We are what we click

Understanding time and content-based habits of online news readers

Makhortykh, Mykola; Helberger, Natali; Harambam, Jaron; Bountouridis, Dimitrios

DOI
10.1177/1461444820933221

Publication date
2020

Document Version
Final published version

Published in
New Media and Society

Article

new media & society
2021, Vol. 23(9) 2773–2800

DOI: 10.1177/1461444820933221
journals.sagepub.com/home/nms


We are what we click: Understanding time and content-based habits of online news readers

Mykola Makhortykh
University of Bern, Switzerland

Claes de Vreese and Natali Helberger
University of Amsterdam, The Netherlands

Jaron Harambam
University of Amsterdam, Amsterdam, The Netherlands

Dimitrios Bountouridis
Delft University of Technology, The Netherlands

Abstract

The article contributes both conceptually and methodologically to the study of online news consumption by introducing new approaches to measuring user information behaviour and proposing a typology of users based on their click behaviour. Using as a case study two online outlets of large national newspapers, it employs computational approaches to detect patterns in time- and content-based user interactions with news content based on clickstream data. The analysis of interactions detects several distinct timelines of news consumption and scrutinises how users switch between news topics during reading sessions. Using clustering analysis, the article then identifies several types of news readers (e.g. samplers, gourmets) and examines their news diets. The results point out the limited variation in topical composition of the news diets between

Corresponding author:
Mykola Makhortykh, Institute of Communication and Media Studies, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland.
Email: makhortykhn@yahoo.com

different types of readers and the tendency of these diets to align with the news supply patterns (i.e. the average distribution of topics covered by the outlet).

Keywords
Clickstream, digital news, information behaviour, legacy media, news consumption, news diets

Introduction

The increasing adoption of digital technologies by legacy media has significant implications for news dissemination and consumption. The formation of a high-choice information environment (Van Aelst et al., 2017) challenges users with the unprecedented amount of news content, whereas the rise of mobile devices enables them to consume news at a different pace and in different contexts compared with predigital times (Westlund and Färdigh, 2015). These factors fundamentally transform the media–audience relationship, yet their impact on user information behaviour online and the long-term societal consequences remain unclear. Using computational methods, this article contributes an observation of online reading habits of legacy media users and introduces new ways of studying them based on clickstream data.

Until now, empirical studies of online news consumption remain relatively limited in number and usually rely on self-reported or small-scale experimental data (see, for review, Mitchelstein and Boczkowski, 2010). While important, these studies provide limited possibilities for identifying the impact of digitalisation on news reading habits as participants are often unable to recognise their consumption patterns (Möller et al., 2020). More large-scale approaches are required to assess how legacy media users consume news online and in which ways digital environments change their reading habits, in particular considering the ongoing debate on the societal effects of the 'algorithmic' (Anderson, 2013) turn in news distribution.[1] Without this knowledge, it is hardly possible to evaluate the impact of more targeted ways of news distribution on how users inform themselves and assess if algorithms enclose users in 'echo chambers' (Sunstein, 2017) and facilitate 'masked censorship' (Makhortykh and Bastian, 2020) or diversify their information diets (Eskens et al., 2017) and enable more control over their information diets (Harambam et al., 2018).

Besides limited possibilities for identifying generalisable patterns of online news consumption, the current scholarship (Gil De Zúñiga et al., 2014; Möller et al., 2020; Trilling and Schoenbach, 2013) often focuses on the effects of personal characteristics (such as gender, age or race) of users on online news consumption. While the importance of these characteristics can hardly be questioned, in particular considering their prominent role in discussions about online consumption and selective news exposure, the role of other factors such as time- and content-based reading habits remains under-investigated. Yet, these factors are of particular importance for two reasons: first, they are used as major elements of user behaviour models used by algorithmic systems of news distribution (Karimi et al., 2018; Möller et al., 2018). Second, unlike personal characteristics,

time- and content-based reading habits are particularly susceptible to the change because of algorithmic curation of users' information diets.

To address this gap, we extend the few earlier studies (Epure et al., 2017; Esiyok et al., 2014) by employing more computational approaches for studying news readers' habits. To do so, we use a large set of data on user click behaviour during their interactions with news content coming from two major legacy newspapers. Using a combination of clustering (TBCA and K-means) and stochastic modelling (Markov processes) techniques, we analyse these interactions to identify and compare time- and content-based patterns of news consumption between the users of the two newspapers. While doing so, we ask what types of users can be identified based on their click behaviour and discuss how these findings can advance our understanding of user information diets. In addition to the conceptual contribution to the field of online news consumption, our observations can be used for improving existing models of users' interactions with news stories (e.g. for developing recommender systems) and evaluating the impact of algorithmic systems on information diets.

Literature review

Online news consumption and legacy media

The adoption of digital technologies has led to significant changes in the relationship between mass media and news consumers. Purcell et al. (2010) note three major changes brought about by the growing consumption of news online: (1) users are able to choose when and where to consume news, (2) news offers are increasingly personalised and (3) consumption mode switches from a passive to an active one. The proliferation of new channels for news consumption, in particular social media platforms, has been discussed intensively by scholars (Boczkowski et al., 2018; Gil De Zúñiga et al., 2014; Hermida et al., 2012). In our article, however, we focus on the legacy media and discuss the implications of digitalisation for information behaviour of their users.

A large number of studies discuss the impact of online news consumption on information diets and how it is influenced by individual characteristics of news readers, varying from age and gender to political interest and education level (Gil De Zúñiga et al., 2014; Möller et al., 2020; Ohlsson et al., 2017; Trilling and Schoenbach, 2013). Many of these studies examine the relationship between online news consumption and audience fragmentation as well as selective exposure and explore to what degree these phenomena can be affected by user characteristics (Helberger and Wojcieszak, 2018; Stroud, 2010). A related strand of research looks at the effects of algorithmic distribution systems on user information diets and discusses what is the role of individual characteristics in this process (Bodó et al., 2019; Möller et al., 2016).

At the same time, time- and content-based patterns of information behaviour remain under-studied despite them being essential for the above-mentioned debates. Time-based reading habits influence the spread of information within the society and determine the amount of information received by the users from the news (Dunaway et al., 2018; Molyneux, 2019). Similarly, content-based reading habits define the composition of individual news repertoires and impact user engagement with the public sphere by

informing (or not informing) them about societal matters (Kim, 2016; Molyneux, 2019; Taneja et al., 2012).

Besides the above-mentioned reasons, time- and content-based reading habits are essential for implementation and optimisation of algorithmic distribution systems for news content. Content features and their attractiveness to the users serve as a basis for the most common implementations of news recommender systems, namely user- and content-based collaborative filtering approaches (Karimi et al., 2018; Möller et al., 2018). Similarly, the integration of time-based consumption patterns is viewed as an important condition for improving the quality of news recommendations by accounting for contextual information related to news consumption (Lommatzsch et al., 2017).

Time-based consumption habits

The transition towards online news consumption and the distribution of mobile devices led to significant changes in the ways users receive news on a daily basis (Westlund and Färdigh, 2015). Increasing mobility allows users to engage with news updates more frequently and leads to diversification of reading behaviour compared with predigital spatial and social configurations of news consumption (Van Damme et al., 2015). Instead of consuming news during fixed time slots (e.g. in the morning and in the evening), online consumption enables more spontaneous interactions with news as users expect to have access to news content throughout the day and independently of their physical location (Dimmick et al., 2011).

A number of studies (Dimmick et al., 2011; Van Damme et al., 2015) discuss the emergence of new time-based reading habits among users interacting with legacy media online. For instance, Schrøder (2015) demonstrates that the proportion of different media consumed depends on the location of the user (e.g. at home, at work, or commuting) and, thus, varies significantly throughout the day. Van Damme et al. (2015) shows that there is a dependency between time of news consumption and the device used for this purpose as well as the type of news consumed.

While some of the above-mentioned studies (Dimmick et al., 2011; Van Damme et al., 2015) look at the concrete distribution of news consumption throughout the day, they usually deal only with aggregated data about group consumption. A few studies that discuss time-based reading habits on the individual level do so via small samples of users providing self-reported data (Incollingo, 2018; Yadamsuren and Erdelez, 2011). By contrast, the question of a large-scale assessment of individual time-based reading habits and how these habits vary between different news outlets remains under-studied and leads us to our first research question.

- RQ1: What kind of time-based news consumption habits can be identified based on clickstream data?

Content-based consumption habits

Another aspect of news consumption affected by digitalisation of news media is related to the changes in the composition of information diets. The increased number of

available media channels enables more choices in terms of the format and content of news, including a greater supply of information associated with niche or partisan views (Van Aelst et al., 2017). Such a change in news supply raises concerns related to the potential fragmentation of the public sphere and the subsequent ideological segregation enhanced by the algorithmic systems of content distribution (Pariser, 2011). Yet, existing studies (Flaxman et al., 2016; Möller et al., 2016) suggest that users still consume predominantly mainstream content online and that possible shrinking of the common core does not necessarily threaten the public sphere.

