Frames beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue

Burscher, B.; Vliegenthart, R.; de Vreese, C.H.

DOI
10.1177/0894439315596385

Publication date
2016

Document Version
Final published version

Published in
Social Science Computer Review

Article

Social Science Computer Review
2016, Vol. 34(5) 530-545

DOI: 10.1177/0894439315596385
ssc.sagepub.com
SAGE

Frames Beyond Words: Applying Cluster and Sentiment Analysis to News Coverage of the Nuclear Power Issue

Bjorn Burscher[1], Rens Vliegenthart[1], and Claes H. de Vreese[1]

Abstract

Methods to automatically analyze media content are advancing significantly. Among others, it has become increasingly popular to analyze the framing of news articles by means of statistical procedures. In this article, we investigate the conceptual validity of news frames that are inferred by a combination of *k*-means cluster analysis and automatic sentiment analysis. Furthermore, we test a way of improving statistical frame analysis such that revealed clusters of articles reflect the framing concept more closely. We do so by only using words from an article's title and lead and by excluding named entities and words with a certain part of speech from the analysis. To validate revealed frames, we manually analyze samples of articles from the extracted clusters. Findings of our tests indicate that when following the proposed feature selection approach, the resulting clusters more accurately discriminate between articles with a different framing. We discuss the methodological and theoretical implications of our findings.

Keywords

cluster analysis, news framing, sentiment analysis, information retrieval

News media can shape public opinion regarding an issue by emphasizing some elements of the broader controversy over others (Jasperson, Shah, Watts, Faber, & Fan, 1998; Shah, Watts, Domke, & Fan, 2002). Research shows that aspects of an issue, which are more salient in the media, cause individuals to focus on these aspects when constructing their opinions. Shah, Watts, Domke, and Fan (2002), for example, showed that public approval of President Clinton depended on whether news coverage during the Lewinsky sex scandal focused on the sexual nature of Clinton's indiscretion or the attacks of Republicans on Clinton's behavior. Furthermore, Sniderman, Brody, and Tetlock (1991) found that a majority of the public supports the rights of a person with HIV when the role

[1] University of Amsterdam, Amsterdam, the Netherlands

Corresponding Author:
Bjorn Burscher, University of Amsterdam, Nieuwe Achtergracht 166, Amsterdam 1018 WV, the Netherlands.
Email: b.burscher@uva.nl

of civil liberties is stressed in the news and supports mandatory testing when the importance of public health is stressed. In the literature, this phenomenon is referred to as emphasis framing.

In this article, we introduce and evaluate a method to automatically analyze emphasis framing in news coverage. We use this method to identify a set of news frames within the nuclear power debate and study developments in the frames' prevalence and tone over time. Therefore, this study fits into a broader line of research investigating the use of automated content methods in social science research (Grimmer & Stewart, 2013).

More specifically, we apply a combination of *k*-means cluster analysis and automatic sentiment analysis. Cluster analysis can be used to group articles according to their word use. As articles, which contain the same words, most likely also stress the same elements of a controversy, this technique can reveal groups of articles with a similar framing. Sentiment analysis is a way to automatically determine the polarity of the tone of an article. Together, these techniques provide a powerful tool to study dynamics in the news framing of social and political issues. To our knowledge, cluster and sentiment analysis have not been combined before in framing research.

Furthermore, we explore a novel way of performing cluster analysis such that the resulting clusters more accurately differentiate between articles with a different framing. To do so, we apply natural language processing to select parts of a news article, which we consider highly relevant to capture the meaning of frames, and only use these parts to present articles in the analysis. In our approach, we abstained from human involvement like defining frame elements (e.g., Motta & Baden, 2013) or manually (pre)coding data (e.g., Matthes & Kohring, 2008). This has the advantage that the framing analysis is largely automated and mostly unaffected by potential biases of human researchers and/or coders.

In order to validate results of our analyses, we conduct a manual content analysis. We conclude that the combination of cluster and sentiment analysis can be used to identify and code emphasis frames in news coverage automatically and validly. We discuss extensively the theoretical and methodological implications of our findings.

The Nuclear Power Debate

Our study addresses *emphasis framing* (e.g., Chong & Druckman, 2007)—a rather broad form of news framing, which is particularly prominent in the field of communication science. Throughout this article, we define framing as *emphasis in salience of some elements of a story above others* (e.g., De Vreese, 2005; Nelson, Clawson, & Oxley, 1997). In order to investigate the application of cluster and sentiment analysis to framing research, we must choose an issue to study. Ideally, this would be an issue that has created ample (controversial) news coverage in the past and which has been extensively studied in framing research before. Such a case would allow us to compare the frames we find by means of cluster analysis to frames that have been identified in previous studies, by means of different methods. The nuclear power debate provides such a case. Various studies have analyzed the nuclear power debate in the past 50 years (e.g., Bickerstaff, Lorenzoni, Pidgeon, Poortinga, & Simmons, 2008; Gamson & Modigliani, 1989; Nisbet, 2009). In Table 1, we provide an overview of nuclear power frames. To create this overview, we reviewed the most-cited journal articles[1] that study and/or discuss the media framing of nuclear power.

Frame Analysis

Frame analysis requires (1) the identification of frames that news media focus on when covering an issue and (2) coding the presence of these frames in news articles (e.g., Jasperson et al., 1998). Traditionally, scholars identify emphasis frames by qualitatively analyzing rather small samples of articles (e.g., Simon & Xenos, 2005). Afterward, to measure their usage, each frame

Table 1. Nuclear Power Frames in the Literature.

