Taking stock of the toolkit: an overview of relevant automated content analysis approaches and techniques for digital journalism scholars

Boumans, J.W.; Trilling, D.

DOI
10.1080/21670811.2015.1096598

Publication date
2016

Document Version
Final published version

Published in
Digital Journalism

TAKING STOCK OF THE TOOLKIT
An overview of relevant automated content analysis approaches and techniques for digital journalism scholars

Jelle W. Boumans[1] and **Damian Trilling**[1]

When analyzing digital journalism content, journalism scholars are confronted with a number of substantial differences compared to traditional journalistic content. The sheer amount of data and the unique features of digital content call for the application of valuable new techniques. Various other scholarly fields are already applying computational methods to study digital journalism data. Often, their research interests are closely related to those of journalism scholars. Despite the advantages that computational methods have over traditional content analysis methods, they are not commonplace in digital journalism studies. To increase awareness of what computational methods have to offer, we take stock of the toolkit and show the ways in which computational methods can aid journalism studies. Distinguishing between dictionary-based approaches, supervised machine learning, and unsupervised machine learning, we present a systematic inventory of recent applications both inside as well as outside journalism studies. We conclude with suggestions for how the application of new techniques can be encouraged.

KEYWORDS: automated content analysis; computational social science; digital data; journalism studies; review

Introduction

The current digital age brings about substantial changes to the field of journalism studies. Evidently more and more researchers from other fields have started analyzing journalistic content. Computer scientists, for instance, analyze data from journalistic websites, social media, or blogs (e.g., Mishne 2007; Morgan, Zubair Shafiq, and Lampe 2013; Stieglitz and Dang-Xuan 2012). The availability and volume of digital journalistic content make it interesting material for various academic fields, from behavioral finance (Uhl 2014) to wildlife studies (Houston, Bruskotter, and Fan 2010). Studies of journalistic content in these fields typically benefit from advanced computer-assisted methods that find their origin in computer science.

Given computer scientists' expertise in dealing with digital data, taking stock of their techniques can be very rewarding for journalism scholars. Automated content analysis (ACA) can identify patterns in journalistic data that traditional analysis would

[1]Both authors contributed equally to this work.

Digital Journalism, 2016
Vol. 4, No. 1, 8–23, http://dx.doi.org/10.1080/21670811.2015.1096598

not, or only with greater effort (Flaounas et al. 2013). Additionally, it can provide "harder" evidence for what journalism scholars might already have suspected based on qualitative or small-scale quantitative research, help to sketch the bigger picture, and—last but not least—save time and money. Other fields testify to this potential: relevant applications of automated techniques are found in fields as various as linguistics (Schneider 2014), management studies (Illia, Sonpar, and Bauer 2014), behavioral finance (Uhl 2014), and political science (Grimmer and Stewart 2013). To a certain extent, journalism scholars do rely on the knowledge and insights from computer science. Examples range from assessing news formats (Sjøvaag and Stavelin 2012) to gender bias (Flaounas et al. 2013), and from sentiment analysis (Uhl 2014) to framing analysis (Hellsten, Dawson, and Leydesdorff 2009). But still, as noted in an earlier issue of this journal: "The introduction of computer science into the social sciences is still at an immature stage" (Flaounas et al. 2013, 102; similarly Freelon 2014, 71).

Becoming more familiar with the available ACA toolkit is increasingly necessary as our object of study changes. In addition to the traditional mass communication channels and their predictable, static content, journalism nowadays takes place via a variety of communication channels, including blogs, news aggregators, social media, apps, and news websites. It is not only the channels that have changed: technological advances have also created new quantitative forms including computer-assisted reporting, data journalism, and computational journalism (Coddington 2014). Yet, while scholars have paid ample attention to the increasing impact that computational applications have on the field of journalism (e.g., Carlson 2014; Fink and Anderson 2014; Flew et al. 2012), they rarely take advantage of the potential of these tools themselves.