The question of the changing composition of news diets and how it is impacted by digitisation, however, remains rather important. A number of recent studies (Kim, 2016; Taneja et al., 2012; Van Damme et al., 2015) look at the online user-defined news repertoires and discuss their impact on news distribution. Taneja et al. (2012) demonstrate that the growing supply of media content leads to the formation of distinct repertoires determined by the availability of specific media to users at a given time (e.g. more consumption of web news via desktop devices outside work). Kim (2016) shows that different media repertoires are indicative of different groups of media users and this leads to the significant variations in the news content consumed by these groups.

At the same time, the majority of existing studies focus on cross-media repertoires and little is known about different news repertoires within the same outlet. Esiyok et al. (2014), for instance, demonstrate that different categories of news have a strong effect on the user clicking behaviour that suggests the presence of content-based reading habits. Similarly, Epure et al. (2017) show how user preferences towards specific forms of thematic content provided by the single news outlet change over time. None of these studies, however, look at the content-based reading habits in the comparative perspective, thus raising the question of how similar/different are these habits between the users of different news outlets and leading us to the second research question:

- RQ2: What kind of content-based consumption habits can be identified?

Time- and content-based typologies of news consumers

A growing number of academic studies look at the possible typologies = of online news consumers based on their reading habits. Tewksbury et al. (2008) propose the differentiation between the two types of users: (1) selectors (focused primarily on a specific topic) and (2) browsers (sampling across different topics). Van Damme et al. (2015) suggest three categories: (1) omnivores (with intense news diets relying on multiple news sources), (2) traditionals (relying on traditional news formats and sources) and (3) serendips (rarely engaging with news routinely, but mostly checking for updates). Bos et al. (2016) identify four types of media use profiles: minimalists (rarely interacting with news), public news consumers (actively consuming public news broadcasts), popular news consumers (actively consuming news via commercial channels) and omnivores (frequently engaging with public broadcasts/online news media).

These and other typologies, however, tend to focus on content-based reading habits, in particular cross-media ones. Little has been done to integrate observations on content- and time-based habits, even while several studies suggest a possible relationship between

them. Dimmick et al. (2011) show that during certain time slots users rely on different devices for news consumption that leads to various ratios of specific types of news content consumed. Similarly, Van Damme et al. (2015) trace differences in consumption of hard, soft and service news throughout the day that can be attributed both to external factors and user preferences. Yet, the above-mentioned studies usually rely on a small selection of general content categories (e.g. general news and sport news). Thus, a typology acknowledging both content- and time-based reading habits is important for articulating differences in user information behaviour in relation to news and examining the consequences of these habits for information diets. This leads us to our last research question:

- RQ3: What types of users can be identified based on time- and topic-based interactions with news content?

Methodology

Data acquisition

For implementing our study, we acquired data from Persroep, a Belgian publishing company owning news organisations in Belgium, Denmark, and the Netherlands. Persgroep provided us with clickstream data generated from 1 June to 31 August 2018 by users who accessed online versions of two Dutch legacy newspapers: *Trouw* and *AD*. Both newspapers are distributed in printed and digital formats and constitute important elements of the news ecosystem in the Netherlands.

Trouw (n.d.) is a daily quality newspaper rooted in Protestant tradition and pays 'particular attention to democracy, sustainability and all forms of religion, philosophy and philosophy of life'. *AD* is a daily tabloid newspaper and one of the two largest Dutch newspapers in terms of subscribers together with *Telegraaf*. Unlike *Trouw, AD* has a strong regional focus and provides readers with a number of regional supplements which are also available online. In addition to the nation-wide edition of *AD*, there are 59 regional supplements each collecting local news from a particular region of the Netherlands.[2] These regional supplements and user interactions with them were also included in our dataset. While there is some variation in the content published by the regional outlets (for instance, for the period we obtained information about, the Amsterdam supplement put more emphasis on politics- and crime-related news, whereas the Rotterdam supplement published more on economics, crime and culture, and the Hague supplement published more sport stories than the other two supplements), for the majority of outlets the distribution of the top topics followed the aggregated pattern for *AD* described in Table 1. For future research we consider examining the potential differences in user interactions with news content between regional supplements, but for now we just treated them as part of the *AD* corpus (hence, its larger size compared with that for *Trouw*).

The distinct feature of clickstream data is that these data capture all readers' interactions – that is, clicks – with the newspapers' websites. Because of this, it allows tracing how users interact with specific news items over time and identify patterns in their interactions. Unlike earlier studies which usually rely on clickstream data acquired for a short period of

Table 1. Corpus composition for *AD* and *Trouw* (articles).

Topic	No. of articles (*AD*)	No. of articles (*Trouw*)
Accidents	6% (7942)	1% (108)
Crime	10% (13,838)	4% (501)
Culture	12% (15,505)	18% (2044)
Economics	18% (23,672)	15% (1622)
Education	2% (2522)	2% (238)
Environment	2% (2055)	3% (316)
Health	2% (2553)	4% (400)
Politics	8% (10,291)	20% (2216)
Religion	0.5% (701)	4% (420)
Science	0.5% (660)	2% (220)
Sport	16% (20,329)	8% (888)
Society	22% (28,259)	17% (1901)
War	0.5% (548)	2% (241)
Weather	0.5% (655)	0% (18)
Total no. of articles	100% (129,530)	100% (11,115)

time and only from a single newspaper, we acquired data for 3 months and from two digital outlets. It allowed us to compare online information behaviour between *AD* and *Trouw* users and examine similarities and differences in how users consume news online.

Clickstream data provided to us by Persgroep included user interactions with two news websites via desktop/mobile browsers and mobile news applications. The process of data collection was consistent between the two newspapers. The data included four fields: a timestamp (i.e. when the click was made), a user id (i.e. a unique id automatically generated by the system for each user based on cookies),[3] a clicked item id (i.e. a unique id of each news story) and a brand (i.e. to which newspaper the clicked item belonged). The data were collected for all users who visited the online outlets during the period of study, including both registered and non-registered users. No further details about users (e.g. their demographic data) or the type of devices used were available. While these limitations do not allow us to relate our findings to specific demographic groups, it is still sufficient for examining time and content-based reading behaviours, which is the major focus of our study.

In addition to clickstream data, we were provided data about news stories published by *Trouw* and *AD* from 1 January until 31 August 2018.[4] Besides article texts, we acquired associated metadata including the text of the article, the author(s) name and the thematic tags (e.g. 'war', 'politics', 'football') automatically generated by the automated classification system used to identify the article's subject by Persgroep. Both *Trouw* and *AD* used the same classification method and the same set of categories based on IPTC media topics taxonomy[5] to classify the articles. We used the metadata for dividing articles into thematic categories; before doing it, however, we went through the automatically detected categories in order to group related categories together and decrease the overall number of categories from a 100 to 14.

The decision to decrease the overall number of categories was related to the way the automated enrichment of articles with IPTC categories was implemented by Persgroep. Instead of assigning each article to a single top-level (or core) IPTC media topic,[6] the articles received multiple topic tags related to different levels of IPTC taxonomy (e.g. 'sport' and 'soccer'; also the article could be assigned to several top-level IPTC topics such as 'politics' and 'crime') together with a score indicating the probability that the article belongs to this particular topic.

We assigned each article to the IPTC topic with the highest score – that is, the one classified as the most representative for a particular article. After doing so, however, we ended up with a large number of rather narrow categories (e.g. 'ice hockey' or 'astronomy'), including only a few dozens of articles as contrasted by several metacategories (e.g. 'politics') constituted by thousands of articles and usually related to top-level IPTC media topics. We assumed that such disparity between categories together with their high granularity could have a detrimental effect on our analysis of content-based reading habits.[7]

Hence, we examined all the categories occurring in our sample and manually combined the related categories to produce larger categories. The decisions concerning the relatedness of categories were made based on the discussion between the co-authors. The final set of categories generally followed the top-level categories of IPTC taxonomy with the exception of topics 'lifestyle and leisure', 'human interest', 'labour' and 'society', which were merged into a single category 'society'. The decision to implement such a merge was related to these topics being thematically close and often overlapping (for instance, the same news story might have very similar if not the same scores for 'lifestyle and leisure' and 'human interest' topics or 'society' and 'labour' topics) as well as the scarce presence of some of these categories in our corpora (e.g. labour-related news).

Data analysis

To analyse time- and content-based consumption habits of *AD* and *Trouw* readers, we first used descriptive statistics to examine the key features of the articles' corpus and the clickstream dataset for the two newspapers. Our goal here was to identify how similar or different are the general characteristics of data related to each of the two newspapers in order to avoid possible biases on the later stage of our analysis. An example of such biases is the different rate of supply of specific thematic categories of news stories: for instance, if *AD* publishes a significantly larger number of sport-related news than *Trouw*, then we can expect *AD* users to read more about sport just because the newspaper supplies them with the respective content.

