Through a different gate: An automated content analysis of how online news and print news differ

Burggraaff, C.; Trilling, D.

DOI
10.1177/1464884917716699

Publication date
2020

Document Version
Final published version

Published in
Journalism

Article

Journalism
2020, Vol. 21(1) 112–129

DOI: 10.1177/1464884917716699
journals.sagepub.com/home/jou
SAGE

Through a different gate: An automated content analysis of how online news and print news differ

Christiaan Burggraaff
University of Amsterdam, The Netherlands

Damian Trilling
University of Amsterdam, The Netherlands

Abstract

We investigate how news values differ between online and print news articles. We hypothesize that print and online articles differ in terms of news values because of differences in the routines used to produce them. Based on a quantitative automated content analysis of $N = 762{,}095$ Dutch news items, we show that online news items are more likely to be follow-up items than print items, and that there are further differences regarding news values like references to persons, the power elite, negativity, and positivity. In order to conduct this large-scale analysis, we developed innovative methods to automatically code a wide range of news values. In particular, this article demonstrates how techniques such as sentiment analysis, named entity recognition, supervised machine learning, and automated queries of external databases can be combined and used to study journalistic content. Possible explanations for the difference found between online and offline news are discussed.

Keywords

automated content analysis, computational methods, news values, online journalism, print journalism

Corresponding author:
Damian Trilling, Department of Communication Science, University of Amsterdam, Postbus 15791, 1001NG, Amsterdam, The Netherlands.
Email: d.c.trilling@uva.nl

Introduction

Over the last years, citizens have become more and more reliant on online news at the expense of print news.[1] While the news coverage on these two channels may seem interchangeable to the average news consumer, it is, in fact, not at all clear to what extent online and print news content are different.

Based on the hierarchy of influences model (Shoemaker and Reese, 2014), we argue that the routines of online news production can be expected to differ from the routines of offline news production. Journalism is going through a process of commercialization, which McManus (2009) defines as 'any action intended to boost profit that interferes with a journalists or news organizations best effort to maximize public understanding of those issues and events that shape the community they claim to serve' (p. 219). One example for this is journalists who try to reach as large an audience as possible to attract advertisers and investors. By analyzing website metrics, journalists have detailed insights into the preferences of the readers of their online content (Anderson, 2011; Tandoc and Thomas, 2014; Welbers et al., 2015), which makes it comparatively easy to strive to maximize their audience. To gather similar information about their print audience, journalists have to rely on occasional surveys or focus groups, which provide a much less fine-grained picture. Especially popular outlets (as opposed to quality outlets) can be expected to make use of this and to select stories that promise commercial success (Barnhurst and Nerone, 2009). Others have argued that in an online environment, journalists have a much higher workload than in a print environment. Journalists have to produce more stories per day and usually have to work in a smaller team. As a consequence, they have less time for important journalistic tasks like fact-checking (Witschge and Nygren, 2009), which also can be seen as a change of journalistic routines. In a print environment, journalists find it more important to fulfill an interpretative and investigative role than in an online environment, while in an online environment, they attach most importance to the dissemination of news (Cassidy, 2005). Online journalists want to publish news stories as soon as possible, while in the case of print newspapers with only one daily deadline, such extreme time pressure is less of an issue.

But do these differences in routines really lead to a difference in the journalistic gatekeeping process, 'the process of selecting, writing, editing, positioning, scheduling, repeating and otherwise massaging information to become news?' (Shoemaker et al., 2009). To shed light on this question, we conducted a quantitative content analysis of Dutch online and print news articles. It aims to answer the following overarching question: *To what extent does the content differ between online and print news?* In addition, as we included both popular and quality outlets in our sample, we will also address differences between them.

Theoretical background and related research

In the latest version of their hierarchy of influences model, Shoemaker and Reese (2014) distinguish five levels of what shapes media content: social systems, social institutions, media organizations, routine practices, and individuals (p. 9). When we compare online news and print news,[2] it seems obvious that the two outer levels, *social systems* and *social institutions*, are identical for both.

On the *organizational level*, undoubtedly, the introduction of online news has led to changes. When the first news sites emerged in the 1990s, their production was organizationally distinct from the production of a newspaper, featuring different newsrooms and different routines. Over the last couple of years, however, the division has become less clear (Avilés et al., 2009; Tameling, 2016; Vobic, 2011). In an effort to cut costs, media companies aim for convergence between the online and the print crew. Journalists of different platforms cooperate, stories are published on both channels, and some journalists work for both platforms. However, as Avilés et al. (2009) point out, many media companies struggle with finding the right level of convergence. For instance, Tameling (2016) describes how Dutch quality newspaper *de Volkskrant* moved back-and-forth between integration, de-integration, and again integration of online and offline newsrooms.

A direct consequence of the convergence is that, as many journalists work for both online and print, differences on the *individual level* are a problematic starting point for an analysis. Historically, although, role conceptions (an important point on the individual level, see Shoemaker and Reese, 2014: 232) of online and print journalists have shown to differ (e.g. Beam et al., 2009; Paulussen, 2006). More specifically, online journalists more often than print journalists saw themselves as disseminators and less often as interpreters (Carpenter et al., 2016; Cassidy, 2005; Deuze and Dimoudi, 2002; Hartley, 2013). We argue that even though print and online articles are nowadays often written by the same journalists, the results of such studies still can inform us about the differences in the routines that are applied when writing content produced for publishing on the online outlet, the offline outlet, or both.

Summing up, we argue that differences between online and offline news production are most likely to be explained by studying the level that Shoemaker and Reese (2014) refer to as *routine practices*.

Different routines in online and offline news production

Regardless of the state of newsroom convergence, there is at least one organizational goal that journalists can act upon to a very different degree, depending on whether they publish online or offline. Entman (2005) mentions both 'reporting on important events, people, and issues' and 'generating […] revenue' (p. 58) as central aim of media organizations (see also Shoemaker and Reese, 2014: 139). The hierarchy of influences model suggests that these organizational level goals will constrain and shape the next level, the level of routines.