Both the new channels as well as new journalistic practices require a critical inquiry into the methods that we use to study journalism. Traditional content-analytical tools generally cannot account for the dynamic and often interactive characteristics of the online news environment. In addition, the vast size of many digital journalism datasets make manual approaches unfeasible. Generally speaking, a discipline should strive for expanding its methodological toolkit when both the subject of study as well as the technological actuality changes. The journalism researcher's toolkit lags behind the state of the art of other fields that study journalistic content. To encourage exploring automated approaches, this article provides a systematic inventory of ACA approaches that have demonstrated their usefulness for the study of journalistic content.

Of course, automated methods also have drawbacks and this review will discuss them. Foremost, automated methods are not equivalent to manual methods. Because language is so multifaceted, automated methods will inevitably fall short on reducing a text to a model that represents the text in its entirety (Simon 2001). As Grimmer and Stewart (2013, 269) state: "All quantitative models of language are wrong—but some are useful." While human coders are not flawless either, they are generally better able to recognize the various meanings that words and phrases can have.

In the remainder of this article, we propose a classification of such techniques along a continuum ranging from deductive to inductive. Deductive techniques, as we propose to call them, define some coding criteria in advance (e.g., lists of relevant words), while inductive techniques rather seek to identify patterns in the data, without prescribing in advance what to look for. One could therefore also call the deductive techniques "top-down approaches" and the inductive techniques "bottom-up approaches."

Sorting Out the ACA Toolkit: From Deductive to Inductive

The approaches that are most easy and straightforward to apply are typically deductive. These are based on predefined categories or rules. The researcher has a large—theoretically motivated—say in deciding what content features the technique should extract. *Visibility analysis*, applied when the researcher is interested in how many times a specific actor or event occurs in the media, is a common example of a deductive type of research.

At the other end of the continuum, where we find the inductive techniques, one could say that it is the computer rather than the researcher which makes the decisions about what is meaningful in the dataset. Following the rationale that for some tasks, computers are better equipped than humans, it is up to the analytical technique to extract meaningful features from—often large—datasets. Implicit *framing analysis*, where the computer seeks patterns of co-occurring words that convey meanings in a collection of texts, is an example of such an application (Hellsten, Dawson, and Leydesdorff 2009). To structure our overview, we distinguish three general methodological approaches that can be arranged along this continuum: *counting and dictionaries*, *supervised machine learning*, and *unsupervised machine learning*. Figure 1 presents these approaches and provides examples of related research interests and statistical procedures.

Before we describe the methodological approaches and the types of problems they typically address, it is necessary to understand that most of them do not analyze the original, but pre-processed versions of the texts. Such pre-processing can include normalizing the text in terms of spelling and punctuation, *stopword removal* (removing words that are not meaningful, like "the," "a," "an"), and *stemming* (reducing words to their stem, so that "voting" and "vote" are recognized as the same concept). While these techniques are fairly simple, more advanced options like *named entity recognition* —where an algorithm tries to detect named actors in a sentence, or *parsing* (to include only specific parts of speech) can also be applied.

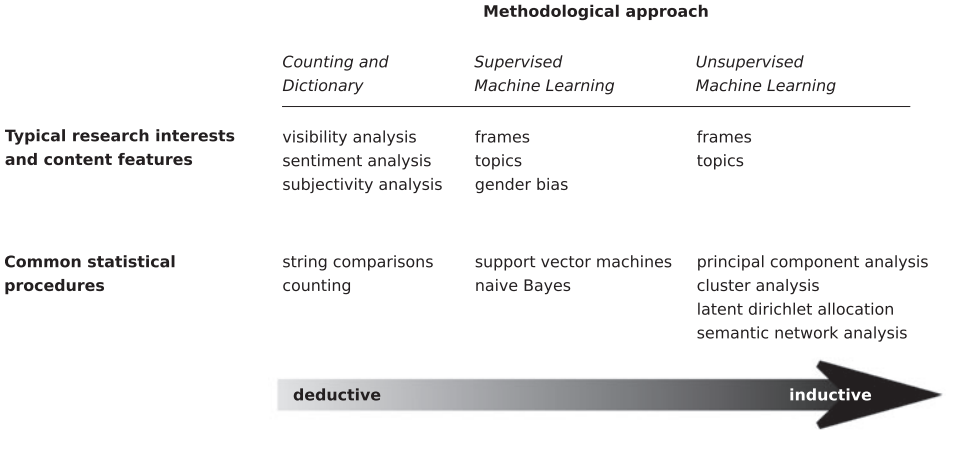

FIGURE 1
Proposed classification of ACA approaches.