Following the examination of the general features of the datasets, we looked at the time-based patterns of user interactions with news content. We started by examining a general distribution of interactions with news content depending on the time of day and the day of the week. Following this general overview of time-based news interactions, we employed trajectory-based clustering analysis (TBCA) (Genolini et al., 2016) to identify clusters of users based on the usual trajectories of their interaction with news throughout the day. Unlike other clustering approaches, TBCA is adapted for clustering longitudinal data based on the shape of its trajectories and not distances between points.

Thus, we constructed trajectories of clicking behaviour on the hourly basis for each user in our dataset and clustered them using TBCA to examine the resulting clusters.

Following our analysis of time-based interactions, we looked at the content-based consumption patterns. For this purpose, we enriched our clickstream data by constructing reading sessions out of user clicks. In order to do so, we treated all user interactions with news content occurring in less than 15 minutes from each other as part of the same reading session. After constructing reading sessions, we again started with descriptive analysis and looked at the distribution of session lengths among users of *AD* and *Trouw* to see whether there are significant differences between them. We also examined the distribution of news topics between sessions of a minimal length (i.e. of one click) and the longer sessions to check whether very short 'peeking' sessions are characterised by the consumption of certain types of thematic news content.

Following this descriptive analysis, we used a first-order Markov model to estimate transition probabilities between news categories; specifically, we were interested in verifying the earlier claim by Esiyok et al. (2014) about varying degrees of loyalty towards different topics as well as the relationship between topics as a part of user information diets (e.g. if sport news tend to go together with politics news and so on).

Finally, we looked at what kind of user types can be identified based on time- and content-based reading habits. For this purpose, we again enriched our dataset with two additional features: the average length of the reading session for each user and the average number of different news topics consumed during a single session. We then used K-means clustering to identify distinct groups of users based on these two features. The idea of K-means clustering is to divide data points into k groups depending on the mean distance from the other points; because of its simplicity and robustness, this algorithm is frequently used to detect groups in the uncategorised data. To determine k, we calculated squared distance and then used an elbow method. The resulting graphs are provided in the supplementary materials; based on their examination, we decided to use eight clusters for *AD* and *Trouw*.

Findings

General characteristics of the datasets

We started our analysis by examining the general characteristics of two datasets: one describing the corpus of articles and the other on user interactions with the articles. As shown in Table 1 (in the 'Methodology' section), the two newspapers differ significantly in terms of the volume and topical distribution of content produced from 1 January to 31 August 2018. During the 8 months for which we acquired data, *AD* produced on average 533 stories per day, whereas *Trouw* produced on average 44 stories. Such a gap can be attributed to different newspaper profiles, in particular the large number of regional supplements of *AD*. These supplements serve as newspapers in themselves, thus resulting in the significantly higher number of articles published by *AD*.

In addition to the different volume of articles, Table 1 shows that the two newspapers differed in terms of the subjects covered. Compared with *AD, Trouw* published more content on politics, culture and religion. By contrast, *AD* devoted more attention to news

Table 2. Corpus composition for *AD* and *Trouw* (clicks).

No. of clicks	No. of users (*AD*)	No. of users (*Trouw*)
1	32.5% (11,254,377)	48% (2,888,704)
2–5	41.5% (14,364,486)	44% (2,685,946)
6–10	11% (3,691,530)	5% (299,324)
11–50	11% (3,743,017)	3% (164,774)
51–100	2% (714,365)	0% (8489)
101+	2% (810,663)	0% (4733)
Total no. of users	34,578,438	6,051,970

about accidents, economics, sport, crime and society.[8] These distinctions again can be attributed to the different profiles of the two newspapers, in particular of *Trouw* being a quality newspaper with a strong historical focus on religious subjects and *AD* being a tabloid focusing on entertainment and regional news.

After this simple descriptive examination of the articles' corpus, we looked at the data on user interactions with the articles from 1 June to 31 August 2018. As shown in Table 2, the large number of users who accessed online versions of *Trouw* and *AD* made only one click during the 3 months. The second largest category of users was those who made between two and five clicks during the period of study. These observations indicate the extremely high drop rate of potential news readers, many of whom do not return to the news outlet for a long time after checking a single news story. While both newspapers used cookie walls, their presence did not explain the drop rate: in the case of *Trouw*, users were allowed to read seven articles before being prompted to subscribe, whereas *AD* limited access only to premium articles. Such a high dropout has significant implications for the commercial model of digital distribution of news content, but it is even more important in terms of the long-term effects of online consumption for the societal role of legacy media.

At the same time, *AD* had significantly more active users compared with *Trouw* both in absolute and relative numbers. The higher user engagement in the case of *AD* can be attributed to the newspaper's strong regional following; however, it also raises the question of whether higher rates of news consumption lead to higher diversity in terms of content consumed.

In addition, these observations point out the importance of considering news consumption habits in the context of deploying algorithmic news recommenders. On the one hand, the high number of one-time news peekers emphasises the importance of the cold-start problem for news recommendations – that is, the task of making initial content suggestions under the condition of the lack of information of user consumption preferences and with a purpose of keeping the user engaged with the content. It can also be viewed as an indicator of the importance of helping users to deal with information overload and preventing them from being discouraged from news exploration because of it.

Time-based reading habits

Following our examination of the general characteristics of the datasets, we proceeded with the analysis of the time-based reading habits of *AD* and *Trouw* users. We started by

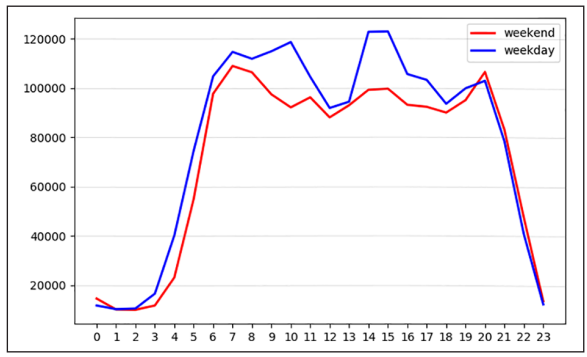

Figure 1. Time-based interactions by the hour and the weekday/weekend for *Trouw*.

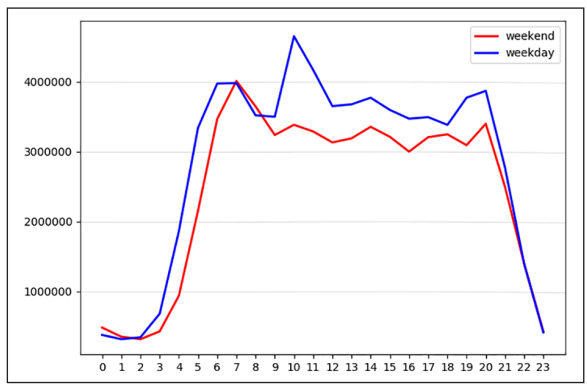

Figure 2. Time-based interactions by the hour and the weekday/weekend for *AD*.

using a descriptive approach, which is commonly used in existing studies on news consumption (Van Damme et al., 2015). Specifically, we looked at the distribution of user clicks on the hourly basis during weekdays and weekends and compared it between *AD* and *Trouw* as shown in Figures 1 and 2.

The comparison shows that for the two newspapers, content consumption during the weekends was rather similar: in both cases, there were two activity peaks (from 7 a.m. to 8 a.m. and from 9 p.m. to 10 p.m.) with a decrease of activity in the afternoon. By contrast, the consumption during the weekdays showed more differences: for *AD*, news consumption peaked from 10 a.m. till 11 a.m. and then dropped, whereas for *Trouw* the second major peak was observed in the afternoon from 14 p.m. to 16 p.m. Such a difference in the weekday reading habits can be attributed to several factors, varying from the various rates of online content publication by *AD* and *Trouw* to different demographics of newspaper users. Specifically, these differences can be related to different social media strategies used by the two newspapers: while both of them use newsletters, *Trouw* tends to send its newsletter in the late afternoon (i.e. around 4 p.m.), whereas the majority

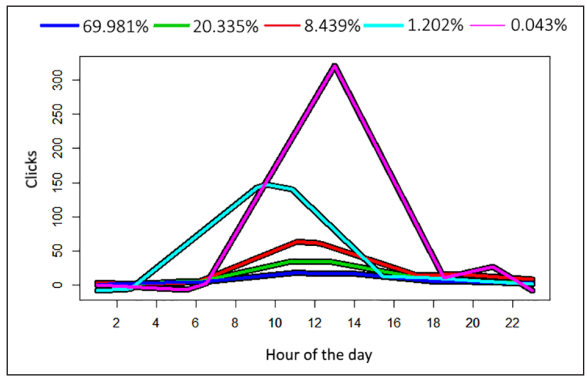

Figure 3. User clusters by time-based reading habits for *Trouw*.