- Risks of nuclear weapon development[a,b]
- Health and environmental risks of radioactive waste[a,b,c,d]
- Social progress and economic development due to nuclear power usage[b,e,f]
- Terrorism threats and risks of nuclear accidents[b,c,d,e,f]
- Economic risks of nuclear power production, not cost-effective[b,f]
- Nuclear power to cut greenhouse gas emissions and prevent climate change[b,d]
- Nuclear power to satisfy energy demands and provide energy independence[b,c,e,f]
- Renewable energies as alternative to nuclear power[b,e,f]

[a]Joppke (1991). [b]Culley, Ogley-Oliver, Carton, & Street (2010). [c]Pidgeon, Lorenzoni, & Poortinga (2008). [d]Bickerstaff Lorenzoni, Pidgeon, Poortinga, & Simmons (2008). [e]Gamson & Modigliani (1989). [f]Nisbet (2009).

is operationalized—either in the form of indicator questions in manual content analysis (Semetko & Valkenburg, 2000) or search strings in automatic content analysis (Shah et al., 2002; Ruigrok & Van Atteveldt, 2007).

Alternatively, frames can be identified by means of statistical analysis. The most basic approach is to interpret word co-occurrences. Hellsten, Dawson, and Leydesdorff (2010), for example, plotted cosine distances between words in a network graph and then interpreted agglomerations of words within the network as frames. More sophisticated approaches applied either factor analysis (Motta & Baden, 2013; Van der Meer & Verhoeven, 2013) or cluster analysis (Matthes & Kohring, 2008; Miller, 1997). Factor analysis describes variability among observed variables in terms of a potentially lower number of unobserved variables, which can be interpreted as frames.

Cluster analysis groups a set of articles in such a way that articles in the same group are more similar to each other than to those in other groups (Kaufman & Rousseeuw, 2009). In other words, based on the similarity of articles, a number of clusters are created and each article is assigned to one cluster. The clusters present groups of articles with a different framing. By interpreting the most prototypical words of articles from each cluster, one can infer frames. Cluster analysis results in a classification model, which can be used to automatically code future articles according to the created cluster structure. The method can thus be used for further analyses: One can, for example, easily compare the popularity of different frames over time and across news sources. The assignment of articles to clusters is a critical difference between factor analysis and cluster analysis. Factor analysis reduces the dimensionality of a data set and provides information about how each factor corresponds to the original variables (e.g., words in the corpus) but does not classify the articles into groups.

In this study, we use cluster analysis to identify and code emphasis frames in news coverage about nuclear power from the past 20 years. Furthermore, we apply automatic sentiment analysis to analyze the tone of coverage. This allows us to study dynamics in the prevalence of different frames as well as dynamics in the tone of news articles containing a specific frame over time. So far, scholars have not combined cluster analysis and sentiment analysis to identify frames. We expect that the analysis of tone improves the interpretation of clusters as emphasis frames, because earlier research has shown that news coverage of nuclear power generally focuses on benefits or risks of its usage (Gamson & Modigliani, 1989). Moreover, frames often contain moral evaluations of policy issues (Semetko & Valkenburg, 2000). All in all, we present a method to show how the portrayal of an issue changes over time—in terms of topical elements that are emphasized and in terms of their valence.

To validate this automatic analysis, we conduct a manual content analysis for a sample of articles and compare automatic codings to manual codings. In addition, we compare outcomes of the cluster analyses to outcomes of previous studies that investigated the framing of the nuclear power debate (e.g., Gamson & Modigliani, 1989). This leads to the following research question: To what extent can cluster analysis be used to infer emphasis frames from a collection of issue-specific news articles?

By answering this question, we can determine the ability of cluster and sentiment analysis to identify and code emphasis frames in future research and we can draw conclusions about whether cluster analysis leads to similar frames as manual approaches. To our knowledge, previous studies have not explicitly cross-validated the use of statistical frame identification with manual approaches.

Building Blocks of Frames

In cluster analysis, the quality of resulting clusters depends on the selection of document features (Kaufman & Rousseeuw, 2009). There are various document features that might be used to compare two texts—but not all are important for the classification of interest. Some features may be redundant or irrelevant and others can misguide results of the cluster analysis. Various articles (e.g., Dy & Brodley, 2004; Gnanadesikan et al., 1995; Hatzivassiloglou, Gravano, & Maganti, 2000) have studied the question of which set of document features is most useful for several classification tasks (e.g., topic or sentiment). Among others, scholars selected words based on their frequency, part of speech, or position in the document. Furthermore, word features have been enriched by adding semantic features using Wikipedia (Hu, Zhang, Lu, Park, & Zhou, 2009) or WordNet (Sedding & Kazakov, 2004).

In statistical frame analysis, in order to find clusters (or factors) that discriminate between different frames, one must represent documents in terms of features that are indicative of such frames. According to Entman (1993, p. 52), news frames manifest themselves in certain text attributes as "the presence or absence of certain keywords, stock phrases, (and) stereotyped images." Therefore, we used word frequencies as features in our cluster analyses, which is called the "bag-of-words" approach (e.g., Hellsten, Dawson, & Leydesdorff, 2010; Miller, 1997). This has two advantages: First, using words is highly reliable, because words are manifest features (Riff, Lacy, & Fico, 2014) and consequently, frame analysis becomes a replicable process that is unlikely to be biased by the subjective input of individual researchers. Second, it is cost-efficient, because no manual analysis is involved.

The key issue, however, is construct validity: To what extent do word-based clusters actually reflect different emphasis frames? In the literature, this is a highly debated question. On the one hand, words are widely used as features in statistical frame analysis. On the other hand, critiques have repeatedly objected to its use (Carragee & Roefs, 2004; Hertog & McLeod, 2001). The main point of criticism is that not all words are equally important to a news frame. As Cappella and Jamieson (1997) put it, considering any production feature of verbal or visual texts as a candidate for news frames is a too broad view. As a response, scholars started using higher level frame elements as features (e.g., Matthes & Kohring, 2008; Motta & Baden, 2013). Matthes and Kohring (2008), for example, used Entman's popular operational definition of news frames and manually coded all articles for *problem definitions*, *causal interpretations*, *moral evaluations*, and/or *treatment recommendations* (Entman, 1993). Afterward, the authors used these frame elements as features in a cluster analysis.

Using higher level frame elements as features has brought significant advancements to statistical frame analysis, because such features are generally more conclusive building blocks of frames and, consequently, lead to a higher construct validity when identifying frames. However, as the used frame elements are usually issue-specific, they must be defined and coded individually before each analysis. Our aim, in contrast, is to explore a way in which we can improve statistical frame analysis but keep the analysis as inductive as possible without relying on a priori made decisions on the side of researchers or human coders. For doing so, we apply natural language processing to select such parts of a news article, which we consider highly relevant to capture the meaning of emphasis frames and only use these parts as features.