Often, a so-called bag-of-words representation of each text is created, which is based on the frequency with which each word is occurring, disregarding the order in which the words appear in the text. Just as stopwords are regarded irrelevant for understanding a text's meaning, there are also words that are considered particularly relevant for a text. Therefore, it is common to attach a specific weight to words. The rationale behind this is that words that occur rarely in language are more informative than words that are very frequently used.[1] For example, one can then use measures like the log likelihood (which compares the actual frequency of each word with its expected frequency) to compare collections of texts (e.g., Rayson and Garside 2000). Mishne (2007), for instance, uses this measure to distinguish different genres of blogs. However, as we will show in the next sections, the ACA toolkit includes much more.

Basic Counting and Dictionary-based Methods

Many content analyses in the realm of journalism studies address questions like "how often is actor X mentioned?" or "what are the most frequent topics in outlets Y and Z?" Such studies involve coding tasks that come down to counting words from a pre-defined list. A large-scale analysis that manually counted references to the European Union and political actors in 52,009 news stories (Schuck et al. 2011) is just one of the many examples of studies that involve counting and that could benefit from an automated approach. When counting is automated, success or failure largely reflects a correct (excluding irrelevant data) and exhaustive (including all the relevant data) definition of the search criteria. Essentially, a given string of text (a pre-defined word or a search string, often a so-called "regular expression") is compared to another string (e.g., a paragraph of a newspaper article). Requiring only a searchable database and basic software like Excel, SPSS, or Stata, this very basic method can be employed by journalism scholars with no previous experience and can result in insightful visibility analyses. One can, for instance, show how often a topic or actor is in the news, and how this differs between outlets or evolves over time. Vliegenthart, Boomgaarden, and Boumans (2011) used such an approach to analyze the relative visibility of politicians in British and Dutch newspapers over time.

The basis of a visibility analysis is in essence an example of a dictionary in its most simple form: a list of key words, used to determine the category a document belongs to. The simple principle behind a dictionary-approach makes it easy to measure a variety of concepts. Within journalism studies, the approach has, for instance, successfully been applied to capture metaphors from news articles (Krennmayr 2014) or hostility in news comments (Ksiazek, Peer, and Zivic 2014). One of the most common research interests to which it is applied, is *sentiment analysis*. Sentiment analysis aims at assessing whether the tone of news content is positive or negative. The dictionary in this case consists of a manually constructed list of words with attached tone scores, which can be either dichotomous or measured on a detailed scale. Sentiment analysis is widely used in marketing and market research to inform organizations on how their brands are evaluated in (social) media (e.g., Mostafa 2013) and is applied to all types of text, including genres as idiosyncratic as movie reviews (e.g., Taboada, Brooke, and Stede, 2009) or suicide notes (Huang, Goh, and Liew, 2007; Pestian et al. 2012).

Despite the wide application in both industry and other scientific fields, by comparison journalism scholars have barely used sentiment analysis (one of the few exceptions being Kleinnijenhuis et al. 2013). In the related field of political communication, measuring tone or sentiment in a computer-assisted fashion is increasingly common, for example to measure how parties are portrayed in the news (Junqué de Fortuny et al. 2012; Van Atteveldt et al. 2008) or to predict election outcomes (but see the critical review of such endeavors by Gayo-Avello 2013). Similar questions are addressed by computer scientists: Stieglitz and Dang-Xuan (2012), for instance, showed how emotions in a political tweet can be used to predict the number of retweets.