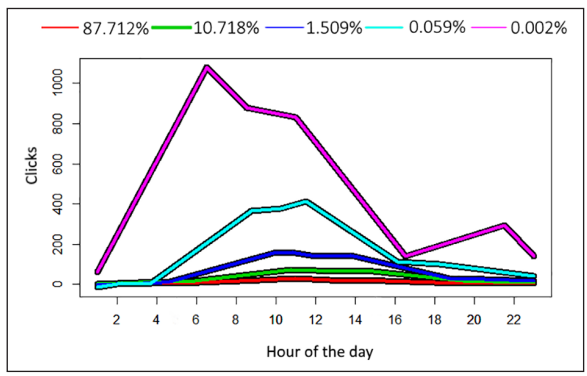

Figure 4. User clusters by time-based reading habits for *AD*.

of *AD's* numerous newsletters are disseminated in the first half of the day (from 7 a.m. to 9 a.m. with some appearing around 1 a.m.).

The above-mentioned differences also emphasise the importance of acknowledging contextual factors when developing automated systems of content distribution. The comparison between *Trouw* and *AD* suggests that users of both newspapers might have different preferences in relation to the time slot when they consume news. Such differences can have implications for their willingness to engage with content recommendations as well as the scope of such engagement (e.g. it can be more beneficial to offer a broader selection of recommended items for *Trouw* users around the lunch break than in the evening).

After examining the general distribution of reading activity by time, we examined different reading trajectories of active readers (100+ clicks) using trajectory-based clustering analysis. The results of the clustering are shown in Figures 3 and 4: in both the cases we observed small groups of power users[9] (i.e. pink and teal clusters), which constituted slightly more than 1% of users for *Trouw* and less than 1% for *AD* users. The shape of these power user clusters, however, was different: in the case of *Trouw*, there

was a strong '9 to 5' reader cluster peaking around noon, whereas its analogue for *AD* was a 'dawn' cluster peaking around 6 a.m.–7 a.m. and then going down throughout the day. It is worth noting that two smaller – that is, teal – power user clusters reproduced the shapes of the larger purple clusters from the other newspaper – for *Trouw*, the teal cluster also had a 'dawn'-like shape, whereas for *AD* the teal cluster was close to the '9 to 5' shape.

In addition to the distinct clusters of power users, both *AD* and *Trouw* included clusters of less active news consumers. In both the newspapers, these clusters represented time-based consumption of the majority of users and had rather similar shapes. With the exception of a more pronounced red cluster for *Trouw* related to the 'morning coffee' readers, both the newspapers shared several low-profile clusters uniting users who consumed news in small sessions distributed throughout the day. This observation supports the earlier suggestions about online news consumption facilitating serendipitous news consumption (Van Damme et al., 2014) or the 'news snacking' (Molyneux, 2018).

Content-based reading habits

After examining time-based reading habits, we moved towards analysing content-based reading habits. To do so, we constructed reading sessions based on the criteria explained in the methodology: for *Trouw*, we constructed 9,059,748 sessions and for *AD* we constructed 203,129,610 sessions. The shortest session for the two newspapers consisted of a single click; the longest session consisted of 521 clicks for *Trouw* and 1653 for *AD*.[10]

The distribution of sessions according to their frequency follows a power law (see the figures in the supplementary materials). Out of the constructed sessions, 6,793,747 (74%) of *Trouw* sessions consisted of a single click. For *AD*, the proportion was slightly different: 111,760,006 (55%) sessions were made of a single click. The distribution of topics between sessions consisting of 1 click and 2+ clicks is shown in the supplementary materials; to calculate it, we divided the number of clicks on articles from a specific thematic category by the overall number of clicks for the respective session type (i.e. 1 click type or 2+ clicks type). The tables show little difference between content consumed during the 1 and 2+ click sessions; however, there was quite some difference between the content published by the news outlets and the content consumed by the users.

In the case of *Trouw*, content related to the human interest such as news related to war, environment and health received more attention. Health in particular attracted more attention from *Trouw* readers: while stories on the topic constituted only 3% of content published during the period of study, they attracted 7% of clicks in 1-click session and 5% of clicks in 2+ sessions. Similar discrepancies were found for *AD*, where content related to weather (0.5% published stories attracted 3% of all clicks) and society (22% published stories attracted 32% and 31% clicks respectively) generated more engagement from the users.

Following examination of the lengths and content composition of constructed news reading sessions, we moved towards examining the transition probabilities between different news categories. We did not include 1-click sessions in the analysis of transitions between topics, because in the case of a single-click session no transition has occurred. Thus, we used information about sessions which consisted of 2 or more clicks (91,369,604 for *AD* and 3,122,916 for *Trouw*).

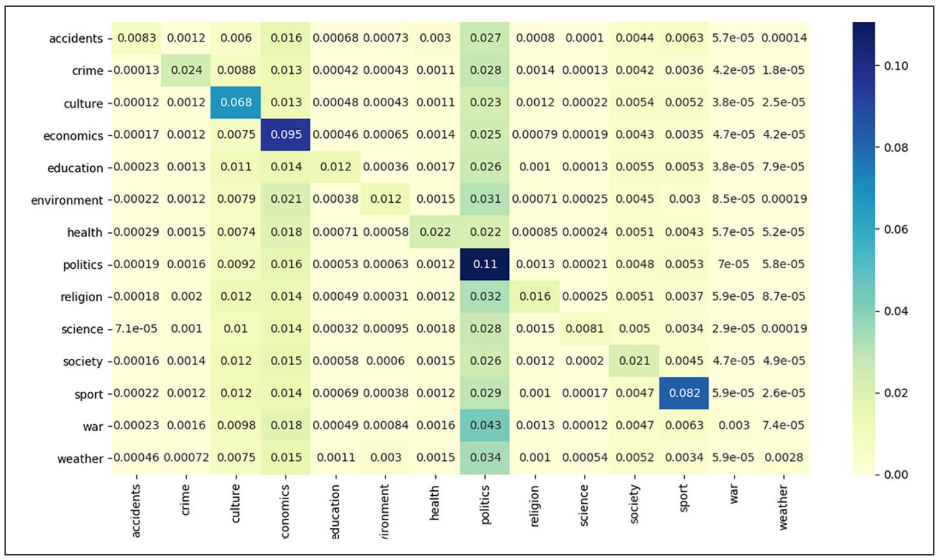

Figure 5. Normalised topic transition probabilities for *Trouw*.

We first looked on non-normalised transition probabilities between topics using first-order Markov chains (i.e. by considering the conditional probability of topic B [vertical row] being clicked directly after topic A [horizontal row] in the same session). Our findings (for visualisations, see supplementary materials) support the earlier observations by Esiyok et al. (2014), who used data for 1 month of observations of user clicks on a major German news portal and found that the user clicking behaviour varies between the thematic categories of news content. The conditional probabilities we observed suggest the significant degree of consistency in user preferences – that is, users often tend to stick to the same topic while reading the news. The same tendency was observed by Esiyok et al. (2014), who labelled it as the topic 'loyalty'.

While examination of the conditional probabilities of topic transitions highlights the significant differences between the two newspapers both in terms of preference consistency and directional relations between topics, these differences are also influenced by the unequal distribution of topical content between different topics in *AD* and *Trouw* and, most importantly, the varying degrees of user interaction with this content.[11] To address these discrepancies, we normalised the observed probability of transitioning into a topic with the expected probability of this topic apprearence based on the sample of sessions made of 2+ clicks.

The comparison provided by Figures 5 and 6 suggests that sport news seems to be particularly 'sticky' among both *AD* and *Trouw* users despite the significant difference in terms of their frequency (4% of clicks for *Trouw* and 16% for *AD*). The other consistently read topics tend to be more newspaper-specific: for instance, *Trouw* users tend to consistently read about politics and economics (and, to a lesser degree, about culture), whereas *AD* users tend to keep reading about society and, albeit less frequently, about crime and

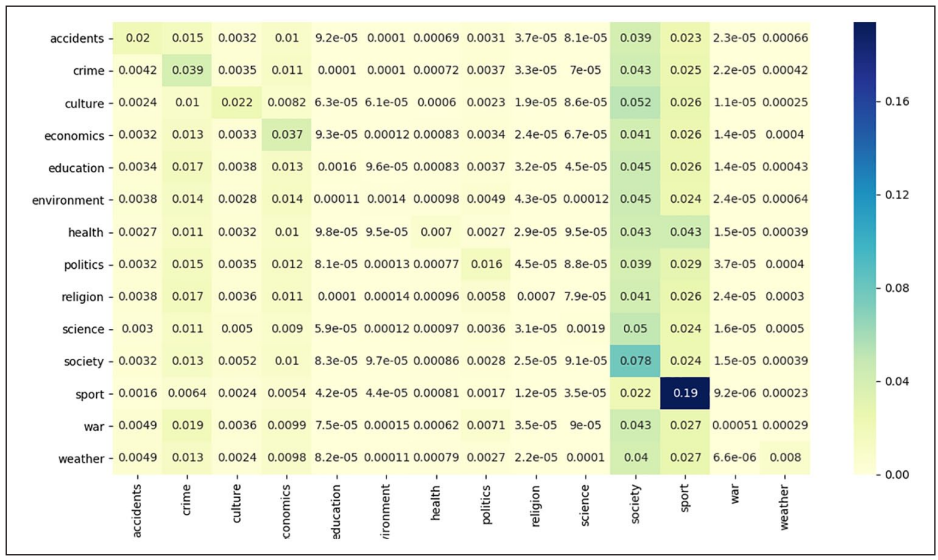

Figure 6. Normalised topic transition probabilities for *AD*.

economics. While to a certain degree, these preferences reflect the expected probabilities of topic occurrence, this pattern did not hold true in all cases. For instance, despite society news composing a large chunk of *Trouw* content (17%) and a common target of 2+ click sessions (19%), the users did not stick to this specific news category.