Accordingly, while they may be reluctant to admit it, economic considerations can be a factor in the daily work of journalists (Tandoc, 2014). For example, when journalists publish a story online, they often monitor the number of readers and aim to maximize it. The trend to increasingly take these organizational demands into account in journalistic work routines can be described as *commercialization* (see also McManus, 2009).

In online news websites, journalists have a wide set of website metrics which enable them to see exactly which stories are read a lot and which are not (Karlsson and Clerwall, 2013; Tandoc and Thomas, 2014; Tandoc and Vos, 2015). Accordingly, several studies suggest quite a strong influence from audience clicks on the gatekeeping process (Anderson, 2011; Jacobi et al., 2015; Lee et al., 2012; Welbers et al., 2015). In the case

of print media, such numbers are not available, and therefore it is inherently impossible to adapt news selection decisions immediately to the readers' preferences. In online, it is very easy for news consumers to visit another website, whereas in case of print newspapers, readers are less flexible because they often have a subscription, canceling of which is a much higher hurdle than just directing your browser to a different site. Furthermore, online news consumers often visit news sites through links they find on other platforms like GoogleNews, Facebook, and Twitter, which increases competition. Thus, for online journalists, it is important to monitor and live up to the preferences of news consumers.

We can thus state that online news websites act in a much more competitive environment than traditional newspapers. In particular, journalists working in an online environment have to produce more stories and have to work much faster than their print colleagues (Bivens, 2008; Witschge and Nygren, 2009). Also, they have to work on several tasks at the same time (Boczkowski, 2009; Mitchelstein and Boczkowski, 2009; Quandt, 2008) and have less time for fact-checking because of the high speed in which stories must be published (Cassidy, 2006). Occasionally, even ethical standards are violated in order to be fast enough (Agarwal and Barthel, 2013; Cassidy, 2006). In short, their workload is much higher.

Shoemaker and Reese (2014) highlight that routines 'stem from three domains: audiences, organizations, and suppliers of content' (p. 168). This implies that routines differ between outlets with different audiences. Media research regularly distinguishes between quality outlets and popular outlets, and as Trilling and Schoenbach (2015) show for the Dutch case, their audiences overlap only to a limited extend. This will be reflected in different work routines of the journalists. Indeed, Boukes and Vliegenthart (2016) show that news values in quality and popular outlets differ, and Esser (1999) even suggests that there are distinct 'tabloid news values' (p. 293). Similarly, Barnhurst and Nerone (2009) argue that 'the mass circulation press included more event-oriented news, especially crime news, and also more reporting on social and cultural concerns, or so-called human interest stories' (p. 20).

Therefore, when studying the differences between online and offline news, it makes sense to also take into account whether the outlets in question can be categorized as popular or quality, to eliminate a possible confounding factor.

News values as dependent variables on the routine level

As Shoemaker and Reese (2014) point out, content characteristics can be studied both as dependent and as independent variables. In particular, they suggest that *news values* found in news articles can be analyzed as an outcome of journalistic work routines. Consequently, the differences outlined above can be expected to be reflected in the presence or absence of news values. News values can be seen as properties of an event, although – depending on their epistemological background – scholars differ in whether they see these properties as inherent to the event or as constructions made by journalists to justify their choices (e.g. Eilders, 2006). But regardless of whether we regard them as ascribed or inherent, we can state that news values are properties of an event that explain to which extent it is considered newsworthy.

We focus on a subset of news values (based on Harcup and O'Neill, 2001) on which online and print news are likely to differ and which were feasible to measure.

Power elite. Focus on power elite is probably one of the most clear-cut news values: if a story deals with important, relevant, well-known entities, it is newsworthy. This also makes sense, because in a democracy one of the most important tasks of the media is to control the elites.

RQ1. To what extent is there a difference in references to the power elite between news outlets?

Reference to persons. While Harcup and O'Neill (2001) initially considered this factor, they dropped it from their definite list of news values. Nevertheless, it has been investigated by many researchers since. Örnebring and Jönsson (2004) found that popular newspapers are more likely to focus on people than quality newspapers. Boukes and Vliegenthart (2016) even identify reference to persons as the one news value that differs most strongly between popular and quality newspapers. Furthermore, the difference in focus on political leaders between quality and popular newspapers is larger online (Jacobi et al., 2015). Thus, it may be expected that popular newspapers are more likely to apply a personalized reporting style, and that this difference is larger online.

H1a. Articles from popular outlets contain more references to persons than articles from quality outlets.

H1b. The difference between quality outlets and popular outlets in references to persons is larger in online news than in print news.

Celebrity news and entertainment news. Celebrity news and entertainment news may be the most despised categories of news by those who see news standards declining. It deals with topics like sex, show business, and human interest (Harcup and O'Neill, 2001). Examples from the tabloid press (see also Barnhurst and Nerone, 2009) show that it attracts large audiences. Since, especially online, journalists focus on an article's reach, the prevalence of such stories may be higher. Maier (2010) found that celebrity/entertainment news was one of the only three news categories (out of a total of 19) in which online newspapers published more stories than print newspapers, and also Van Der Wurff et al. (2008) found that news websites publish more entertainment stories than print newspapers. Furthermore, such stories often do not require a lot of research, which makes them easy to produce (Bird and Dardenne, 2009; Lehman-Wilzig and Seletzky, 2010).

H2. The likelihood that an article contains celebrity news is higher in online news than in print news.

H3. The likelihood that an article contains entertainment news is higher in online news than in print news.

Bad news/good news. Journalists have a tendency to cover mainly bad news (e.g. Leung and Lee, 2014). Galtung and Ruge (1965) offer several explanations for this: negative news is usually unexpected, unambiguous, it has a higher frequency and it fits into most peoples' picture of the world. It generally tends to attract a larger audience than positive news. Thus, especially routines that are shaped by commercial needs can be expected to

lead to the publishing of bad news. In addition, Shoemaker and Cohen (2006) argue that negative news is newsworthy because it is 'deviant' and thus important to monitor.