The level of sophistication of sentiment analyses varies (Young and Soroka 2012). The most simple, non-statistical approaches count the number of positive and negative words in a text. More powerful algorithms involve multiple features of the text and can deal with, for instance, negation or punctuation use, like the SentiStrength algorithm (Thelwall et al. 2010). In addition, while some algorithms only allow assessments based on a positive–negative dimension, others such as LIWC[2] (Pennebaker, Booth, and Francis 2007) offer additional dimensions like subjectivity or even affective components (i.e., anxiety, anger or sadness; a review of 121 studies using LIWC is provided by Tausczik and Pennebaker 2009).

A study published in *Human Dimensions of Wildlife* on the portrayal of wolves in print media illustrates the wide variety of fields that study journalistic content by means of a dictionary approach. It complements a dictionary with a set of "idea transition" rules, which specify how words and phrases together can create new meanings. For example, when the word "should" appeared near the word "protect," the paragraph was scored as an instance of the concept "wolves should be protected" (Houston, Bruskotter, and Fan 2010, 394–395).

Three remarks should be made concerning the use of dictionary-based techniques. First, it all starts with availability. Dictionaries are manually constructed, which is a very labor-intensive task. Having access to a predefined dictionary is thus of great value. Second, often the applicability of a dictionary is limited to the specific domain within which it is developed. Applying it to a dataset outside this domain can lead to erroneous results (see, e.g., Loughran and McDonald 2011). The development of reliable, freely available dictionaries that are valid across domains needs to be continued. A third issue is that, while internationalization efforts have been undertaken recently, dictionaries are often tailored to the English language, rendering them irrelevant for datasets in other languages.

Above all, however, these lexicons have to be tested and validated (see, e.g., Burscher et al. 2014; Monroe, Colaresi, and Quinn 2008). Obviously, approaches that come down to identifying specific words are deterministic processes and thus extremely reliable, but their validity can be low (e.g., Conway 2006): sometimes, a given word can have multiple meanings, resulting in wrongfully counting an item as relevant (false positive); and sometimes, an unanticipated synonym or a paraphrase might be used which results in not recognizing a relevant item (false negative). "Validate, Validate, Validate," as Grimmer and Stewart (2013, 271), stress, is therefore a key principle of successful ACA. An interesting solution has been employed by Mohammed and Turney (2010), who showed that even a platform like *Amazon Mechanical Turk* can be used to create valid emotion lexicons, with limited financial means.

Supervised Methods

In contrast to ready-to-use techniques that require little manual effort (like using existing dictionaries to measure sentiment), there are more advanced ACA methods that are highly useful for deductive coding of large-scale datasets, but that nevertheless require a relatively high degree of initial manual labor. *Supervised machine learning* is such an approach. It comes in various flavors (like *support vector machines* or *naïve Bayes classifiers*), which cannot be discussed within the scope of this article. Very broadly speaking, techniques under this umbrella are suitable for coding implicit variables in a large dataset, when a smaller sub-set can be (or already has been) hand-coded, but the dataset of interest is so large that it is not feasible to code each article manually. Journalism researchers might be interested in classifying a large amount of digital journalism texts according to their genre (e.g. to find out how a certain issue or actor is treated in editorials compared to news reports). For a human coder, this classification is a rather easy task, but it is difficult to specify an explicit rule for what constitutes an opinion piece. A supervised machine learning algorithm learns from a human coder's decisions and would allow the journalism researcher to solve the classification problem for a virtually unlimited amount of articles. These techniques can be employed for a wide range of research problems including the coding of readability, subjectivity, and gender imbalances (Flaounas et al. 2013), but also, as we will discuss below, frames and topics.

In contrast to simple automated coding, which we discussed above and where the researcher specifies some rules for coding (e.g., code as A if words X or Y are mentioned—which, obviously, is a good way to code explicit and manifest[3] variables), supervised machine learning does not require formulating explicit rules. In fact, the idea is quite the opposite: human coders classify a number of texts, and the machine learning algorithm tries to infer which characteristics of a text lead to which classification. For example, one could ask human coders to classify 1000 articles according to their genres. Then, one would use 500 articles to train a machine learning algorithm and test it by letting it predict the classification of the remaining 500 articles. If the classification of the latter matches the classification of the human coders, one can use the algorithm to classify an unlimited number of new texts.