First, we only use words from the headline and the lead as features. Generally, news stories present information in terms of relative importance (Poettker, 2003). This structure is called the inverted

pyramid style. We infer from this style that the article's dominant perspective on the issue is presented at the beginning. Pan and Kosicki (1993) argued the following: "A headline is the most salient cue to activate certain semantically related concepts in readers' minds; it is thus the most powerful framing device of the syntactical structure. A lead is the next most important device to use. A good lead will give a story a newsworthy angle" (p. 59). Similarly, Tankard (2001) counts headline and lead as two important framing mechanisms. We expect that only using headline and lead as features leads to clusters that more clearly differentiate between distinct emphasis frames. This is because other elements in the remaining paragraphs of an article, which do not address the dominant frame, would act as noise in the analysis.

Related research on topic clustering has shown that giving higher weight to the title of a news article can increase the accuracy of topic clusters, because the title is more representative of the topic than the main text. In their experiments, Banerjee et al. (2007) obtained best results by doubling the weights of the terms appearing in the title of a given news article. Similarly, Bouras and Tsogkas (2012) increased the weights of terms that also appeared in the title of an article when analyzing topic clusters of news articles. We expect similar effects for frame clusters.

Second, we conduct part-of-speech tagging (Toutanova, Klein, Manning, & Singer, 2003) to select words that are a noun, an adjective, or adverb. In linguistics, part-of-speech tagging is the process of marking up a word in a text as corresponding to a particular part of speech. We believe that words from the selected classes (nouns, adjectives, and adverbs) are most indicative of frames. This is because other word classes, like verbs, conjunctions, or pronouns, are much less likely than the selected classes to add meaning to a frame. Previous research has shown that giving higher weights to nouns than other word classes can increase the quality of topic clusters (e.g., Bouras & Tsogkas, 2012; Hatzivassiloglou, Gravano, & Maganti, 2000).

Third, we apply named-entity recognition (Nadeau & Sekine, 2007) to remove all names of persons, organizations, and locations as well as times and dates. Names of countries and organizations, for example, refer to very specific events, while frames are more abstract semantic concepts. Therefore, it is more likely to get clusters, which actually discriminate between emphasis frames, when we remove named entities from the feature space. To our knowledge, this has not been tested before in document clustering.

When representing articles in the cluster analysis, we only use the above-mentioned parts of each article as features and ignore all other words. We call this the *selection approach*. In order to see whether this way of selecting features improves the validity of the cluster analysis, we conduct a baseline analysis where we use all words from each article as features (*baseline approach*). We compare cluster centers in the selection approach with clusters centers in the baseline approach. Furthermore, we conduct a manual content analysis to compare the accuracy of frame codings in both approaches. This leads to the following research question: To what extent does selecting frame-related document features improve the construct validity and coding accuracy of statistical frame analysis?

In sum, contrasting the approaches aims at finding a way of representing news articles in terms of features that are highly indicative of frames. We expect that selecting frame-specific features (selection approach) does a better job in discriminating between emphasis frames than using all words as features (baseline approach).

Data and Method

Data

Our data consisted of English-language news articles covering the issue of nuclear power, which were published in *The New York Times, The Washington Post,* or *The Guardian* between 1992 and

2013. We used LexisNexis to search all three sources for articles that contain the key words "nuclear power" or "nuclear energy" at least 2 times in total and at least once in the headline or lead. By applying these rather strong restrictions, we made sure that nuclear power actually is the main topic of the article. This led to 4,286 articles, which we used in the analyses.

Automatic Content Analysis

Based on this collection of news articles, we created two data sets—one for the baseline approach and one for the selection approach. In both data sets, we used all 4,286 articles and applied the following preprocessing steps. We converted all words to their lemmas (Bird, 2006) and removed numbers and common English stop words. Furthermore, we removed words that appeared in less than five documents or in more than 40% of all documents. Due to their frequency of use (very high or very low), such words do not differentiate well between clusters of news articles. As explained in the previous section, in the selection approach data set, we also removed (a) words that did not appear in the title or lead; (b) words with a part-of speech other than noun, adjective, or adverb; and (c) names of persons, organizations, and countries. For all of the above-mentioned steps, we used the Python natural language toolkit (Bird, 2006).

Afterward, we created document vectors with TF.IDF weighted word frequencies (Manning, Raghavan, & Schütze, 2008) for news articles in both data sets. Each word was assigned the number of times it occurs in the document (TF), weighted by the inversed frequency of articles in the data set containing the word (IDF). The idea behind TF.IDF weighting is to evaluate the power of a word to discriminate between articles. Rare words are assumed to be more discriminating and, therefore, are assigned higher weight. We standardized the document vectors using L2 normalization (Ng, 2004).

To reveal clusters from our data sets, we applied *k*-means clustering—a centroid-based clustering technique, where the number of clusters (k) must be specified a priori (Hartigan & Wong, 1979). Given a set of articles ($x_1, x_2, \ldots, x_n$), where each article is a d-dimensional vector, *k*-means clustering separates the n articles into k ($\leq n$) clusters $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares. More formerly, it aims at finding:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} x - \mu_i{}^2$$

where μ_i is the mean of points in S_i.

Each cluster is represented by a cluster center, which is described by the mean of the articles in the cluster. The algorithm defines the cluster centers and assigns each article to the cluster for which its distance to the cluster center is the smallest. We conducted separate cluster analyses for the baseline and the selection approach and used the cluster center vectors in order to identify emphasis frames in both approaches. In doing so, we listed for each cluster center the 15 document features with the highest means, that is, the most prototypical words for the cluster. Then, we gave each cluster center a frame label based on these 15 words.