On the other hand, the rise of infotainment and soft news may also contribute to the production of positive news. As Leung and Lee (2014) found, journalists tend to believe that touching positive stories are popular; and Shoemaker and Cohen (2006) concede that even though 'deviant' bad news is newsworthy, positive news is also asked for by the audience. It is possible that online news contains not only more negative emotions but also more positive emotions compared to print news. In order to write entertaining articles, journalists may make use of an emotional tone of voice rather than a neutral writing style or select stories that lend themselves for such writing. Emotionality, in this context, can be seen as the presence of positivity and/or negativity as opposed to the absence of both.

RQ2a. To what extent is there a difference in the relative amount of negative news between print and online news?

RQ2b. To what extent is there a difference in the relative amount of positive news between print and online news?

RQ2c. To what extent is there a difference in the degree of emotionality between print and online news?

Follow-up news. When writing articles for an online outlet, journalists have to produce more articles than when writing for a print outlet (Witschge and Nygren, 2009). They also want to publish stories as quickly as possible (Agarwal and Barthel, 2013). As an event unfolds, news sites place updates on the issue. In contrast to a newspaper that is published once a day, a website therefore can have several stories on the same item within the same time period. Such stories on the same news topic that already has been in the news, can be called 'follow-up' news.

While such follow-up news can involve investigative reporting, we may speculate that more often, they are comparatively cheap to produce, as they usually do not involve digging up a completely new story. Online journalists report that they feel like they are not executing proper (i.e. investigative) journalism, because the workload is too high (Witschge and Nygren, 2009). This trend is in line with statements from newsmakers themselves who say that in online they strive for the ideal of ongoing 24-hour coverage, where being just minutes behind the competitors is already seen as failure (Bivens, 2008) – again giving an incentive to rather quickly publish small piecemeal stories than waiting for the one, great story to write.

H4. The amount of follow-up news is higher in online news than in print news.

Methods

Sample

In the period between 2 January 2014 and 31 December 2015, we collected all available news items from a selection of major Dutch news outlets, both online and print. For the

print data, we used the LexisNexis database. For the online data, every hour during the whole research period, we executed a Python script to check whether the main Really Simple Syndication (RSS) feeds of the news sites contained new items. All information from the feeds and the full webpage containing the article were downloaded, parsed, and stored in a database.

The original dataset consisted of a total of $N = 899,607$ articles. We removed $n = 20,939$ duplicates, $n = 3851$ articles with an original publishing date outside of the research period, $n = 108,888$ articles with a publishing date for which we did not have access to *both* online and offline articles, and $n = 3834$ articles with a length of less than 100 characters. The final dataset therefore consisted of $N = 762,095$ articles, roughly half of them online, half print. Table 1 gives a detailed overview.

Nu.nl is the largest Dutch news site. Although it has no offline counterpart, it is owned by publishing house Sanoma. *De Volkskrant, NRC Handelsblad*, and *Trouw* are the three largest quality papers. *Algemeen Dagblad* (AD) and *De Telegraaf* are the two newspapers with the highest overall circulation figures. *Geenstijl* is a weblog-style site owned by the same publisher as *De Telegraaf. Nederlandse Omroep Stichting* (NOS) is a Dutch public-service broadcaster, and *Metro* is the leading free daily newspaper. We explicitly did not include small regional papers or niche media, to avoid confounding the comparison by too large differences in the resources available.

Independent variables

Article length. The articles in the sample varied greatly in length. After some impossible values were removed, the overall average article length was $M = 1693.73$ characters ($SD = 1936.55$, $min = 100$, $max = 134,544$).

Platform. We created a dummy variable that was coded as 1 for all print articles and 0 for all online articles. In addition, we created a second dummy variable that was 1 if the article was non-exclusive, that is, if a highly similar article was published on the other platform on the same day or the day before. If the cosine similarity between two articles was $> .7$, they were considered non-exclusive. The same cut-off value (.7) was used by Welbers et al. (2016); Boumans (2016) used a marginally less conservative value of .65.

Popular and quality news. We created a dummy variable that was coded 0 for quality outlets and 1 for popular outlets. *Telegraaf, AD, Geenstijl*, and *Metro* were considered popular outlets. Sparks (2000) and Deuze (2005) see the lack of clear distinction between information and entertainment as characteristic for the latter; however, as Deuze (2005) notes, the Dutch news media landscape lacks extreme forms of tabloid papers. Therefore, we also relied on an additional indicator, namely the self-description of the outlets and their target audiences. *Telegraaf, AD*, and *Metro* all define themselves as papers for 'the ordinary citizen', and *GeenStijl* clearly lacks the information/entertainment distinction.

Dependent variables

In order to investigate the prevalence of different news values on these various platforms, we used Python to conduct an automated content analysis (see Boumans and

Table 1. Number of articles per source.

Source	Online*	Print*	Non-exclusive**	Date range***
nu.nl	26,746 (26,746)	(online-only)	(online-only)	2 January 2014 to 23 December 2015
AD	30,392 (28,174)	32,437 (30,218)	2,218/2,219	2 January 2014 to 1 September 2014
Telegraaf	183,376 (172,963)	110,960 (100,724)	10,413/10,236	2 January 2014 to 31 December 2015
nos.nl	23,844 (23,844)	(online-only)	(online-only)	1 January 2015 to 31 December 2015
Volkskrant	76,229 (67,916)	66,105 (57,576)	8,313/8,529	2 January 2014 to 31 December 2015
NRC	26,358 (21,759)	71,060 (66,725)	4,599/4,335	2 January 2014 to 31 December 2015
Trouw	5,926 (4,986)	10,614 (9,731)	940/883	31 December 2014 to 18 May 2015
Metro	58,742 (53,192)	35,311 (29,648)	5,550/5,663	2 January 2014 to 23 December 2015
geenstijl.nl	3,995 (3,995)	(online-only)	(online-only)	30 December 2014 to 31 December 2015
Total	435,608 (403,577)	326,487 (294,622)	63,896	

AD: *Algemeen Dagblad*.
*Numbers between brackets exclude articles published both online and print. **Number of online articles followed by number of print articles. The numbers can differ because based on one online article two print articles may be written, or vice versa. ***Because of data availability issues, some sources were analyzed for less than the 2-year period of this study.

