Removing Boilerplate and Duplicate Content from Web Corpora." PhD. diss. <https://is.muni.cz/th/o6om2/phdthesis.pdf?so=nx>.
- Pöttker, H. 2003. "News and Its Communicative Quality: The Inverted Pyramid—When and Why Did It Appear?." *Journalism Studies* 4 (4): 501–511.
- Prior, M. 2005. "News vs. Entertainment: How Increasing Media Choice Widens Gaps in Political Knowledge and Turnout." *American Journal of Political Science* 49 (3): 577–592.
- Prior, M. 2007. *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections*. New York, NY: Cambridge University Press.
- Prior, M. 2009. "Improving Media Effects Research through Better Measurement of News Exposure." *The Journal of Politics* 71 (3): 893–908.
- Richardson, J. E., and J. Stanyer. 2011. "Reader Opinion in the Digital Age: Tabloid and Broadsheet Newspaper Websites and the Exercise of Political Voice." *Journalism: Theory, Practice & Criticism* 12 (8): 983–1003.
- Ruggiero, T. E. 2000. "Uses and Gratifications Theory in the 21st Century." *Mass Communication and Society* 3 (1): 3–37.
- Scherpenzeel, A. 2011. "Data Collection in a Probability-Based Internet Panel: How the LISS Panel Was Built and How It Can Be Used." *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 109 (1): 56–61.
- Shin, D., and Y. J. Park. 2019. "Role of Fairness, Accountability, and Transparency in Algorithmic Affordance." *Computers in Human Behavior* 98: 277–284.
- Shoemaker, P. J., and A. A. Cohen. 2005. "News around the World: Content." In *Practitioners, and the Public*. New York, NY: Routledge.
- Shoemaker, P. J., and T. P. Vos. 2009. *Gatekeeping Theory*. London: Routledge.
- Strömbäck, J., K. Falasca, and S. Kruike-meier. 2018. "The Mix of Media Use Matters: Investigating the Effects of Individual News Repertoires on Offline and Online Political Participation." *Political Communication* 35 (3): 413–420.
- Sunstein, C. R. 2007. *Republic.com 2.0*. Princeton, NJ: Princeton University Press.
- Tandoc, E. C. 2014. "Journalism is Twerking? How Web Analytics is Changing the Process of Gatekeeping." *New Media & Society* 16 (4): 559–575.
- Taneja, H., J. G. Webster, E. C. Malthouse, and T. B. Ksiazek. 2012. "Media Consumption across Platforms: Identifying User-Defined Repertoires." *New Media & Society* 14 (6): 951–968.
- Tewksbury, D. 2003. "What Do Americans Really Want to Know? Tracking the Behavior of News Readers on the Internet." *Journal of Communication* 53 (4): 694–710.

- Tewksbury, D., A. J. Weaver, and B. D. Maddex. 2001. "Accidentally Informed: Incidental News Exposure on the World Wide Web." *Journalism & Mass Communication Quarterly* 78 (3): 533–554.
- Thorson, K., and C. Wells. 2016. "Curated Flows: A Framework for Mapping Media Exposure in the Digital Age." *Communication Theory* 26 (3): 309–328.
- Thurman, N., J. Moeller, N. Helberger, and D. Trilling. 2019. "My Friends, Editors, Algorithms, and I: Examining Audience Attitudes to News Selection." *Digital Journalism* 7 (4): 447–423.
- Trilling, D., P. Tolochko, and B. Burscher. 2017. "From Newsworthiness to Shareworthiness: How to Predict News Sharing Based on Article Characteristics." *Journalism & Mass Communication Quarterly* 94 (1): 38–60.
- Vermeer, S. 2018. A Supervised Machine Learning Method to Classify Dutch-Language News Items.
- Vermeer, S., and D. Trilling. 2020. "Toward a Better Understanding of News User Journeys: A Markov Chain Approach." *Journalism Studies* 21 (7): 879–894.
- Wallace, J. 2018. "Modelling Contemporary Gatekeeping: The Rise of Individuals, Algorithms and Platforms in Digital News Dissemination." *Digital Journalism* 6 (3): 274–293.
- Welbers, K., W. van Atteveldt, J. Kleinnijenhuis, and N. Ruigrok. 2018. "A Gatekeeper among Gatekeepers: News Agency Influence in Print and Online Newspapers in The Netherlands." *Journalism Studies* 19 (3): 315–333.
- Welbers, K., and M. Opgenhaffen. 2018. "Social Media Gatekeeping: An Analysis of the Gatekeeping Influence of Newspapers' Public Facebook Pages." *New Media & Society* 20 (12): 4728–4747.
- Wonneberger, A., K. Schoenbach, and L. van Meurs. 2012. "Staying Tuned: TV News Audiences in The Netherlands 1988-2010." *Journal of Broadcasting & Electronic Media* 56 (1): 55–74.
- Zuiderveen Borgesius, F. J., D. Trilling, J. Möller, B. Bodó, C. H. de Vreese, and N. Helberger. 2016. "Should we Worry about Filter Bubbles?" *Internet Policy Review* 5 (1): 1–16.