In spite of the obvious promises of this approach (once trained, it can be used over and over again without any additional costs), journalism scholars have largely neglected this technique. One of the few exceptions is the work by Scharkow (2011), who relied on supervised machine learning to code the topic of articles from 12 German news sites. More recently, Burscher and colleagues have shown that supervised machine learning can be used to code frames in Dutch news articles (Burscher et al. 2014; Odijk et al. 2013) as well as policy issues (Burscher, Vliegenthart, and De Vreese 2015). One should note that there is no inherent limitation as to which classifications can be coded and which cannot. While topics and frames might be of particular interest, one could also think of coding the tone of an article based on supervised machine learning instead of using dictionaries only (e.g., Thelwall et al. 2010).

In the long run, using supervised machine learning does not only increase efficiency, but also transparency and reproducibility. In fact, for English news, there already exist manually annotated corpora that can be used to successfully train a classifier to automatically code the topic of news articles (Flaounas et al. 2013). In this case,

researchers do not need to code the training material themselves, making the technique effectively a fully automated one (for more details, we refer to standard textbooks like Manning, Raghavan, and Schütze 2008). The researcher has to evaluate, though, how close the predicted probability (e.g., of an article being about a given topic) is to the empirical observed probability. A number of metrics are available for this, and it is also possible to graphically examine the reliability of the classifier. When assessing a classifier, the researcher has to make a tradeoff between precision and recall. In the case of a binary classifier (which has only two categories: an item either matches the criteria or not), precision signifies how many of the selected cases are truly relevant, while recall signifies the fraction of all relevant cases that have been identified. For example, a classifier can have a perfect recall (find all relevant cases), but a low precision (it incorrectly finds a number of irrelevant cases as well); or it might have a high precision (only relevant cases are found), but a low recall (many relevant cases are not found).

In political science, the machine learning approach is becoming increasingly common (see, e.g., Grimmer and Stewart 2013). For example, research on the coverage of political speeches has shown that hand-coding between 100 and 500 texts is sufficient to train an algorithm that can distinguish between articles opposing or supporting a policy or between letters to the editor and other articles (Hopkins and King 2010). Similar, in political communication research, supervised machine learning has been applied in the analysis of political tweets (e.g., Roback and Hemphill 2013; Vargo et al. 2014). Also within computer science it is a common approach to use supervised machine learning for classifying the content of blog posts (see, e.g., Husby and Barbosa 2012).

Taken together, supervised machine learning promises that journalism scholars will make more efficient use of the limited resource of human coders. While traditionally, the work of human coders only had value for one single study, their efforts now serve a purpose that extends over multiple studies. Once the discipline has trained solid classifiers, these can be used over and over again. It would also enable *ad hoc* studies on emerging topics to be conducted in a timely fashion, allowing researchers to contribute to recent public debates. To do so, however, the technique has to become more accessible. Fortunately, in the last years, tools have been developed that can help journalism scholars applying supervised machine learning. For example, several R packages have been developed to this end (e.g., Hopkins and King 2010; Jurka et al. 2013).

Yet, researchers have to keep in mind that as a consequence of being dependent on a training dataset, human annotation is still the "gold standard"—by definition, the classifier cannot perform better. In fact, many classifiers perform better than random, but still considerably worse than humans. In addition, they have a certain black-box nature: in contrast to simple dictionary-based methods, it can be difficult to fully understand why a specific classification was made or not. For example, an unobserved oddity in the training data could result in systematic erroneous judgments by the classifier.

Unsupervised Methods

The supervised machine learning approach, by design, is used to code pre-determined categories—and, in fact, is the best way to do so. However, as Grimmer

and Stewart (2013, 281) note, "Supervised methods will never contribute to a new coding scheme." Where such an inductive approach is needed, *unsupervised methods* come into play. Unsupervised methods might be especially interesting for those who want to address questions in the realm of digital journalism that traditionally would have been researched using qualitative methods; researchers interested in questions that aim at describing discourses, frames, or topics in an open way, without having any predefined categories, but who struggle with the amount of data and are looking for tools to help them make sense of the material.