Trilling, 2016; Grimmer and Stewart, 2013; Günther and Quandt, 2016). In doing so, we extend earlier work by Trilling et al. (2017), who used such techniques to automatically code a smaller set of news values than we do in this article.

Power elite. News on power elite was defined in terms of political, economic, and geographical power. In order to find characteristic words of political power, from a set of articles that were known to be political news, we compiled a list of words that unambiguously referred to political power. The coding was iterative and was continued until no new words came up. The same process was repeated to add characteristic words of economic power to the list.

In order to find countries that can be considered to be part of the power elite, for those countries belonging to the global G20, and for those with the highest gross domestic product (GDP) per citizen and for those with the highest overall GDP, the country name, capital, seat of government, and the biggest city were added to the list.

Every occasion of one of the words from the list in the body of an article meant an increase in power elite score by 1 point, an occurrence in the title meant an increase by 2 points ($M = 3.44$, $SD = 4.10$, $min = 0$, $max = 77$) .

References to persons. References to persons were determined using named entity recognition (NER) (e.g. Nadeau and Sekine, 2007). NER identifies entities (like locations or persons) in a text. We used the Python Natural Language Toolkit (Bird et al., 2009) to train a Naïve Bayes classifier and chunker. As training data, the Dutch version of the conll2002 data were used (Tjong Kim Sang and De Meulder, 2003). The trained model achieved an F1-score of 69.3 percent (precision: 66.9%, recall: 71.9%). Using this NER-model, we counted references to persons ($M = 2.98$, $SD = 4.40$, $min = 0$, $max = 771$ (sic)).

Celebrity news. In order to find characteristic words of celebrity news, from a set of articles that were known to contain celebrity news, we compiled a list of roles or 'jobs' (like actor, anchorman, …) that unambiguously referred to celebrities. Again, the coding was iterative and was carried out until no further references could be found. In order to make sure that only well-known celebrities were included, the word 'famous' was added for most keywords. With the resulting list of potential jobs, a SPARQL Protocol and RDF Query Language (SPARQL) query was set up in order to find DBpedia articles (the machine-readable form of Wikipedia). The objective of this query was to find Dutch Wikipedia articles about persons, where one of the 'celebrity jobs' was mentioned in the abstract. For some of the more specific keywords (like TV-anchormen), a prerequisite to be included was that they were Dutch in order to avoid an unnecessary high number of false positives, while in the case of, say, actors (which in Dutch has a less broad meaning than in English), this requirement was not set. Still, the query resulted in a list of almost 10,000 celebrities. The dataset was then searched for the appearance of three or more of those celebrities, resulting in a celebrity score for every article by simply counting the number of occurrences ($M = 0.24$, $SD = 0.91$, $min = 0$, $max = 67$). The sample contained a total of $n = 23,502$ celebrity articles.

Entertainment news. Our database also stored the subdirectories of the website where a given article was published. Not all news sources categorize their articles reliably, but *nu.nl* does: Each article has a label attached that identifies its category, for example, 'economic news', 'sports', 'entertainment', and so on. These labels were used for a supervised machine learning approach. After comparing the performance of different classifiers, we chose to train a Naïve Bayes classifier ($N_{train} = 14,000, N_{test} = 14,000$). This classifier was able to well distinguish between entertainment and non-entertainment articles (accuracy: 0.98, precision: 0.81, recall: 0.85). However, when applying the classifier to other newspapers, we realized that often sports articles where mistakenly classified as entertainment. We therefore trained another classifier on the *nu.nl* dataset to identify the characteristics of sports news (accuracy: 0.99, precision: 0.90, recall: 0.87). We therefore only regard articles as entertainment articles if they are classified as entertainment, but not as sports.

Positive/negative news. In order to measure the amount of positive and negative news, a sentiment analysis was carried out for each article using the Sentistrength software for Dutch (Thelwall et al., 2010). Each article was assigned a score for the amount of positivity ($M = 2.01$, $SD = 1.00$, $min = 1, max = 5$) and negativity ($M = -2.81$, $SD = 0.90$, $min = -5$, $max = -1$) which makes it possible to compare the emotionality of different

articles. As Thelwall et al. (2010) point out, sentiment is not a two-dimensional scale formed by positivity on the one and negativity on the other end: rather both are concepts that do not necessarily have to be correlated strongly and as such can (and have to be) measured individually. Therefore, by adding up the absolute values of positivity and negativity, we were able to determine emotionality ($M = 4.82$, $SD = 1.50$, $min = 2$, $max = 10$).

Follow-up news. An article was considered a follow-up article if its topic was covered in another article from the same source that was published up to 2 days in advance. In order to find such articles, the cosine similarity between a published article and all articles on the following 2 days were calculated. If the cosine similarity was higher than 0.5, an article was considered a follow-up article of the previously published article. This threshold was determined by trying several thresholds on a random dataset of 100 different articles, and it turned out that a threshold of 0.5 yielded the best results. In total, $n = 42,226$ (5.54%) of the articles were follow-up articles.

Results

To give a first overview of the results, we present the descriptive statistics of all variables of interest, split by category, in Tables 2 and 3. While these are interesting to get a general understanding of the data, we employ regression models to answer our research questions and to test our hypotheses.