A common technique to select the number of clusters (k) is the "elbow method." We repeatedly run the analysis with different numbers of clusters (1–15) and added the amount of explained variance for each value to a scree plot (see Figure 1). Because the scree plot depicted an elbow at seven clusters in the baseline approach analysis, we decided to use a seven-cluster solution. In order to make both analyses more comparable, we also used a seven-cluster solution in the selection approach analysis. Our implementation of *k*-means clustering makes use of the *mini-batch k*-means algorithm (Sculley, 2010) and the *k-means++* optimization (Arthur & Vassilvitskii, 2007).[2] We

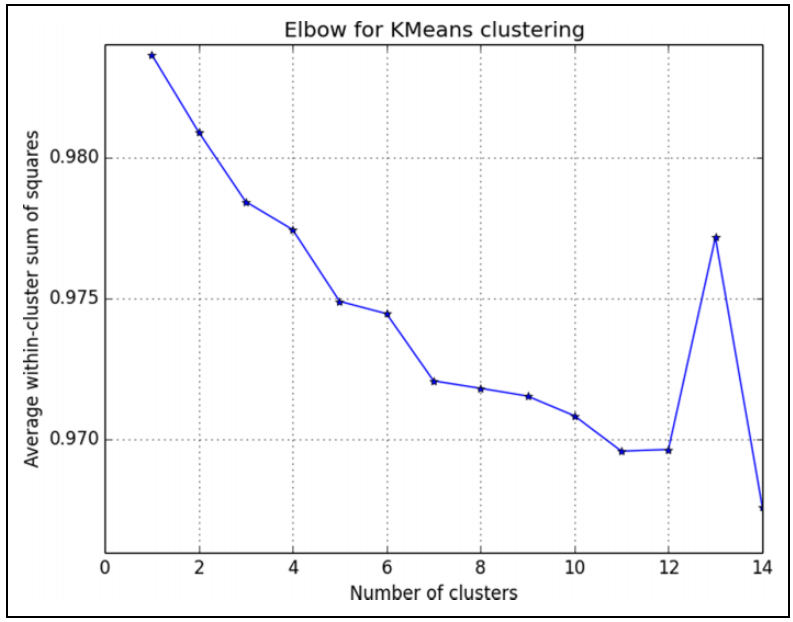

Figure 1. Scree plot of explained variance for baseline approach.

used the scikit-learn machine learning library in Python for document vectorization and the cluster analyses (Pedregosa et al., 2011).

Finally, we applied the SentiWords[3] tool to automatically code the tone of articles in the selection approach data set. SentiWords is a lexical resource containing roughly 155,000 words associated with a sentiment score between −1 (*negative*) and 1 (*positive*). Scores are learned from SentiWordNet and represent state-of-the-art computation of words' prior polarities (Baccianella, Esuli, & Sebastiani, 2010). See Guerini, Gatti, and Turchi (2013) for information about the method used to build SentiWords and Warriner, Kuperman, and Brysbaert (2013) for a detailed description of the used data set. We annotated each word from the title and lead of all articles with the corresponding sentiment scores from the SentiWords lexicon. Then, we computed the mean of sentiment scores over all annotated words in each article and used it as a summary score for the article's tone. Words from the articles, which were not included in SentiWords, received a sentiment score of zero.

Manual Content Analysis

We also conducted a manual content analysis to test whether the articles in each frame cluster actually contained the predicted frame. For the baseline and the selection approach, we sampled a random subset of 15 articles from each cluster and asked human coders to indicate the most relevant emphasis frame per article. Because we had two approaches with seven clusters each, 210 articles were manually coded. Per article, coders could choose one frame from a list containing all unique frames that we identified for the corresponding approach. In the Results section, we describe these sets of frames more closely. Additionally, coders could code an article as "containing none of the listed frames/not primarily dealing with nuclear power." We used two trained coders, who were fluent in English. In order to assess intercoder reliability, both coders coded 15% of the articles ($N = 32$). Krippendorff's α for intercoder agreement is equal to .82.

Table 2. Clusters Baseline Approach.

B1	B2	B3	B4	B5	B6	B7
British	Commission	Indian	Carbon	Iran	Japan	India
Company	Chernobyl	Point	Gas	Iranian	Fukushima	Korea
Pound	Safety	Entergy	Emission	Russia	Tokyo	North
EDF	Waste	County	Wind	Weapon	Radiation	Treaty
Station	Company	Emergency	Climate	Uranium	Tepco	Weapon
Industry	Station	Westchester	Electricity	Program	Japanese	China
Price	Utility	Buchanan	Coal	Tehran	Water	Test
Cost	Official	Plan	Industry	Enrichment	Tsunami	Pakistani
Share	Site	Commission	Oil	Bushehr	Daichi	United
Million	Fuel	Siren	Station	Russian	Accident	Korean
Billion	Radiation	Steets	Renewable	United	Earthquake	Ban
Electricity	People	Federal	Cost	Sanction	Disaster	Official
Britain	Million	Evacuation	Waste	International	Safety	Agreement
BNFL	Industry	Official	Fuel	Official	Radioactive	Administration
Market	Public	Hudson	Renewables	Ahmadinejad	Worker	South
$N = 558$	$N = 1{,}918$	$N = 250$	$N = 648$	$N = 233$	$N = 383$	$N = 296$

Results

Baseline Approach

We performed two *k*-means cluster analyses: one in which we used all words of each article as features (baseline approach) and one in which we used selected parts (selection approach) of each article as features. Per analysis, we looked at the 15 features with the highest means for each of the seven cluster centers to infer emphasis frames. See Tables 2 and 3 for an overview of the cluster centers for each approach.

When using the baseline approach, we found multiple clusters that refer to the same element of the nuclear power controversy but relate to different geographical contexts. Clusters B5 and B7 are good examples of this phenomenon. Cluster B5 refers to nuclear power and the issue of weapon development in Iran. Cluster B7 also refers to nuclear power and weapon development, but in India and North Korea. B2 and B6 are another pair of examples, both clusters refer to safety issues and radiation risks of nuclear accidents. However, Cluster B2 does so in the context of the Chernobyl catastrophe and Cluster B6 in the context of the Fukushima disaster. Furthermore, Cluster B3 refers to a very specific incident, instead of a more general emphasis frame—emergency evacuations of the Indian Point nuclear plant in Buchanan, New York. We can explain this clustering around specific events by the fact that the centers of the mentioned clusters mainly contain names of countries and organizations (e.g., Iran, Bushehr, India, and Fukushima). This indicates that the clusters do not primarily discriminate between distinct emphasis frames but between geographic contexts and particular incidents.