We also looked on probabilities of transition between specific topics in the course of the reading sessions. Similar to Esiyok et al. (2014), we observed the presence of some 'follow-up' thematic categories – that is, those news categories that were particularly often switched to. For *Trouw* these categories were politics and, to a lesser degree, society and economics news; for *AD* it was primarily society news and, to a lesser degree, sport news. Considering that the majority of these categories were also the most numerous content-wise, these observations emphasise the importance of the content supply that defines the boundary conditions for user interactions with news. At the same time, we also observed some exceptions, in particular related to the society news category in the case of *Trouw*, which was not switched to that often, despite it being rather well-represented via the outlet.

We also observed differences in the probabilities of topic transition between particular topics. Some of these probabilities were similar for both *Trouw* and *AD*: for instance, the tendency to switch to the weather news following the environment news (and to the accident news following the weather news) or to read about politics following stories on religion and war. Other transition probabilities were more newspaper-specific: for instance, *AD* readers interested in sport were less eager to read about crime next compared with other possible transitions that was not the case for *Trouw* readers, whereas *Trouw* readers consuming news about war and religion more commonly switched to crime than readers consuming other types of news first.

Content- and time-based reading habits

After examining content- and time-based reading habits, we looked at how these habits interact and whether we can identify distinct groups of users based on the length and topical variety of a typical news reading session. To achieve this purpose, we calculated the average length of a reading session (in clicks) for each user and the average number of switches between topical categories within a single session. We then used K-means clustering to identify groups of users based on these two parameters.

Figures 7 and 8 show the visualisations of user clusters for *AD* and *Trouw* users. Both the newspapers had clusters of users who consumed news in short sessions (2–4 clicks) on a single topic (#6 for *AD* and #4 for *Trouw*) or with one topic switch (#1 for *AD* and #0 for *Trouw*). We labelled these users as *nibbers*[12] according to their short and focused reading sessions. As shown in Table 3, nibbers were the most frequent type of users and constituted 68% of all users for *AD* and 72% for *Trouw*. Content-wise, single-topic nibbers focused on society (*AD*) and culture/economics news (*Trouw*) and consumed less political content (see Tables 4 and 5[13]). The diet composition of two-topic nibbers was quite similar: in the case of *AD*, the only difference was slightly higher consumption of sport news, whereas for *Trouw* two-topic nibbers consumed more politics- and less culture-related news.

Similarly, for both the newspapers we identified clusters of users who still had short sessions (3–5 clicks), but switched topics rather frequently (2–4 switches per reading session) (#3 for *AD* and #7 for *Trouw*). These users were labelled as *samplers* according to their short, but fluctuant reading sessions. Like nibbers, samplers constituted a rather common user group: 20% users for *AD* and 8% for *Trouw*. In terms of content, *AD* samplers primarily focused on sport and society, whereas *Trouw* samplers preferred reading about politics and society.

Finally, both the newspapers had clusters of users with medium (4–10 clicks) and long sessions (16–256+ clicks) and the high variety of topic switches during a single session (from 0 to 6 and from 0 to 14 switches) (#0 and #4 for *AD* and #2 and #3 for *Trouw*, respectively). Based on the medium-to-high session lengths with the varying degree of topic fluctuations, we labelled these user groups as *buffeteers* and *foodies*. Buffeteers – that is, users with medium-sized sessions and the broad range of topic switches – constituted 8% of users for *AD* and 5% for *Trouw. AD* buffeteers focused on society news and showed slight less interest in sport compared with samplers. In the case of *Trouw*, we observed the same usual focus on politics and society news.

The last cross-newspaper cluster was made of foodies who consumed news via long reading sessions. Some of these sessions were rather focused and included only 1–2 topic switches, whereas other foodies switched up to 14 topics per session. Unlike earlier clusters, foodies were rather rare and constituted less than 0.1% of users for *AD* and approximately 0.1% for *Trouw. AD* foodies consumed the large number of news on sport as well as a slightly higher number of news related to culture and economics compared with earlier clusters. For *Trouw*, politics and society news categories remained prevalent.

The rest of the identified clusters were newspaper-specific ones. *Trouw*, for instance, included two distinct clusters of power users: one with particularly long sessions (256+ clicks) and the high number of topic switches (12–14 per session) (#1) and shorter sessions (64–256 clicks) and frequent topic switches (10–12 per session) (#5). Because of their extensive session lengths, we labelled these users as *gourmets*. Both the gourmet

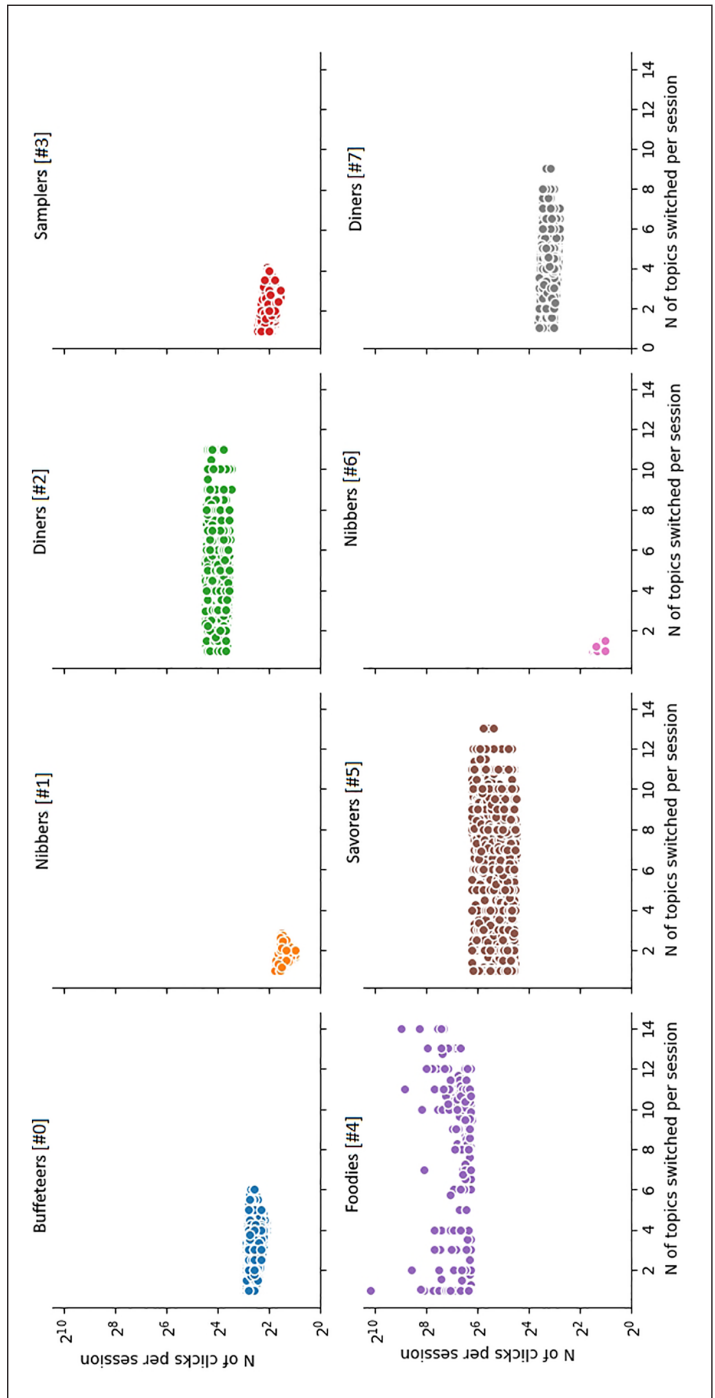

Figure 7. User clusters for *AD*.