For example, where Burscher et al. (2014) aimed at deductively identifying pre-defined frames, other studies (which we discuss below) attempted to inductively identify and extract frames from text. Just as both approaches to framing analysis exist in manual coding (Matthes and Kohring 2008), there is also an automated equivalent to *inductive frame analysis*. Inductive frames are usually extracted based on the idea that the co-occurrence of words can be interpreted as a frame. This can be done by calculating similarity measures and/or applying statistical techniques like *principal component analysis* or *cluster analysis*. Often, these co-occurrences of words are then graphically visualized as networks of words (e.g., Vlieger and Leydesdorff 2012).

While to our best knowledge no studies that explicitly relate to the field of journalism studies have employed these techniques yet, a number of studies from related fields analyze journalistic coverage in this way. For example, within science communication, the journalistic coverage of artificial sweeteners has been analyzed (Hellsten, Dawson, and Leydesdorff 2009); and within crisis communication, the relationship between coverage of disasters and PR releases has been assessed (Jonkman and Verhoeven 2013; Van der Meer et al. 2014).

Although there is no need for manual coders, like in the case of supervised methods, inductive frame extraction still requires some manual efforts in setting up the analysis and making sense of the results. Comparing the method described by Vlieger and Leydesdorff (2012) with manual content analysis, De Graaf and van der Vossen (2013) state that considerable manual effort is still necessary, while at the same time, reliability problems arise. Yet these techniques have a lot to offer for journalism research, especially with the development of easier to use software packages that can import data from different sources. First of all, the possibility to quickly visualize coverage even when a huge amount of data is to be analyzed allows for a better and deeper understanding than manual methods alone. If the dataset consists of thousands of articles, it is not feasible to read a substantial amount of them to get a grasp of the patterns present in the data. Second, as this approach in fact transforms each text to a set of numbers, it allows the application of all kinds of statistical approaches. This means that questions can be answered that could not be asked before, including the quantitative assessment of similarity overlap between collections of texts.

Related to this is the field of *topic modeling* (e.g., Řehůřek and Sojka 2010). One of the most powerful techniques in this field is Latent Dirichlet Allocation (LDA), first described only 12 years ago (Blei, Ng, and Jordan 2003). For example, LDA has been shown to help identifying important news items (Krestel and Mehta 2010). Grimmer and Stewart (2013) give a detailed overview of the application of LDA topic models in political science.

The approaches to inductive frame analysis and to the assessment of document similarity discussed thus far largely rely on a bag-of-words approach, in which it is

considered necessary for the computer to understand the function or meaning of each word. Mishne (2007), for instance, shows that the mood of blog posts can be accurately predicted by the occurrence of a few characteristic words. Some criticize such an approach as overly simplistic. Indeed, it is easy to point to cases where this modeling of the data would fail, for instance with respect to negation ("not good" is counted as "good") or with the incapability of distinguishing between different meanings of a word. For example, the word "Amstel" could refer either to one of the largest beer brands in the Netherlands, to a renowned hotel in Amsterdam, or to the river after which both are named.

In response to the limitations that arise from disregarding the syntactic structure of sentences, methods developed in the field of computer linguistics use more advanced representations of texts, in which the relationship between elements is modeled. Techniques used for this purpose include *part-of-speech tagging* and *named entity recognition*. To be able to understand which word is a proper noun, and thus refers to a potentially interesting actor, or to distinguish between "Israel attacked Hamas" and "Hamas attacked Israel" (Sheafer et al. 2014) can be of vital importance to a researcher. A possible application for journalism researchers could lie in studies that do not only measure whether actors are mentioned, but also precisely in what context they are referred to. Building on such ideas, Van Atteveldt (2008) presents a method called auto- mated semantic network analysis and shows that it is possible to use a computer to code semantic relationships and disentangle the syntactical function of the elements of a sentence (see also Van Atteveldt, Kleinnijenhuis, and Ruigrok, 2008). This can, for instance, be a powerful tool to assess the role in which certain actors appear in the news.