Nonetheless, several clusters uniquely refer to general elements of the nuclear power controversy. Cluster B4, for example, clearly refers to the impact of nuclear power on the climate and Cluster B1 refers to economic aspects of nuclear power usage. Cluster sizes are fairly unequally distributed ($SD = 552.5$) with Cluster B2 as big as all other clusters combined, suggesting a residual category that includes articles not be properly assigned to other clusters. Overall, we identified five unique frames here, which are listed in Table 4. We collapsed Clusters B5 and B7 (weapon development) as well as Clusters B2 and B6 (nuclear safety and accidents), because they referred to identical frames.

Table 3. Clusters Selection Approach.

S1	S2	S3	S4	S5	S6	S7
Station	Energy	Company	Weapon	Commission	Fuel	Reactor
State	Gas	Government	Program	Regulatory	Uranium	Radiation
Mile	Government	Price	State	Federal	Waste	Radioactive
First	Oil	Industry	President	Reactor	Plutonium	Accident
Official	Source	Pound	Country	Safety	Radioactive	Safety
Government	Renewable	Cost	Official	Regulator	Spent	Water
Plan	Climate	Reactor	Agreement	State	Reactor	Disaster
Security	Electricity	Electricity	Test	License	Rod	Leak
People	Policy	Share	Energy	Agency	State	Level
World	Emission	Plan	Foreign	Company	Enrichment	Worker
Last	Change	State	Treaty	Official	Storage	Exposure
Federal	Carbon	Utility	Nation	Problem	Company	Earthquake
Reactor	Coal	Generation	International	Utility	Material	Official
Attack	Minister	Last	World	Attack	Site	Station
Former	Generation	Energy	Uranium	Mile	Government	Operator
$N = 1{,}296$	$N = 645$	$N = 609$	$N = 568$	$N = 548$	$N = 328$	$N = 292$

Table 4. Identified Frames Baseline and Selection Approach.

Baseline approach	
Frame 1	Economic aspects of nuclear power production
Frame 2	Safety of nuclear plants, nuclear waste, nuclear power accidents and radiation risks
Frame 3	Nuclear power and weapon development
Frame 4	Role of nuclear power in electricity production and effects on climate change
Frame 5	Evacuation of nuclear reactors
Selection approach	
Frame 1	*Safety* of nuclear plants
Frame 2	Role of nuclear power in electricity production and effects on *climate* change
Frame 3	*Economic* aspects of nuclear power production
Frame 4	Nuclear power and *weapon* development
Frame 5	Processing of nuclear materials and nuclear *waste*
Frame 6	Nuclear power *accidents* and radiation risks

Selection Approach

When using the selection approach, we got a clearer cluster structure (see Table 3). Six of the seven clusters have coherent and unique cluster centers, all of which refer to distinct elements of the nuclear power controversy. There is little overlap as regards content between the clusters, which means that different clusters do not refer to the same emphasis frame. The cluster centers are, furthermore, easy to interpret, because they contain mostly substantial words and no names of places, persons, or organizations.

The primary question is whether we found different frames with this representation. On the one hand, some clusters are exactly the same as in the baseline approach. Two examples are Cluster S3, which deals with economic aspects of nuclear power, and Cluster S2, which deals with the effects of nuclear power on the climate.

On the other hand, we also found clusters that did not appear in the baseline approach. In the baseline approach, Cluster B2 refers to safety issues, nuclear accidents, and nuclear waste altogether.

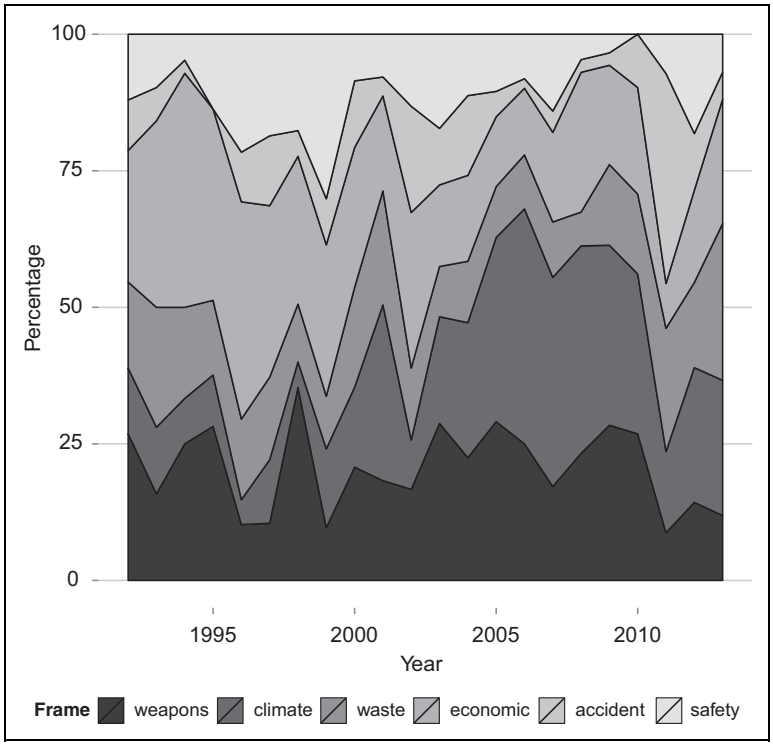

Figure 2. Stacked area plot of frame prevalence between 1990 and 2013.

In contrast, in the selection approach, we found separate clusters for safety issues (S5), nuclear accidents (S7), and nuclear waste processing (S6). This might be explained by the fact that we only used the title and lead of each news article as features here. When different elements of an issue are often referred to in the same article and one uses all parts of the article as features, it is likely that the elements are grouped together in one cluster. However, it is less likely that all elements are mentioned together in the title or lead. The selection approach thus provides a more nuanced grouping of articles around unique elements of the controversy.