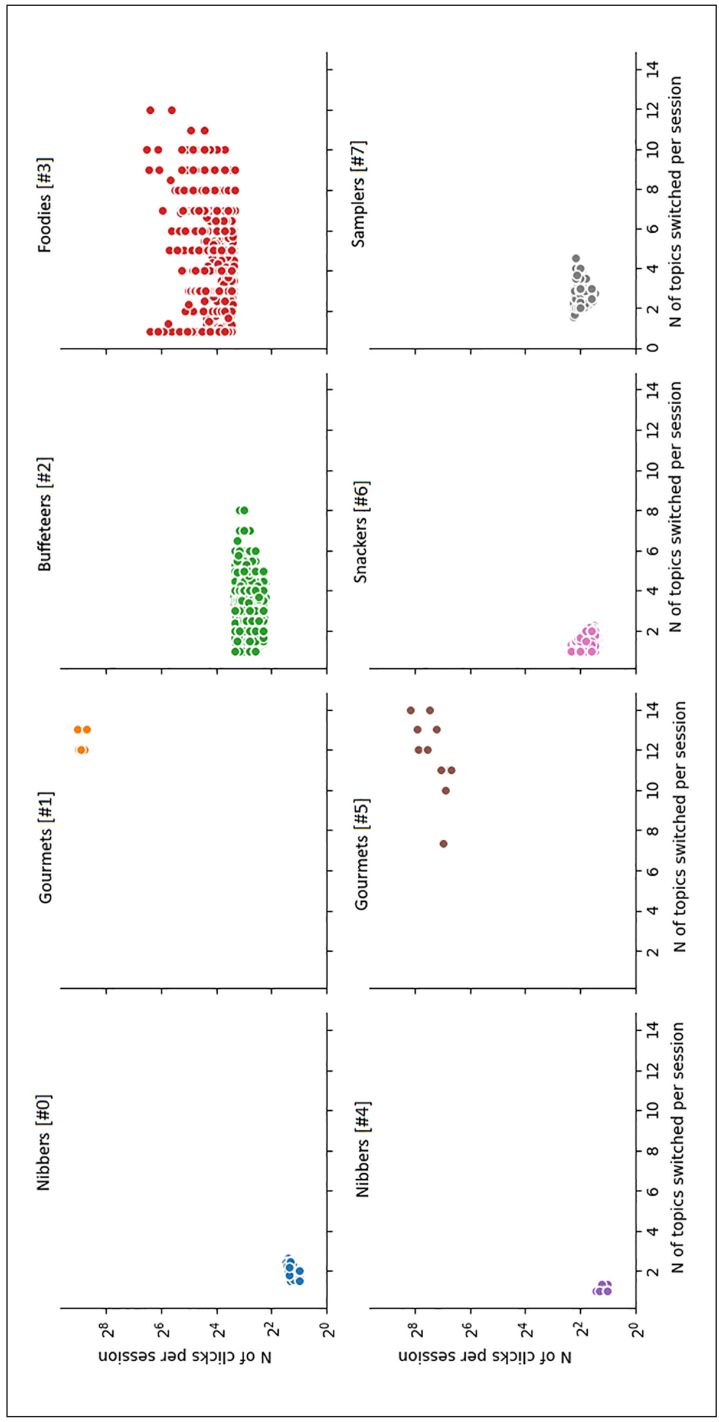

Figure 8. User clusters for *Trouw*.

Table 3. No. of users per cluster for *AD* and *Trouw*.

Cluster	*AD* label	*AD*	*Trouw* label	*Trouw*
0	Buffeteers	8% (1,505,116)	Nibbers	22% (318,073)
1	Nibbers	35% (6,480,702)	Gourmets	0% (9)
2	Diners	0.9% (136,687)	Buffeteers	5%(65,481)
3	Samplers	20% (3,611,781)	Foodies	0.1% (2209)
4	Foodies	0% (260)	Nibbers	50% (737,381)
5	Savorers	0.1% (13,914)	Gourmets	0% (13)
6	Nibbers	33% (6,118,606)	Snackers	14.9% (218,652)
7	Diners	3% (513,051)	Samplers	8% (117,486)

clusters for *Trouw* occurred rather infrequently and composed less than 0.1% of *Trouw* users. Content-wise, gourmets consumed more news on 'high' topics such as culture, economics and religion and read less frequently about weather or crime.

Finally, *Trouw* had a variation of a sampler user type, but the one switching topic less frequently during medium-(3–5 clicks) sized sessions (#6). Because of their slightly longer and more focused sessions, we labelled these users as *snackers*. Snackers were the second-most common *Trouw* cluster after nibbers and constituted 14.9% of all users. Content-wise, snackers were the most active politics readers for *Trouw* and the second-most active consumers of economics-related news.

In the case of *AD*, we also observed two newspaper-specific groups of readers. The first of these groups included two clusters of users (#2 and #7) with medium-sized reading sessions (10–30) and a varying number of topic switches (0–10). Labelled as *diners*, these clusters included 3.9% of *AD* users. The topical composition of diets for the two clusters was similar: diners with longer session lengths (#2) consumed slightly less society-related and more economics-related news, whereas for diners with shorter sessions (#7), society was a more prevalent subject similar to the majority of other *AD* clusters.

The last *AD* cluster (#5) was composed of users with medium- to long-sized sessions (30–70) and varying number of topic switches (0–12). Because of the similarities with the previous cluster (i.e. diners) with a single exception of the longer session lengths, we labelled this user group as *savorers*. Similar to other longer session user groups, savorers were not numerous and constituted around 0.1% of users for *AD*. Society was a prevalent news topic among savorers, seconded by sport news and economics.

Overall, we did not observe significant differences in terms of the aggregate distribution of topics between different groups of users following the same outlet. While there was some variation between certain user groups, these variations tend to concern 1–2 content categories, whereas the rest of the categories usually follow the average distribution within the articles' corpus and, thus, are defined by content supply (with a few exceptions of, for instance, lesser consumption of sport-related news compared with supply for *Trouw* and higher consumption of society-related and lower consumption of culture-/economics-related news compared with supply for *AD*). This observation suggests that variations between reading behaviours do not necessarily lead to profound differences in the composition of information diets (at least on the aggregate level). While more active users acquire volumes of information that is slightly different from the average ones published by the newspaper, these distinctions are minor.

Table 4. Distribution of thematic categories per cluster for *Trouw*.

Topic	Nibbers	Gourmets	Buffeteer	Foodies	Nibbers	Gourmets	Snackers	Samplers	Corpus
Accidents	2	0.5	1	2	2	1	1	2	1
Crime	5	2	5	4	5	3	6	5	4
Culture	**12**	19	**12**	**15**	16	18	**12**	**12**	18
Economics	16	**12**	16	16	**20**	**12**	17	14	15
Education	3	1	3	2.5	3	1	3	3	2
Environment	4	3	3	3	4	5	3	3	3
Health	5	3	5	4	5	3	5	5	4
Politics	22	21	**23**	22	18	**23**	**24**	**23**	20
Religion	5	**8**	5	5	3	**7.9**	4	5	4
Science	2	2.5	2	2	1	2	1	2	2
Sport	**3**	9	**3**	**4**	**3**	**3**	**3**	**4**	8
Society	19	18	**20**	19	19	**20**	19	**20**	17
War	1	1	1	1	0.5	1	1	1	2
Weather	1	0	1	0.5	0.5	0.1	1	1	0

All figures are in percentages. Bold numbers highlight categories, where consumption is different from the supply for more than 2%. "Corpus" column refers to distribution of topics in the corpus of articles published by the newspaper from January 1 to August 31 2018.

Table 5. Distribution of thematic categories per cluster for *AD*.

Topic	Buffeteers	Nibbers	Diners	Samplers	Foodies	Savorers	Nibbers	Diners	Corpus
Accidents	6	7	6	7	5	5	7	6	6
Crime	**14**	**13**	**13**	**13**	10	12	12	**14**	10
Culture	**7**	**7**	**7**	**7**	**8**	**8**	**7.5**	**7**	12
Economics	**12**	**11**	**13**	**11**	**13**	**13**	**11**	**12**	18
Education	1	1	1	1	1	1	1	1	2
Environment	1	1	1	1	1	1	1	1	2
Health	3	3	3	3	2	3	3	3	2
Politics	7	6	7	6	6	7	**5**	7	8
Religion	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Science	1	1	1	1	1	1	1	1	0.5
Sport	15	16	16	17	**28**	**20**	14.9	15	16
Society	**30**	**31**	**29**	**31**	23	**27**	**33**	**30**	22
War	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.5	0.5
Weather	2	**3**	2	2	1	1	**3**	2	0.5

All figures are in percentages. Bold numbers highlight categories, where consumption is different from the supply for more than 2%. "Corpus" column refers to distribution of topics in the corpus of articles published by the newspaper from January 1 to August 31 2018.