Research Question 1 asked in how far different outlets differ in their references to elites. The descriptive statistics (Table 2) suggested that on average, the number of elite references is higher in print news ($M = 3.72$, $SD = 4.40$) than in online news ($M = 3.23$, $SD = 3.85$). Yet, as our regression model in Table 4 reveals, when article length is controlled for, print articles turn out to score lower than online articles. This means that the prevalence of elite references in online articles is more dense, but overall the number of elite references is still higher in print news. Popular outlets, on average, score lower on elite references than quality newspapers. Non-exclusive articles did not differ significantly from other articles.

Hypothesis 1a predicted that the degree of personalization is higher in popular outlets compared to quality outlets. The descriptive statistics in Table 2 indicate that popular outlets score lower on personalization than quality outlets. However, the print articles were considerably longer than online articles ($M = 2201.56$, $SD = 2257.42$ compared to $M = 1313.12$, $SD = 1550.30$), and in longer articles the chance that such references are found is almost by definition higher than in shorter articles. Accordingly, the results of the regression analysis in Table 4 show that – when controlling for article length and platform – popular outlets score higher on personalization than quality outlets, which is in line with H1a. Further, personalization was found to be used more often in print news than in online news. Also H1b receives support. We found an interaction effect of platform and news type, as expected, in online the difference between popular news and quality news is bigger than in print.

Hypothesis 2 predicted that online articles are more likely to be about celebrity news than print articles. Our logistic regression model in Table 5 shows that – while the model exhibits a poor model fit – there even is a small effect in the opposite direction. H2 was

Table 2. Overall means and standard deviations of continuous variables.

	N	Elite	Positivity	Negativity	Emotionality	Persons
Overall	762,095	3.44 (4.10)	2.01 (1.00)	−2.81 (0.90)	4.82 (1.50)	2.99 (4.40)
Online	435,608	3.23 (3.85)	1.88 (0.96)	−2.77 (0.87)	4.64 (1.38)	2.36 (3.89)
Print	326,487	3.72 (4.40)	2.20 (1.03)	−2.86 (0.94)	5.06 (1.62)	3.80 (4.92)
Exclusively online	403,577	3.21 (3.82)	1.87 (0.96)	−2.76 (0.87)	4.63 (1.37)	2.34 (3.79)
Exclusively print	294,622	3.73 (4.41)	2.20 (1.03)	−2.86 (0.94)	5.05 (1.62)	3.80 (4.91)
Non-exclusive	63,896	3.57 (4.28)	2.08 (1.01)	−2.84 (0.91)	4.92 (1.53)	3.23 (4.86)
Popular	455,213	2.66 (3.35)	1.96 (1.00)	−2.69 (0.91)	4.65 (1.48)	2.44 (3.43)
Quality	306,882	4.59 (4.79)	2.09 (1.00)	−2.98 (0.86)	5.07 (1.49)	3.78 (5.44)

Means with standard deviations between brackets.

Table 3. Overall percentages of dichotomous variables.

	N	Celebrity (%)	Entertainment (%)	Follow-up (%)
Overall	762,095	3.08	17.27	5.54
Online	435,608	2.90	12.94	6.79
Print	326,487	3.32	23.05	3.88
Exclusively online	403,577	2.89	12.88	6.91
Exclusively print	294,622	3.30	23.02	3.82
Non-exclusive	63,896	3.33	18.50	4.83
Popular	455,213	3.12	16.56	5.44
Quality	306,882	3.04	18.33	5.69

Table 4. Ordinary least squares (OLS) regressions for RQ1 and H1.

	Power elite: RQ1		Personalization: H1	
	b (SE)	β	*b* (SE)	β
Length	0.001 (0.000)	0.46***	0.001 (0.000)	0.55***
Platform = print	−0.45 (0.01)	−0.05***	0.70 (0.01)	0.08***
Platform = print and online	−0.04 (0.01)	0.00	0.01 (0.02)	0.00
Outlet = popular	−0.89 (0.01)	−0.11***	0.35 (0.01)	0.04***
Print × popular			−0.64 (0.02)	−0.06***
R^2	0.24		0.32	

SE: standard error.
Reference categories: platform = online, outlet = quality.
N = 762,095 .
*$p < .05$ **$p < .01$ ***$p < .001$.

not supported. Further analysis showed that popular outlets are more likely to contain celebrity news than quality outlets. No significant difference was found between non-exclusive articles and the other articles. To ease interpretation, we calculated the

Table 5. Logistic regressions.

	Celebrity news: H2	Entertainment news: H3	Follow-up: H4
	Odds ratio (SE)	Odds ratio (SE)	Odds ratio (SE)
Length	1.00 (.000)***	1.00*** (.000)	1.00 (.000)
Platform = print	1.03 (.01)*	2.01*** (.01)	.55 (.01)***
Platform = print and online	1.06 (.02)*	1.04*** (.01)	.89 (.02)***
Outlet = popular	1.25 (.02)***	.94*** (.01)	.90 (.01)***
Cragg-Uhler pseudo-R^2	.015	.029	.012

SE: standard error.
Reference categories: platform = online, outlet = quality.
$N = 719,294$.
*$p < .05$ **$p < .01$ ***$p < .001$.

likelihood for an article to contain celebrity references.[3] Based on the regression model, we estimated these likelihoods as follows: 2.95 percent for online articles, 3.03 percent for print articles, 2.62 percent for quality outlets, and 3.26 percent for popular outlets.

Hypothesis 3 predicted that online articles are more likely to be about entertainment news than print articles. In contrast with the hypothesis, Table 5 shows that, all other variables being equal, the odds for a print article to be about entertainment news are twice as high as for an online article. We find such a strong effect only for articles that are *exclusively* published in print. Interestingly, quality outlets are slightly more likely to publish entertainment articles than quality papers. Again, we calculated the likelihood for an article to be about entertainment news: online 12.94 percent, print 23.03 percent, quality outlets 17.27 percent, and popular outlets 16.34 percent.