Appendix

List of Websites

Table 1 presents a list of all domain names included in our sample.

Table 1. List of white-listed websites.

Type of website	Domain names
Tabloids	www.telegraaf.nl (also telegraaf.tcdn.nl , m.telegraaf.nl , krant.telegraaf.nl , content.tmgvideo.nl , vaarkrant.telegraaf.nl , beta.telegraaf.nl), www.ad.nl (also video.ad.nl , krant.ad.nl , social.ad.nl , nieuw.ad.nl) (12)
Broadsheets	www.volkskrant.nl (also krant.volkskrant.nl , ims.volkskrant.nl , verhalen.volkskrant.nl , inclusief.volkskrant.nl), www.nrc.nl (also retro.nrc.nl , digitaleeditie.nrc.nl), www.trouw.nl (also krant.trouw.nl , exclusief.trouw.nl , beta.trouw.nl) (12)
Online-only outlets	www.nu.nl (also media.nu.nl , secure.nu.nl , link.nu.nl), www.nieuws.nl (5)
International outlets	www.lemonde.fr , www.mashable.com , www.reuters.com , www.faz.net , www.livejournal.com , www.ft.com , www.cnn.com , www.huffingtonpost.com , www.dailymail.co.uk , www.standaard.be , www.telegraph.co.uk , www.nytimes.com , www.hbvl.be , www.tijd.be , www.diply.com , www.gva.be , www.demorgen.be , www.theguardian.com , www.hln.be , www.nieuwsblad.be , www.forbes.com , www.washingtonpost.com , www.bbc.com , www.deredactie.be , www.spiegel.de (25)
Broadcasters	www.nos.nl (also s.nos.nl , lab.nos.nl , teletekst-data.nos.nl , feeds.nos.nl , teletekst.nos.nl , scroll.lab.nos.nl , nos.nl), www.vara.nl , www.vpro.nl , www.nporadio4.nl , www.sbs6.nl , www.rtlnieuws.nl , www.nporadio5.nl , nporadio2.nl , www.radartv.nl , www.npo3.nl , www.nporadio2.nl , www.nporadio1.nl , www.rtlxl.nl , www.npo.nl (20)
Other news outlets	www.businessinsider.com , www.tweakers.net , www.nujij.nl , www.speld.nl , www.sargasso.nl , www.destentor.nl , www.vn.nl , www.krant.nl , www.jalta.nl , www.lc.nl , www.viralmundo.nl , www.hpdetijd.nl , www.vi.nl , www.geenstijl.nl (also gigapica.geenstijl.nl), www.vice.com , www.dumpert.nl , www.dagelijksstandaard.nl , www.metronieuws.nl , www.dichtbij.nl , www.joop.nl , www.fok.nl , www.tubantia.nl , www.voetbalzone.nl , www.emerce.nl , www.versbeton.nl , www.nd.nl , www.parool.nl (also meer.parool.nl , s.parool.nl , krant.parool.nl , parool.nl), www.haarlemsdagblad.nl , www.ijmuidercourant.nl , www.leidschdagblad.nl , www.noordhollandsdagblad.nl , www.gooieneemlander.nl , www.dvhn.nl , www.barneveldsekrant.nl , www.decorrespondent.nl , www.groene.nl , www.fd.nl , www.omroepbrabant.nl , www.rtvnh.nl , www.rtvutrecht.nl , www.bd.nl , www.bndestem.nl , www.gelderlander.nl , www.ed.nl , www.frieschdagblad.nl , www.powned.tv , www.bnr.nl , www.omroepwest.nl , www.omroepzeeland.nl , www.omroepfryslan.nl , www.rtvooost.nl , www.at5.nl (also at5.nl), www.omroepflevoland.nl , www.elsevier.nl , www.rtvnoord.nl , www.refdag.nl , www.l1.nl (also l1.nl), www.rijnmond.nl , www.ld.nl , www.pzc.nl , www.omroepgelderland.nl , www.rtdvrenthe.nl , www.stand.nl , www.fondsnieuws.nl , tpo.nl (also politiek.tpo.nl , nieuws.tpo.nl , regio.tpo.nl , cult.tpo.nl , media.tpo.nl , biz.tpo.nl), www.vkmag.com (also vkmag.com), www.regio15.nl , 925.nl , www.n1.nl , www.studio040.nl , www.racingnews365.nl , www.regio8.nl , www.pc55.nl , www.breda076.nl , www.regio14.nl , www.sport1.nl , sliedrecht24.nl , www.heeze24.nl (93)
Search engines	www.google.nl , www.yahoo.com , www.google.news , news.google.com , news.google.nl (5)
Social media	www.facebook.com (also l.facebook.com), www.twitter.com (3)