Discussion

Our analysis points out a number of distinct time- and content-based patterns of online news consumption among legacy media users. We found that time-based consumption habits seem to be rather similar between the two newspapers we examined. While there are some differences between *AD* and *Trouw* users in terms of when they consume content, these differences are mostly present only during weekdays and can be (at least partially) attributed to different social media strategies used by the newspapers, whereas news consumption of weekends follows the same pattern. In terms of individual consumption habits, we found that differences mostly concern power users ('9 to 5' readers for *Trouw* and 'dawn' readers for *AD*). For both the newspapers, power users constitute the minority, whereas the majority of users tend to engage in short-time reading sessions throughout the day, thus supporting earlier arguments about online media stimulating spontaneous news consumption.

By contrast, content-based consumption patterns turn to be more different between *AD* and *Trouw*. For both the newspapers, we observed a tendency among readers to stick to a single topic while reading news that is similar to earlier observations by Esiyok et al. (2014) and Epure et al. (2017). In both the cases, sport news was among the most 'sticky' type of content, whereas other consistently read topics varied between newspapers (politics/economics/culture for *Trouw* and society for *AD*). Often, these 'sticky' topics were the ones that were the most extensively covered by the outlet, but we also observed some exceptions from this case (e.g. society news in the case of *Trouw*). We also noted differences in terms of transition probabilities between individual topics as well as which topics users switched to more frequently. Together, these distinctions suggest that unlike time-based reading habits, content-based ones are more specific for certain news outlets.

Finally, we looked at what types of users can be identified based on time- and topic-based interactions with news content. Using clustering analysis, we determined several groups of users based on how long they tend to consume news and how frequently they switch between news categories in the course of a single news session. Some of these groups (e.g. nibbers, samplers and buffeteers) were present among the readers of both newspapers, whereas others (e.g. gourmets or savorers) were specific for a certain news outlet. However, our observations suggest that despite behavioural differences, on the aggregate level there is not much variation in terms of distribution of IPTC-based news categories consumed by different user groups. This does not mean that there is no difference between individual users, in particular those with very short and focused sessions (e.g. nibbers) and those with very long and changeable sessions (e.g. gourmets), especially on the level of individual news items. Instead, our observations indicate that, besides some exceptions, the average proportions of news categories consumed by different groups of users tend to be relatively close and generally follow the overall news supply by the respective outlet.

The results of our analyses point to the presence of distinct time- and content-based news consumption habits in non-personalised online environments. Some of these habits (e.g. time-based ones) seem to be more common between outlets, whereas others (e.g. content-based ones) are more distinct. Our analyses also emphasise the essential role of

content supply that aligns with earlier arguments about the importance of availability factors for the composition of individual news repertoires (Taneja et al., 2012). Independently of how different users consume news, their information diets seem to generally align with the distribution of the supplied content (with the exception of a few news categories which tend to be either under- or over-consumed – e.g., society for *AD* or politics for *Trouw*). These observations can be useful for tracing similarities/differences with studies using self-reported and small-scale experimental data and can be potentially used for modelling/simulating news readers' behaviour to trace how it can be affected by personalised news supply.

Our observations also raise a number of questions concerning the possible impact of algorithmic news recommenders on the user reading habits. For instance, to what degree the goal to increase users' engagement and the rate of return to news websites via personalised content suggestions is contradictory to user information behaviour, which usually involves very infrequent and spontaneous encounters? And, what is even more important, how significant will the impact of news recommenders be on the existing reading habits? To answer these questions, more empirical observations of user information behaviour similar to the ones provided in the current study are required.

Another question that is worthwhile exploring in the future studies is how similar or different the habits of the same users reading different newspapers are. Our observations indicate differences in the way users consume news, but will *Trouw* users keep their bimodal time-based reading habits while consuming content provided by *AD* or will they adapt to the unimodal mode of consumption typical for *AD* readers? The answer to this question is also important both for measuring the impact of algorithmic news recommenders and for designing them as it relates to the matter of how universal (or non-universal) can be algorithmic system designs to remain helpful for the users of specific media outlets.

At the same time, several limitations of our study should be acknowledged. First and foremost, we agree with Kormelink and Meijer (2018), who note that clickstream data provide limited insights into users' information diets as multiple reasons influence users' decisions to click or not to click. The quality of data provided to us also limited our possibilities for news reading habits analysis: for instance, we lacked data about the device used to access content as well as time spent by the user on a specific news page. Similarly, we lacked information about positioning of specific stories on the online newspage that can affect the distribution of user clicks.

The lack of granularity of clickstream data collected by *AD* and *Trouw* is particularly limiting in terms of isolating the effect of different forms of browsing behaviour on news consumption. While both the newspapers measure user clicks in the same way, thus restricting potential measurement biases, the inability to relate information behaviour data generated by the same user via different devices and/or from different browsers is a significant limitation of clickstream data. Even while in the current state of data it was still possible to acquire a number of insights into online news consumption, better data quality is important for improving the quality of future analyses.

The improvement of data quality is also important for enabling more fine-grained comparison of information behaviour between different news outlets. Multiple factors can influence the differences in news consumption between *AD* and *Trouw*, varying from the difference in readership to various promotion techniques and news formats. All these

factors can influence the composition of clickstream data and the insights drawn from them concerning user behaviour. Hence, the enrichment of clickstream data and their matching with other forms of information about the audience is an important prerequisite for future comparative research on news consumption.

Author's note

Jaron Harambam is now affiliated with the Institute for Media Studies at the KU Leuven.

Acknowledgement

The authors would like to thank three anonymous reviewers and Aleksandra Urman (University of Bern) and Olga Krasa-Ryabets (University of Amsterdam) for valuable feedback and insights. Additionally, the authors would like to thank Anne Schuth and Frank Mekkelholt from Persgroep for assisting with data acquisition and data analysis.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Netherlands Organisation for Scientific Research (grant 400.17.605).

ORCID iD


Mykola Makhortykh https://orcid.org/0000-0001-7143-5317

Supplemental material

Supplemental material for this article is available online.

Notes

1. For more information on algorithmically mediated changes in the relationship between the users and the media see Thurman (2011), Helberger (2015), Thorson and Wells (2016), Dörr (2016), Diakopoulos (2019), Bastian et al. (2019).
2. The full list of the supplements is available on the *AD* website: https://www.ad.nl/regio/
3. This specific way of id assignment limits the possibilities for consistent tracking of the behaviour of users who accessed news websites via different devices and/or using VPNs.
4. For *Trouw*, the articles produced from 1 January to 31 August 2018 constituted only 5% of all content interacted with by users during the 3 months of our study; however, these 5% of content attracted 77% of all clicks during summer 2018. In the case of *AD*, the articles produced from 1 January to 31 August 2018 constituted 35% of all content interacted with during the period of study and attracted 99% of clicks in summer 2018.
5. For more information about the taxonomy, see its description on the IPTC website: https://iptc.org/standards/media-topics/
6. These can be checked via the visualisation of the tree structure of the IPTC taxonomy http://show.newscodes.org/index.html?newscodes=medtop&lang=en-GB&startTo=Show
7. For instance, scarcity of topical categories would complicate the interpretability of the probability matrix for topic switches, in particular for topics with the very low number of articles. Similarly, in the case of content- and time-based reading habits, the preservation of the initial number of categories would negate the differences related to switches between related topics

(e.g. transition from a story on biology to a story on neurobiology) and switcher between substantially different topics (e.g. transition from a story on biology to the story on crime).

8. This category also included news about celebrities and lifestyle.
9. Following the work by Olmstead et al. (2011), we identify 'power users' as highly active news readers who engage with news content significantly more often than other users.
10. Most of these anomalously long sessions were the result of a single interaction of the respective user with the outlet. These sessions involved a rather high intensity of clicks (i.e. 3–4 articles per minute) and, in the case of *Trouw*, implied that the users had a subscription that allowed them to pass a cookie wall. In the case of *Trouw*, the sessions were primarily focused on culture and politics; some of the articles' were revisited more than once in the course of the session. In the case of *AD*, we found several extra-long sessions focused on a single economy-related article; the users clicked on it for 8–9 hours every 30 seconds. If for *Trouw* the anomalous activity can potentially be attributed to the human actor (e.g. members of the newsroom or data science team examining the website), then in the case of *AD* these sessions were most probably produced by automated agents used, for instance, to promote the material. Because of the lack of space, we did not investigate these instances of anomalous behaviour in more detail, but we consider looking at it in future research.
11. See the summaries of user interactions with content produced for *AD* and *Trouw* in the course of sessions made of 1 and 2+ clicks in the supplementary materials.
12. The summary of types is given in the supplementary materials. Following the common analogy between news and food diets (Costera Meijer and Groot Kormelink, 2015; Molyneux, 2018), we chose labels based on different modes of food consumption and used various lengths of eating sessions/variety of food consumed as a proxy for click session length/variety.
13. In Tables 4 and 5, we used bold font to highlight thematic categories the frequency of which within specific clusters was significantly different from the average frequency of the category within the whole corpus.

References

Anderson C (2013) Towards a sociology of computational and algorithmic journalism. *New Media & Society* 15(7): 1005–1021.