Research Question 2 asked about the differences in tone. The descriptive statistics in Table 2 indicate that print news score higher on both positivity and on negativity. However, as mentioned before, in the longer print articles the chance that emotional words are found can be expected to be higher than in the shorter online articles. Accordingly, when controlling for length in a regression analysis (Table 6), the standardized β-coefficients show that length has the by far the strongest influence on the tone. Our first impression regarding the differences between online and offline outlets seems to hold only in the case of positivity: print news seem to be more positive, but slightly *less* negative (recall that negativity was coded on a scale from -1 to -5, thus a positive coefficient means less negativity).

We subsequently computed the overall emotionality by adding up the absolute values on positivity and negativity. Our regression model in Table 6 suggests that on average, print articles score higher on emotionality than online articles.

Hypothesis 4 predicted a higher amount of follow-up news in the case of online articles. The results of a logistic regression (Table 5) show striking differences: for print articles, the odds ratio to be a follow-up article is 45 percent lower than for online articles. For non-exclusive articles, the odds ratio to be a follow-up article is 11 percent lower compared to other articles. These findings offer support for hypothesis 4, although we have to keep in mind that the model fit is rather poor. This is also illustrated by the

Table 6. Ordinary least squares (OLS) regressions for RQ2.

	Negativity: RQ2a		Positivity: RQ2b		Emotionality: RQ2c	
	b (SE)	β	*b* (SE)	β	*b* (SE)	β
Length	−.000 (.000)	−.34***	.000 (.000)	.37***	.000 (.000)	.45***
PF = print	.06 (.00)	.03***	.16 (.00)	.08***	.10 (.00)	.03***
PF = print and online	−.01 (.00)	.00	.02 (.00)	.01***	.03 (.01)	.01***
Outlet = pop.	.12 (.00)	.06***	.09 (.00)	.05***	−.03 (.00)	−.01***
R^2	.13		.15		.22	

SE: standard error.
Reference categories: platform = online, outlet = quality.
$N = 719,294$.
*$p < .05$ **$p < .01$ ***$p < .001$.

likelihood scores we calculated: the likelihood for an article to be a follow-up article is 6.78 percent in the case of online news, 3.87 percent in the case of print news, 5.13 percent in the case of popular outlets, and 5.67 percent in the case of quality outlets.

Discussion and conclusion

The contribution of this article is twofold: methodological and substantial. Methodologically, we showed how automated content analysis can be employed to analyze news values on a large-scale; substantially, we identified differences between online and offline news as well as between popular and quality news outlets. We will first discuss the strengths, weaknesses, and implications of our methods, before we turn to a discussion of the results themselves.

While there are more and more studies that employ automated content analysis to analyze journalistic products, to the best of our knowledge, only one study (Trilling et al., 2017) has done so with the explicit goal of automatically coding news values, although others have essentially automatically coded the same or similar variables with slightly different goals, for example, to assess the quality of journalism (Jacobi, 2016). In this article, we extended this line of research by showing that automated content analysis can be used to identify a number of news factors that had not been coded in an automated way before.

In particular, to our knowledge, we are the first to employ resources like DBpedia and techniques like NER to identify news values. We believe that such linking of different data sources can be a fruitful way to automatically identify actors and other entities in a text. Nevertheless, more research is needed to further improve and validate our methods.

Reliability is not an issue in a automated content analysis, as re-running the analysis will yield the same results, but validity can be problematic (Boumans and Trilling, 2016; Grimmer and Stewart, 2013). Over the given time frame, all articles were included. External validity should be guaranteed with regard to the current state of the media landscape. Yet, extrapolations to the past and the future are dangerous because the media

environment is in a continuous state of transition. Concerning internal validity, where possible, precision and recall were calculated. While in case of entertainment articles, the results were excellent, the precision and recall for the NER-tagger were acceptable but left room for improvement. With regard to the validity of our coding of positive and negative emotions, Sentistrength (the algorithm we used) has so far mainly been used for analysis of rather short texts. While the software seems capable of analysis of longer texts as well, it would be useful to evaluate the method in comparison to, for example, a supervised machine learning approach (Gonzalez-Bailon and Paltoglou, 2015). Concerning follow-up news and non-exclusive articles, it may be useful for future research to check if the results match manual coding, but no reasons were found to question the results of the applied method, that is, the use of the cosine similarity, which is a very widespread measure of overlap between texts.

The way in which we categorized entertainment news probably can be improved. Most notably, after conducting some additional analyses, we realized that our theoretical ideas about what constitutes entertainment and the empirical classification diverged. In line with the vanishing distinction between low and high culture, we observed that also our classifier picked up both, for instance, rumors about artists, and serious reviews about movies, CDs, or books. This explains the unexpected result of print and quality news having a high amount of what we called entertainment news – what we measured should rather be called 'culture and entertainment'.

On the substantial side, our analyses show that there are significant differences between online and print news. We argued that they can be explained by focusing on the journalistic routines, which can be understood using the hierarchy of influences model (Shoemaker and Reese, 2014). A content analysis cannot provide evidence what has *caused* the difference we found. Nevertheless, for instance, it seems very plausible to assume that it is the high workload in online journalism that affects the form and the content of online news (Boczkowski, 2009; Mitchelstein and Boczkowski, 2009; Witschge and Nygren, 2009). In our data, we saw that some newsrooms even publish more articles online than offline. In order to be able to publish so many articles online, journalists seem to publish shorter articles. This may have to do with the characteristics of the Internet itself, but it seems reasonable to assume that the workload contributes to this.

The working speed of in the production of online news may also explain, for instance, differences in the personalization score. Print articles, on average, score higher, but this does not necessarily mean that print articles focus more on private lives. It can also mean that in print articles, journalists quote more human sources (e.g. spokespersons), while online, they rely more on written information like press releases. This is a faster way of working because no extra research needs to be done. But of course, it is mainly the power elites who have access to such means. If this is the case, it could explain why elite news coverage is so dense in online news, while personalization is rather low. For their print articles, journalists might tend to make more effort in approaching several (human) sources when doing research, while online, they might prefer to mention only the name of a power elite, for example, an organization, instead of doing further research (Witschge and Nygren, 2009). The fact that popular news contains more references to persons is consistent with research showing that popular papers personalize more (e.g. Örnebring and Jönsson, 2004). We showed that especially in online, this is true: while both in

popular and quality outlets, journalists may use less human sources online, in popular online outlets, journalists still write quite some personal stories, while their quality colleagues do not do this.