Performance Measures

In our study, we employed the Passive-Aggressive (PA) learning algorithm, which is known to perform well in various text classification tasks (Crammer et al. 2006), including Dutch-language news items (see e.g. Burscher, Vliegenthart, and De Vreese 2015).

Before training each classifier, we converted the text to a bag-of-words model and used this as the input for the model. Different pre-processing steps have been used resulting in three

Table 2. Performance measures and hyperparameters for the PA algorithm.

	All words	Stopwords	Lead
<i>Performance measures</i>			
Accuracy	.82	.82	.81
Precision	.82	.82	.81
Recall	.83	.82	.82
<i>Hyperparameters</i>			
CountVectorizer()	1,1	1,1	1,1
TfidfTransformer()	True	True	True
Optimization Iteration	15.0	5.0	10.0
Hinge-Loss function	L2-loss	L2-loss	L2-loss

Table 3. Performance measures for the PA algorithm per topic.

	Accuracy	Precision	Recall
<i>Politics</i>			
All words	.85	.85	.85
Stopword removal	.82	.85	.79
Lead	.81	.84	.79
<i>Business</i>			
All words	.67	.69	.65
Stopword removal	.66	.72	.61
Lead	.66	.71	.61
<i>Entertainment</i>			
All words	.89	.86	.92
Stopword removal	.89	.85	.93
Lead	.89	.85	.92
<i>Other</i>			
All words	.57	.73	.47
Stopword removal	.58	.67	.51
Lead	.55	.64	.47
<i>N/A</i>			
All words	.88	.88	.88
Stopword removal	.88	.83	.94
Lead	.89	.82	.97

different text categories. (1) The first category comprises the original text of the news item. (2) Next, we removed Dutch stop words such as articles (e.g. *the, a* and *an*), personal pronouns (e.g. *I, me* and *he*), coordinating conjunctions (e.g. *for, but* and *so*), and prepositions (e.g. *in, towards* and *before*). We retrieved the list of stop words from the Python NLTK package (see Bird and Loper 2016). (3) Finally, after removing stopwords, we examined the lead (i.e. first 75 words) of each news item, as facts are generally presented in descending order of importance (Pöttker 2003).

We split the dataset into a training set (80%; $N=2,963$) and a test set (20%; $N=738$), and used the former for training and the latter for testing. Additionally, we tested various combinations of hyperparameters to find the ultimate combination, for example determining: how to convert a collection of text documents to a matrix of token counts (CountVectorizer), whether to transform a count matrix to a normalized tf or tf-idf representation (TfidfTransformer), and the maximum number of passes over the training data (i.e. epochs). Table 2 presents the ultimate combination of hyperparameters to tune the classifier.