Furthermore, compared to the baseline approach, clusters are more equally distributed with regard to size ($SD = 306.6$). Again, one cluster (S1) is significantly larger than the average cluster size. Overall, we identified six unique frames here, which are listed in Table 4. We collapsed Clusters S1 and S5, which both refer to safety issues of nuclear power.

Validation Analysis

We conducted three additional analyses. First, we calculated Krippendorrf's α as a measure of agreement between computer-based and human frame codings. Higher values of Krippendorrff's α indicate higher agreement between humans and the computer. As shown in Table 4, Krippendorrff's α is equal to .52 for the baseline approach and .71 for the selection approach. This shows that when using the selection approach, significantly more articles actually contained the predicted emphasis frame. The selection approach thus leads to more accurate codings of frames.

Second, we plotted the prevalence of frames from the selection approach over time. Figure 2 shows that frame prevalence varies considerably over time. Several peaks in the graph correspond to real-life events. We see, for example, a peak in the weapon development frame (Frame 4) around

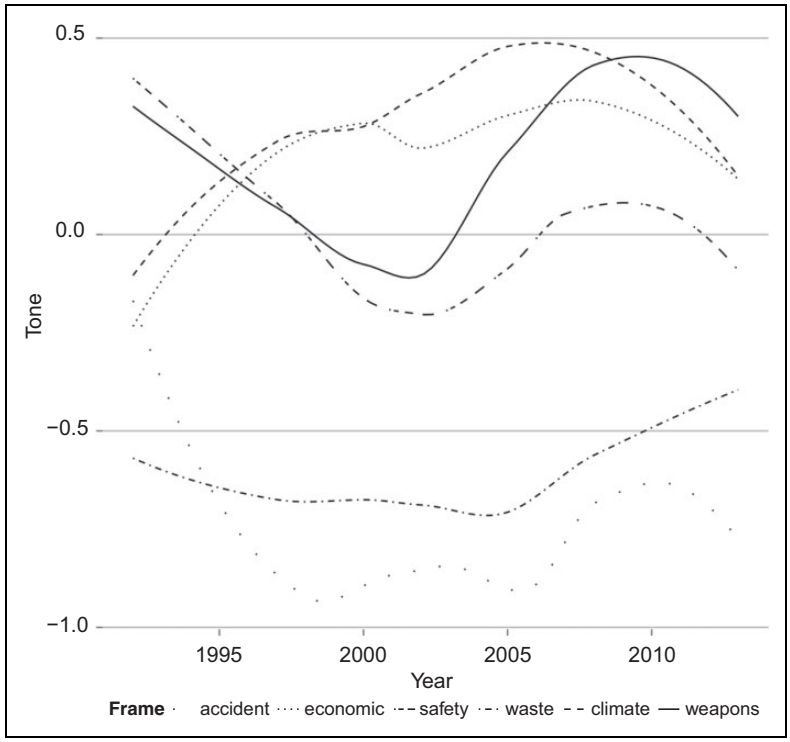

Figure 3. Tone of articles from each frame cluster between 1990 and 2013.

1998, when the Indian government conducted the Pokhran 2 nuclear bomb tests. Moreover, we see a peak in the prevalence of the accidents and radiation risks frame (Frame 6) around 2011, followed by a peak in the safety frame (Frame 1) shortly afterward. This is most probably related to the Fukushima disaster and debates about the safety of nuclear power it caused in the media. All in all, these findings confirm our conclusion that we identified a valid set of emphasis frames.

Third, we analyzed the tone of articles in the selection approach data set. In Figure 3, we show overtime variation in tone for articles from each frame cluster. The graph indicates clear differences in tone across clusters. Articles focusing on the processing of nuclear waste and materials (Frame 5) or accidents and radiation risks (Frame 6), for example, are much more negative than articles that focus on the effects of nuclear power on climate change (Frame 2) or economic aspects of nuclear power (Frame 3). This is in line with the literature, where the former two frames are depicted as risk frames and the latter as opportunity frames (Gamson & Modigliani, 1989). Furthermore, we also see within-cluster variation in tone. Articles in the weapon development cluster (Frame 4) as well as articles in the economic aspects cluster (Frame 3) become much more positive over time. Again, the patterns of variation correspond to actual events. In the aftermath of the Fukushima disaster, news coverage not only focused much more on safety issues (Frame 1) of nuclear plants, but articles focusing on safety issues also became more negative.

Discussion

In this article, we applied cluster and sentiment analysis to identify and code news frames. Statistical frame analysis has several advantages over holistic approaches, all of which can guide future framing research. First, it is more cost-efficient, because no manual content analysis is required. Second, it

scales better to big data sets, which become increasingly available as the use of social media and the availability of digital news content increases. Third, it reduces risks of bias caused by human perceptions and interpretations (Matthes & Kohring, 2008). Cluster analysis, in particular, has the advantage that it automatically classifies articles into groups and that it provides a model for doing so in future research. This is a more efficient and sophisticated way of coding documents for frames, as compared to manually creating coding rules, either on the basis of results from holistic analyses or based on key words from a factor analysis.

Although statistical techniques are widely used among communication scholars to identify news frames, they are criticized for not being able to do so in a conceptually valid manner (Carragee & Roefs, 2004; Hertog & McLeod, 2001). For this reason, we explored a way of improving the cluster analysis of frames such that the resulting clusters more closely resemble emphasis frames. We found that when using all words of an article as features, clusters are often centered on individual actors and events instead of more abstract elements of the controversy. In addition, different elements are referred to in the same cluster, and different clusters overlap as regards the elements they refer to.

In contrast, when using only words from the title and lead as features and when removing all named entities from the feature space, clusters more accurately discriminate between distinct elements of the controversy. In addition, when using this selection of highly indicative features, more articles get accurately coded for frames. Generally, we conclude that the vast majority of articles are correctly classified for the frame they contain when selecting features. In other words, most articles within a frame cluster actually contain the predicted frame.