This findings on source usage match the finding that online, journalists are much more likely to publish follow-up articles. In fact, here, we find one of the biggest differences between print and online articles: online articles are almost twice as likely to be a follow-up article than print articles. Follow-up articles are easier to produce than other articles, because instead of finding a completely new idea, journalists can work on one story for which they only have to look for updates. Often, this is an easy way to produce many articles, even under time pressure. In line with findings of, among others, Cassidy (2006), Bivens (2008), and Carpenter et al. (2016), in an online environment, journalists seem to want to publish news as soon as possible and therefore rather publish a second article with new findings than wait until they had time for further research.

Contrary to our expectations, online news were neither characterized by a high amount of celebrity news nor entertainment news. Given that we used an innovative method for the measurement of these variables, further research should validate this finding to exclude the possibility that this is a measurement artifact.

Overall, we showed that there are visible differences between online and print news in terms of news values. These differences are related to the different routines used in the two news environments. While this article provides a rather general overview and can be seen as a first step in the quantitative analysis of news values in online and offline news, further (also qualitative or comparative) research is necessary. For future research, it may be fruitful to also investigate the actual amount of readers in order to shed further light on the use of website metrics. Given the importance of journalism for democracy, we call for further investigation of differences between online and print journalism, in order to be able to get a more complete picture of the changing nature of journalism in today's changing media environment.

Acknowledgements

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. The authors would like to thank Anne Kroon and Joanna Strycharz for their invaluable contribution to the data collection.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. A representative survey by the authors' research group suggests that 55 percent of the Dutch population reads a newspaper on 3 or more days per week and 50 percent read a news website on 3 or more days per week. On average, newspapers are read on $M = 3.33$ ($SD = 2.84$) days per week and websites on $M = 3.03$ ($SD = 2.77$) days.
2. We use a straightforward criterion: online news is brought to the public via the Internet, while print news reaches the audience via the newspaper. Digital 1:1 copies of newspapers (readable, for example, on a tablet) and mobile apps are beyond the scope of this article.

3. Likelihood was calculated by centering all other independent variables and then calculating the chance for an article from the respective platform to contain to contain celebrity references.

References

Agarwal SD and Barthel ML (2013) The friendly barbarians: Professional norms and work routines of online journalists in the United States. *Journalism* 16(3): 376–391.

Anderson C (2011) Between creative and quantified audiences: Web metrics and changing patterns of newswork in local US newsrooms. *Journalism* 12(5): 550–566.

Avilés JAG, Meier K, Kaltenbrunner A, et al. (2009) Newsroom integration in Austria, Spain and Germany. *Journalism Practice* 3(3): 285–303.

Barnhurst KG and Nerone J (2009) Rethinking news and myth as storytelling. In: Wahl-Jorgensen K and Hanitzsch T (eds) *The Handbook of Journalism Studies*. New York: Routledge, pp. 17–28.

Beam RA, Weaver DH and Brownlee BJ (2009) Changes in professionalism of U.S. journalists in the turbulent twenty-first century. *Journalism & Mass Communication Quarterly* 86(2): 277–298.

Bird S, Klein E and Loper E (2009) *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly.

Bird SE and Dardenne RW (2009) Rethinking news and myth as storytelling. In: Wahl-Jorgensen K and Hanitzsch T (eds) *The Handbook of Journalism Studies*. New York: Routledge, pp. 205–217.

Bivens RK (2008) The Internet, mobile phones and blogging. *Journalism Practice* 2(1): 113–129.

Boczkowski PJ (2009) Rethinking hard and soft news production: From common ground to divergent paths. *Journal of Communication* 59(1): 98–116.

Boukes M and Vliegenthart R (2016) A general pattern of newsworthiness? Analyzing news factors in tabloid, broadsheet, financial and regional newspapers. In: *Annual conference of the International Communication Association*, Fukuoka, Japan, 25–29 May.

Boumans JW (2016) *Outsourcing the news? An empirical assessment of the role of sources and news agencies in the contemporary news landscape*. PhD Thesis, University of Amsterdam. Available at: http://hdl.handle.net/11245/1.532941

Boumans JW and Trilling D (2016) Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism* 4(1): 8–23.

Carpenter S, Boehmer J and Fico F (2016) The measurement of journalistic role enactments: A study of organizational constraints and support in for-profit and nonprofit journalism. *Journalism & Mass Communication Quarterly* 93(3): 587–608.

Cassidy WP (2005) Variations on a theme: The professional role conceptions of print and online newspaper journalists. *Journalism & Mass Communication Quarterly* 82(2): 264–280.

Cassidy WP (2006) Gatekeeping similar for online, print journalists. *Newspaper Research Journal* 27(2): 6–23.

Deuze M and Dimoudi C (2002) Online journalists in the Netherlands: Towards a profile of a new profession. *Journalism* 3(1): 85–100.

Deuze M (2005) Popular journalism and professional ideology: Tabloid reporters and editors speak out. *Media, Culture & Society* 27(6): 861–882.

Eilders C (2006) News factors and news decisions. Theoretical and methodological advances in Germany. *Communications* 31(1): 5–24.

Entman RM (2005) The nature and sources of news. In: Overholser G and Jamieson K (eds) *The Institutions of American Democracy: The Press*. New York: Oxford University Press, pp. 48–65.

Esser F (1999) 'Tabloidization'of news: A comparative analysis of Anglo-American and German press journalism. *European Journal of Communication* 14(3): 291–324.

Galtung J and Ruge MH (1965) The structure of foreign news. *Journal of Peace Research* 2(1): 64–91.