As the PA algorithm examining all words computes a high accuracy for every topic (see Table 3), we used this algorithm to classify news items in our sample (see Vermeer 2018).

Markov Chains

To retrieve a closer examination of the Markov chains, Figure 6 presents the probability distribution for every possible transition in terms of context. Each subplot shows the probability distribution for each possible combination of current (rows) and follow-up (columns) states. These

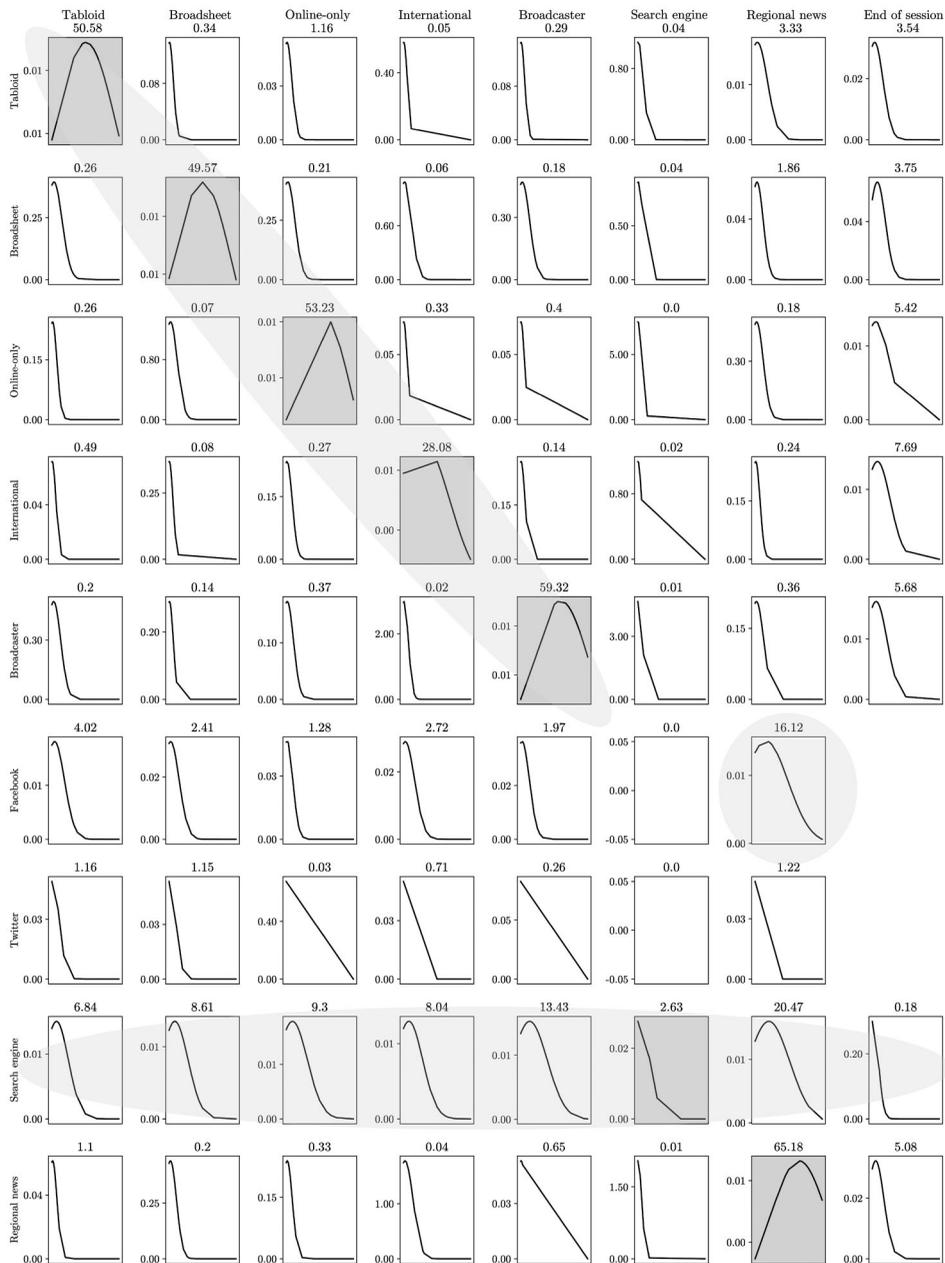


Figure 6. Distribution Markov Chains in terms of Context.