The frames we identified by means of cluster analysis closely match frames that scholars found in earlier research, applying holistic methods (Table 1). From this, we conclude that our method is suited to identify frames. However, the frames we found are less detailed interpretations of the nuclear power controversy as those discussed in holistic studies. First, frames from holistic studies often contain valence elements: They focus on either positive or negative aspects of the issue. The valence of frames is not properly revealed by our cluster analysis, as the cluster centers mostly describe topical aspects. However, when adding sentiment analysis, we can reproduce the valence of holistically identified frames in most cases.

Next to valence, manually identified frames often express causal relations. These are not directly visible in our cluster centers and sentiment scores. We believe that, based on plain word features, a cluster analysis cannot reveal complex semantic and logical relationships like causality. It should be a challenge for future research to improve automatic frame clustering such that causality can be accounted for. In computational linguistics research, this problem has been addressed (e.g., Girju & Moldovan, 2002), but it is still difficult to automatically reveal the exact relation between two concepts in a sentence.

There are several limitations to this research. First, we only focused on three newspapers from two countries. Second, it is challenging to validate the found frames, as there is no ground truth about what is a frame and what is not a frame. Similarly, there is no true reference list of frames that are used in the nuclear power debate. Third, *k*-means clustering is a nondeterministic method and, therefore, results slightly vary each time the analysis is conducted. However, after repeated analyses, we observed comparable results in the sense of similar frames in the majority of the runs. Finally, it might be misleading to argue that our approach is completely inductive, because we interpret the words in the cluster centers as frames. However, this interpretation is very straightforward. When applying the selection approach, for the vast majority of clusters, the words from the cluster centers clearly indicate one (topical) element of the nuclear power debate.

We believe that our approach to statistical frame analysis facilitates the use of mixed-methods designs (e.g., the combination of panel surveys and content analysis) in framing research (e.g., de Vreese, 2012), because it is very cost-efficient. Furthermore, it allows for increases in the scale of frame analysis. This allows scholars to reliably study developments in framing over long time

frames and between different sources in an efficient manner (e.g., Vliegenthart & Roggeband, 2007). This approach is useful for certain applications in particular, including studying the mapping of topical aspects of social and political issues, with an interest in long-term dynamics of how issues are presented in the news. If one is interested in getting a highly detailed and in-depth account of single events that span a limited amount of time, traditional (holistic) approaches might be a better choice. Furthermore, the generalization of this approach is limited to issues that receive a certain amount of coverage and which are sufficiently contested in news coverage.

Finally, findings of this study suggest implications for framing effects on public opinion. Since we identify a different set of frames when only looking at the title and lead of articles, this could have implications if people only read the headline or lead of news stories. Framing research has shown that exposure to different news frames can affect peoples' opinions about an issues and also their behavior (e.g., Nelson & Oxley, 1999; van Spanje & de Vreese, 2014). Furthermore, title and lead are considered the most important framing devices of a news story (Tankard, 2001). Therefore, one can conclude that framing effects might be stronger among people, who only read title and lead of a news story.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Articles cited more than 30 times according to Google Scholar.
2. In each cluster analysis, we run the *k*-means algorithm 10 times with different centroid seeds in each run. The final results were the best output of the 10 consecutive runs in terms of explained variance.
3. https://hlt.fbk.eu/technologies/sentiwords

References

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1027–1035). Society for Industrial and Applied Mathematics, 2007.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Language Resources and Evaluation*, *10*, 2200–2204.

Banerjee, S., Ramanathan, K., & Gupta, A. (2007). Clustering short texts using Wikipedia. In W. Kraaij & A. de Vries (Eds.), *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and development in information retrieval* (pp. 787–788). New York, NY: ACM.

Bickerstaff, K., Lorenzoni, I., Pidgeon, N. F., Poortinga, W., & Simmons, P. (2008). Reframing nuclear power in the UK energy debate: Nuclear power, climate change mitigation and radioactive waste. *Public Understanding of Science*, *17*, 145–169.

Bird, S. (2006). NLTK: The natural language toolkit. In A. Moschitti (Ed.), *Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 69–72). Stroudsburg, PA: Association for Computational Linguistics.

Bouras, C., & Tsogkas, V. (2012). A clustering technique for news articles using WordNet. *Knowledge-Based Systems*, *36*, 115–128.

Cappella, J. N., & Jamieson, K. H. (1997). *Spiral of cynicism: The press and the public good*. New York, NY: Oxford University Press.

Carragee, K. M., & Roefs, W. (2004). The neglect of power in recent framing research. *Journal of Communication*, *54*, 214–233.

Chong, D., & Druckman, J. N. (2007). Framing theory. *Annual Review of Political Science*, *10*, 103–126.

Culley, M. R., Ogley-Oliver, E., Carton, A. D., & Street, J. C. (2010). Media framing of proposed nuclear reactors: An analysis of print media. *Journal of Community & Applied Social Psychology*, *20*, 497–512.

De Vreese, C. H. (2005). News framing: Theory and typology. *Information Design Journal and Document Design*, *13*, 51–62.

De Vreese, C. H. (2012). New avenues for framing research. *American Behavioral Scientist*, *56*, 365–375.

Dy, J. G., & Brodley, C. E. (2004). Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, *5*, 845–889.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*, 51–58.

Gamson, W. A., & Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, *95*, 1–37.

Gnanadesikan, R., Kettenring, J. R., & Tsao, S.-L. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification*, *12*, 113–136.

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

Guerini, M., Gatti, L., & Turchi, M. (2013). Sentiment analysis: How to derive prior polarities from SentiWordNet. In A. Moschitti (Ed.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)* (pp. 1259–1269). Stroudsburg, PA: ACL.

Girju, R., & Moldovan, D. I. (2002) Text mining for causal relations. In S. Haller & G. Simmons (Eds.), *Proceeding of the FLAIRS Conference* (pp. 360–364). Palo Alto, CA: AAAI Press.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, *28*, 100–108.