Bastian M, Makhortykh M and Dobber T (2019) News personalization for peace: how algorithmic recommendations can impact conflict coverage. *International Journal of Conflict Management* 30(3): 309–328.

Boczkowski P, Mitchelstein E and Matassi M (2018) 'News comes across when I'm in a moment of leisure': understanding the practices of incidental news consumption on social media. *New Media & Society* 20(10): 3523–3539.

Bodó B, Helberger N, Eskens S, et al. (2019) Interested in diversity: the role of user attitudes, algorithmic feedback loops, and policy in news personalization. *Digital Journalism* 7(2): 206–229.

Bos L, Kruikemeier S and de Vreese C (2016) Nation binding: how public service broadcasting mitigates political selective exposure. *PLoS ONE* 11(5): e0155112.

Costera Meijer I and Groot Kormelink T (2015) Checking, sharing, clicking and linking: changing patterns of news use between 2004 and 2014. *Digital Journalism* 3(5): 664–679.

Diakopoulos N (2019) *Automating the News: How Algorithms Are Rewriting the Media*. Cambridge, MA: Harvard University Press.

Dimmick J, Feaster J and Hoplamazian G (2011) News in the interstices: the niches of mobile media in space and time. *New Media & Society* 13(1): 23–39.

Dörr K (2016) Mapping the field of algorithmic journalism. *Digital Journalism* 4(6): 700–722.

Dunaway J, Searles K, Sui M, et al. (2018) News attention in a mobile era. *Journal of Computer-Mediated Communication* 23(2): 107–124.

Epure E, Kille B, Ingvaldsen JE, et al. (2017) Recommending personalized news in short user sessions. In: *Proceedings of the 11th ACM conference on recommender systems*, Como, 27–31 August, pp. 121–129. New York: ACM.

Esiyok C, Kille B, Jain B-J, et al. (2014) Users' reading habits in online news portals. In: *IIiX'14: proceedings of the 5th information interaction in context symposium*, Regensburg, 26 August, pp. 263–266. New York: ACM.

Eskens S, Helberger N and Moeller J (2017) Challenged by news personalisation: five perspectives on the right to receive information. *Journal of Media Law* 9(2): 259–284.

Flaxman S, Goel S and Rao J (2016) Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly* 80(S1): 298–320.

Genolini C, Ecochard R, Benghezal M, et al. (2016) kmlShape: an efficient method to cluster longitudinal data (time-series) according to their shapes. *PLoS ONE* 11(6): e0150738.

Gil De Zúñiga H, Molyneux L and Zheng P (2014) Social media, political expression, and political participation: panel analysis of lagged and concurrent relationships. *Journal of Communication* 64(4): 612–634.

Harambam J, Helberger N and van Hoboken J (2018) Democratizing algorithmic news recommenders: how to materialize voice in a technologically saturated media ecosystem. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376(2133): 20180088.

Helberger N (2015) Merely facilitating or actively stimulating diverse media choices? Public service media at the crossroad. *International Journal of Communication* 9: 1324–1340.

Helberger N and Wojcieszak M (2018) Exposure diversity. In: Napoli P (ed.) *Mediated Communication*. Berlin: De Gruyter, pp. 535–560.

Hermida A, Fletcher F, Korell D, et al. (2012) Share, like, recommend: decoding the social media news consumer. *Journalism Studies* 13(5–6): 815–824.

Incollingo J (2018) 'I'ma news junkie. . . I like being informed': mobile news use by a newspaper's digital subscribers. *Newspaper Research Journal* 39(2): 134–144.

Karimi M, Jannach D and Jugovac M (2018) News recommender systems – survey and roads ahead. *Information Processing & Management* 54(6): 1203–1227.

Kim S (2016) A repertoire approach to cross-platform media use behavior. *New Media & Society* 18(3): 353–372.

Kormelink T and Meijer I (2018) What clicks actually mean: exploring digital news user practices. *Journalism* 19(5): 668–683.

Lommatzsch A, Kille B and Albayrak S (2017) Incorporating context and trends in news recommender systems. In: *Proceedings of the international conference on web intelligence*, Leipzig, 23 August, pp. 1062–1068. New York: ACM.

Makhortykh M and Bastian M (2020) Personalizing the war: perspectives for the adoption of news recommendation algorithms in the media coverage of the conflict in Eastern Ukraine. *Media, War & Conflict*. Epub ahead of print 19 February. DOI: 10.1177/1750635220906254.

Mitchelstein E and Boczkowski P (2010) Online news consumption research: an assessment of past work and an agenda for the future. *New Media & Society* 12(7): 1085–1102.

Möller J, Trilling D, Helberger N, et al. (2016) Shrinking core? Exploring the differential agenda setting power of traditional and personalized news media. *Info* 18(6): 26–41.

Möller J, Trilling D, Helberger N, et al. (2018) Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society* 21(7): 959–977.

Möller J, van de Velde R, Merten L, et al. (2020) Explaining online news engagement based on browsing behavior: creatures of habit? *Social Science Computer Review* 38(5): 616–632. DOI: 10.1177/0894439319828012.

Molyneux L (2018) Mobile news consumption: a habit of snacking. *Digital Journalism* 6(5): 634–650.

Molyneux L (2019) Multiplatform news consumption and its connections to civic engagement. *Journalism* 20(6): 788–806.

Ohlsson J, Lindell J and Arkhede S (2017) A matter of cultural distinction: news consumption in the online media landscape. *European Journal of Communication* 32(2): 116–130.

Olmstead K, Mitchell A and Rosenstiel T (2011) Navigating news online: where people go, how they get there and what lures them away. *Pew Research Center's Project for Excellence in Journalism, Washington, DC*, 9 May, pp. 1–30.

Pariser E (2011) *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin Press.

Purcell K, Rainie L, Mitchell A, et al. (2010) Understanding the participatory news consumer. *Pew Internet and American Life Project, Washington, DC*, 1 March, pp. 19–21.

Schrøder K (2015) News media old and new: fluctuating audiences, news repertoires and locations of consumption. *Journalism Studies* 16(1): 60–78.

Stroud N (2010) Polarization and partisan selective exposure. *Journal of Communication* 60(3): 556–576.

Sunstein C (2017) *#Republic: Divided Democracy in the Age of Social Media*. Princeton, NJ: Princeton University Press.

Taneja H, Webster J, Malthouse E, et al. (2012) Media consumption across platforms: identifying user-defined repertoires. *New Media & Society* 14(6): 951–968.

Tewksbury D, Hals M and Bibart A (2008) The efficacy of news browsing: the relationship of news consumption style to social and political efficacy. *Journalism & Mass Communication Quarterly* 85(2): 257–272.

Thorson K and Wells C (2016) Curated flows: a framework for mapping media exposure in the digital age. *Communication Theory* 26(3): 309–328.

Thurman N (2011) Making 'The Daily Me': technology, economics and habit in the mainstream assimilation of personalized news. *Journalism* 12(4): 395–415.

Trilling D and Schoenbach K (2013) Skipping current affairs: the non-users of online and offline news. *European Journal of Communication* 28(1): 35–51.

Trouw (n.d.) Over ons. Available at: https://www.trouw.nl/nieuws/over-ons~b7aea298/ (accessed 19 June 2020).

Van Aelst P, Strömbäck J, Aalberg T, et al. (2017) Political communication in a high-choice media environment: a challenge for democracy? *Annals of the International Communication Association* 41(1): 3–27.

Van Damme K, Courtois C and Afschrift J (2014) Serendipitous news consumption: a mixed-method audience-centred study on mobile devices. In: *Amsterdam conference on social media and the transformation of public space*. Available at: https://biblio.ugent.be/publication/5638432/file/5638436

Van Damme K, Courtois C, Verbrugge K, et al. (2015) What's APPening to news? A mixed-method audience-centred study on mobile news consumption. *Mobile Media & Communication* 3(2): 196–213.

Westlund O and Färdigh M (2015) Accessing the news in an age of mobile media: tracing displacing and complementary effects of mobile news on newspapers and online news. *Mobile Media & Communication* 3(1): 53–74.

Yadamsuren B and Erdelez S (2011) Online news reading behavior: from habitual reading to stumbling upon news. *Proceedings of the American Society for Information Science and Technology* 48(1): 1–10.

Author biographies

Mykola Makhortykh is a postdoctoral researcher at the Institute of Communication and Media Studies at the University of Bern.

Natali Helberger is a professor of Information Law at the Institute for Information Law (IVIR), Faculty of Law of the University of Amsterdam.

Claes de Vreese is a professor and Chair of Political Communication at the Amsterdam School of Communication Research, University of Amsterdam.

Jaron Harambam is a postdoctoral researcher at the at the Institute for Information Law (IVIR), Faculty of Law of the University of Amsterdam.

Dimitrios Bountouridis is a postdoctoral researcher at the Delft University of Technology.