Note. (i) The numbers above every subplot present the transition probabilities; (ii) As Facebook, Twitter, and Other websites are merely taken into account as an entry-point they are not presented as a column; (iii) As Facebook, Twitter, and Other websites are merely taken into account as an entry-point it is impossible to transition from these outlets to the end of a Web session; hence there are no subplots for these outlets in the far right column; (iv) The subplot is empty when transitions between two states do not occur; (v) Highlighted subplots are states that transition back into themselves; (vi) Circled (groups of) subplots provide the most relevant insights and are discussed in more detail in the main text.

probabilities were calculated from a total of 1,175,022 visits. For instance, search engines were on average followed 7% of the time by tabloid visits, 9% by broadsheet visits, and 9% by online-only outlets.

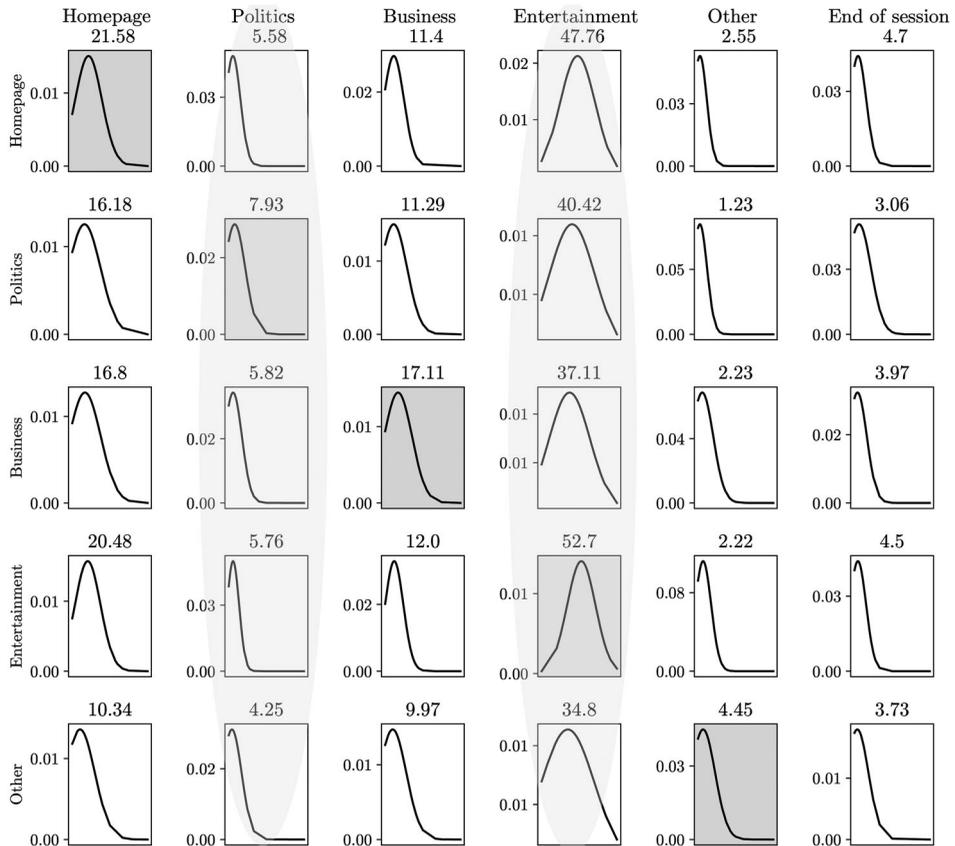


Figure 7. Distribution Markov Chains in terms of Content.

Note. (i) The numbers above every subplot present the mean transition probabilities; (ii) Homepage can also refer to section pages; providing an overview of the news for a specific topic; (iii) Highlighted subplots are states that transition back into themselves; (iv) Circled (groups of) subplots provide the most relevant insights and are discussed in more detail in the main text.

Figure 7 presents the probability distribution for every possible transition in terms of content. Again, each subplot shows the probability distribution for each possible combination of current (rows) and follow-up (columns) states.