Hatzivassiloglou, V., Gravano, L., & Maganti, A. (2000). An investigation of linguistic features and clustering algorithms for topical document clustering. In N. Belkin, M.-K. Leong, & P. Ingwersen (Eds.), *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development information retrieval* (pp. 224–231). New York, NY: ACM.

Hellsten, I., Dawson, J., & Leydesdorff, L. (2010). Implicit media frames: Automated analysis of public debate on artificial sweeteners. *Public Understanding of Science*, *19*, 590–608.

Hertog, J. K., & McLeod, D. M. (2001). *A multiperspectival approach to framing analysis: A field guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hu, X., Zhang, X., Lu, C., Park, E. K., & Zhou, X. (2009). Exploiting Wikipedia as external knowledge for document clustering. In J. Elder & F. S. Fogelman (Eds.), *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 389–396). New York, NY: ACM.

Jasperson, A. E., Shah, D. V., Watts, M., Faber, R. J., & Fan, D. P. (1998). Framing and the public agenda: Media effects on the importance of the federal budget deficit. *Political Communication*, *15*, 205–224.

Joppke, C. (1991). Social movements during cycles of issue attention: The decline of the anti- nuclear energy movements in West Germany and the USA. *British Journal of Sociology*, *42*, 43–60.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. New York, NY: John Wiley.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, England: Cambridge University Press.

Matthes, J., & Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, *58*, 258–279.

Miller, M. M. (1997). Frame mapping and analysis of news coverage of contentious issues. *Social Science Computer Review*, *15*, 367–378.

Motta, G., & Baden, C. (2013). Evolutionary factor analysis of the dynamics of frames: Introducing a method for analyzing high-dimensional semantic data with time-changing structure. *Communication Methods and Measures*, *7*, 48–82.

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

Nelson, T. E., Clawson, R. A., & Oxley, Z. M. (1997). Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, *91*, 567–583.

Nelson, T. E., & Oxley, Z. M. (1999). Issue framing effects on belief importance and opinion. *Journal of Politics*, *61*, 1040–1067.

Ng, A. Y. (2004). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In C. Brodley (Ed.), *Proceedings of the twenty-first international conference on Machine learning* (p. 78). New York, NY: ACM.

Nisbet, M. C. (2009). Communicating climate change: Why frames matter for public engagement. *Environment: Science and Policy for Sustainable Development*, *51*, 12–23.

Pan, Z., & Kosicki, G. M. (1993). Framing analysis: An approach to news discourse. *Political Communication*, *10*, 55–75.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pidgeon, N. F., Lorenzoni, I., & Poortinga, W. (2008). Climate change or nuclear power—No thanks! A quantitative study of public perceptions and risk framing in Britain. *Global Environmental Change*, *18*, 69–85.

Poettker, H. (2003). News and its communicative quality: The inverted pyramid—When and why did it appear? *Journalism Studies*, *4*, 501–511.

Riff, D., Lacy, S., & Fico, F. (2014). *Analyzing media messages: Using quantitative content analysis in research*. New York, NY: Routledge.

Ruigrok, N., & Van Atteveldt, W. (2007). Global angling with a local angle: How us, British, and Dutch newspapers frame global and local terrorist attacks. *The Harvard International Journal of Press/Politics*, *12*, 68–90.

Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of communication*, *50*, 93–109.

Sculley, D. (2010). Web-scale k-means clustering. In M. Rappa & P. Jones (EDs.), *Proceedings of the 19th international conference on World Wide Web* (pp. 1177–1178). New York, NY: ACM.

Sedding, J., & Kazakov, D. (2004). WordNet-based text document clustering. In V. Pallotta & A. Todirascu (Eds.), *Proceedings of the 3rd workshop on robust methods in analysis of natural language data* (pp. 104–113). Stroudsburg, PA: ACL.

Shah, D. V., Watts, M. D., Domke, D., & Fan, D. P. (2002). News framing and cueing of issue regimes: Explaining Clinton's public approval in spite of scandal. *Public Opinion Quarterly*, *66*, 339–370.

Simon, A., & Xenos, M. (2000). Media framing and effective public deliberation. *Political Communication*, *17*, 363–376.

Sniderman, P. N., Brody, R. A., & Tetlock, P. E. (1991). *Reasoning and choice: Explorations in political psychology*. Cambridge, England: Cambridge University Press.

Tankard, J. W. (2001). The empirical approach to the study of media framing. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *Perspectives on media and our understanding of the social world* (pp. 95–106). Mahwah, NJ: Lawrence Erlbaum Associates.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of- speech tagging with a cyclic dependency network. In Hwee Tou Ng & Ellen Riloff (Eds.), *Proceedings of the 2003 Conference of the Association for Computational Linguistics on Human Language Technology* (pp. 173–180). Stroudsburg, PA: Association for Computational Linguistics.

Van der Meer, T. G. L. A., & Verhoeven, P. (2013). Public framing organizational crisis situations: Social media versus news media. *Public Relations Review*, *39*, 229–231.

van Spanje, J., & De Vreese, C. H. (2014). Europhile media and Eurosceptic voting: Effects of news media coverage on Eurosceptic voting in the 2009 European Parliamentary elections. *Political Communication*, *31*, 325–354.

Vliegenthart, R., & Roggeband, C. (2007). Framing immigration and integration Relationships between press and parliament in The Netherlands. *International Communication Gazette*, *69*, 295–319.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance *for* 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191–1207.

Author Biographies

Bjorn Burscher is a PhD candidate in the Department of Communication Science at the Amsterdam School of Communication Research, University of Amsterdam. He works on the INFINITI project, which develops and enables the exploitation of open source and open standards tools to support semantic search. He can be reached at b.burscher@uva.nl

Rens Vliegenthart is a full professor in communication science and the chair in media and organizations in the Department of Communication Science and at the Amsterdam School of Communication Research, University of Amsterdam. He can be reached at r.vliegenthart@uva.nl

Claes H. de Vreese is a professor and chair of political communication, and the director of the Program Group Political Communication & Journalism in The Amsterdam School of Communication Research (ASCoR) in the Department of Communication Science, University of Amsterdam. He can be reached at c.h.devreese@uva.nl