Gonzalez-Bailon S and Paltoglou G (2015) Signals of public opinion in online communication: A comparison of methods and data sources. *The Annals of the American Academy of Political and Social Science* 659(1): 95–107.

Grimmer J and Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267–297.

Günther E and Quandt T (2016) Word counts and topic models. *Digital Journalism* 4(1): 75–88.

Harcup T and O'Neill D (2001) What is news? Galtung and Ruge revisited. *Journalism Studies* 2(2): 261–280.

Hartley JM (2013) The online journalist between ideals and audiences. *Journalism Practice* 7(5): 572–587.

Jacobi C, Kleinen-von Königslöw K and Ruigrok N (2015) Political news in online and print newspapers. *Digital Journalism* 4(6): 723–742.

Jacobi CA (2016) *The quality of political news in a changing media environment*. PhD Dissertation, University of Amsterdam. Available at: http://hdl.handle.net/11245/1.503897

Karlsson M and Clerwall C (2013) Negotiating professional news judgment and 'clicks'. *Nordicom Review* 34(2): 65–76.

Lee AM, Lewis SC and Powers M (2012) Audience clicks and news placement: A study of time-lagged influence in online journalism. *Communication Research* 41(4): 505–530.

Lehman-Wilzig S and Seletzky M (2010) Hard news, soft news, 'general' news: The necessity and utility of an intermediate classification. *Journalism* 11(1): 37–56.

Leung DKK and Lee FLF (2014) How journalists value positive news. *Journalism Studies* 16(2): 289–304.

McManus J (2009) The commercialization of news. In: Wahl-Jorgensen K and Hanitzsch T (eds) *The Handbook of Journalism Studies*. New York: Routledge, pp. 218–233.

Maier SR (2010) Newspapers offer more news than do major online sites. *Newspaper Research Journal* 31(1): 6–19.

Mitchelstein E and Boczkowski PJ (2009) Between tradition and change: A review of recent research on online news production. *Journalism* 10(5): 562–586.

Nadeau D and Sekine S (2007) A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30(1): 3–26.

Örnebring H and Jönsson AM (2004) Tabloid journalism and the public sphere: A historical perspective on tabloid journalism. *Journalism Studies* 5(3): 283–295.

Paulussen S (2006) Online news production in Flanders: How Flemish online journalists perceive and explore the Internet's potential. *Journal of Computer-mediated Communication* 9(4). Available at: http://dx.doi.org/10.1111/j.1083-6101.2004.tb00300.x

Quandt T (2008) News tuning and content management: An observation study of old and new routines in German online newsrooms. In: Paterson C and Domingo D (eds) *Making Online News: The Ethnography of New Media Production*. New York: Peter Lang, pp. 77–97.

Shoemaker P, Vos T and Reese S (2009) Journalists as gatekeepers. In: Wahl-Jorgensen K and Hanitzsch T (eds) *The Handbook of Journalism Studies*. New York: Routledge, pp. 73–87.

Shoemaker PJ and Cohen AA (2006) *News around the World*. New York: Routledge.

Shoemaker PJ and Reese SD (2014) *Mediating the Message in the 21st Century: A Media Sociology Perspective*, 3rd edn. New York: Routledge.

Sparks C (2000) The panic over tabloid news. In: Sparks C and Tulloch J (eds) *Tabloid Tales*. Lanham, MD: Rowman & Littlefield, pp. 1–40.

Tameling K (2016) *En wat doen we online? Crossmediale dilemma's op de nederlandse nieuwsredactie*. PhD Thesis, Rijksuniversiteit Groningen. Available at: http://www.rug.nl/research/portal/files/20147470/Complete_thesis.pdf

Tandoc EC (2014) Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media & Society* 16(4): 559–575.

Tandoc EC and Thomas RJ (2014) The ethics of web analytics. *Digital Journalism* 3(2): 243–258.

Tandoc EC and Vos TP (2015) The journalist is marketing the news. *Journalism Practice* 10: 950–966.

Thelwall M, Buckley K, Paltoglou G, et al. (2010) Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12): 2544–2558.

Tjong Kim Sang EF and De Meulder F (2003) Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition *In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL2003*. Stroudsburg, PA: Association for Computational Linguistics, pp. 142–147.

Trilling D and Schoenbach K (2015) Investigating people's news diets: How online news users use offline news. *Communications: The European Journal of Communication Research* 40(1): 67–91.

Trilling D, Tolochko P and Burscher B (2017) From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics. *Journalism & Mass Communication Quarterly* 94(1): 38–60.

Van Der Wurff R, Lauf E, Balčytienė A, et al. (2008) Online and print newspapers in Europe in 2003. Evolving towards complementarity. *Communications: The European Journal of Communication Research* 33(4): 403–430.

Vobic I (2011) Online multimedia news in print media: A lack of vision in Slovenia. *Journalism* 12(8): 946–962.

Welbers K, Van Atteveldt W, Kleinnijenhuis J, et al. (2015) News selection criteria in the digital age: Professional norms versus online audience metrics. *Journalism* 17: 1037–1053

Welbers K, Van Atteveldt W, Kleinnijenhuis J, et al. (2016) A gatekeeper among gatekeepers. *Journalism Studies*. Epub ahead of print 30 June. DOI: 10.1080/1461670X.2016.1190663.

Witschge T and Nygren G (2009) Journalistic work: A profession under pressure? *Journal of Media Business Studies* 6(1): 37–59.

Author biographies

Christiaan Burggraaf, MSc, works as a freelance journalist. He received his Master of Science at the Department of Communication Science, University of Amsterdam, The Netherlands.

Damian Trilling, PhD, is an Assistant Professor of Political Communication and Journalism in the Department of Communication Science, University of Amsterdam, The Netherlands, where he also received his PhD. He is affiliated with the Amsterdam School of Communication Research. He studies the changing news media landscape using methods of automated content analysis and computational social science.