Even more than in the case of dictionaries (which in principle can be translated), fixation on the English language is a considerable limitation for advanced natural language processing techniques. Parsers do not exist for all languages, and some languages are much more difficult to parse than others. Other limitations of unsupervised methods include the potential openness for different interpretations by the researcher: they do not offer one and only one correct solution. Also, especially when the researchers do not carefully plan necessary pre-processing steps, they can be sensitive to peculiarities of the material that are irrelevant from a theoretical point of view.

Identifying and Overcoming Obstacles

From the previous sections it has become clear that automated approaches have much to offer scholars studying journalistic content and, indeed, the approaches are commonly applied in a number of academic fields. Oddly enough, as yet, relatively few studies within the field of journalism studies make use of them (but see, for instance, Günther and Scharkow 2014; Kleinnijenhuis et al. 2013; Krennmayr 2014; Ksiazek, Peer, and Zivic 2014; Sjøvaag and Stavelin 2012). Particularly inductive and more complex approaches appear to be hardly applied by journalism scholars (a notable exception being Flaounas et al. 2013). To understand why this might be the case, a number of observations can be made.

First, there is a tendency to use familiar methods. While methods are continuously improved and new approaches are created, scholars are human beings and stick to methods they once learned, both in research and in teaching—which implies that it

takes a long time before a methodological innovation actually diffuses and becomes commonly accepted. The younger generation of scholars may very well already be more accustomed to automated approaches. Second, journalism scholars are unaware of each other's research. Journalism scholars may also often simply be unaware of new methods. The reference list of this article illustrates this point: often, studies of core value to journalism research are published in the proceedings of a computer science conference or other venues less familiar to journalism scholars. In addition, citation patterns show little to no overlap. Third, different fields speak different languages. Even if a scholar does read an article from another discipline, it can be difficult to understand the language and technical terms that each discipline applies. Several options can be considered to overcome these barriers, as follows:

- *Cooperations and interdisciplinary teams*: A lexicon to guide the journalism scholar through the computer scientist's world and their jargon would be a welcome start. But more valuable would be to increase cooperation between the fields. The fact that there are more and more research teams involving both computer scientists and journalism or communication scholars (such as Burscher et al. 2014; Flaounas et al. 2013; Morgan, Shafiq, and Lampe 2013; Stieglitz et al. 2014) indicates a growing awareness of the surplus value that such joint projects can offer.
- *Teach new methods*: Some communication and journalism departments have started offering courses on computer programming. While we do not believe that every journalism scholar has to be a programmer, we deem some *code literacy* to be more and more useful: already some basic knowledge of programming can help to get a grasp of the computer science literature. Familiarizing current journalism students with advanced automated approaches will ensure that lack of knowledge and skills are no longer obstacles for the next generation of scholars. The first textbooks on how to use Python and R for these tasks have appeared (e.g., Munzert et al. 2015).
- *Use custom-made tools*: Vis (2013), for instance, stresses the importance of first thoroughly defining the research questions before making any decision on the tools to use. Often, the tool has to be tailored to the task. Luckily, as we have outlined in this article, the building blocks for such an endeavor exist. Indeed, more and more content-analytical journalism research relies on custom-written programs, making use of, for example, the large variety of available Python modules (Lewis, Zamith, and Hermida 2013; Sjøvaag and Stavelin 2012; Trilling 2015).
- *Share not only results, but also code*: As a last measure, we advocate a culture of sharing and acknowledging code. Instead of reinventing the wheel again and again, tools should be accessible—and, even more important, the source code should be available, allowing the researcher to tailor the tool to specific needs. The increased transparency would also help a lot to increase reproducibility of our research. First steps in this direction have already been made in neighboring fields. For example, the journal *Political Analysis* now requires all authors to submit a replication package, including not only the raw data, but also all syntax files and any form of code necessary to reproduce the paper's results. While admittedly some of the above suggestions are more easily implemented than others, it may be clear that there are various opportunities for journalism scholars to become more familiar with automated approaches.

Conclusion

In this article, we have reviewed a number of new methodological approaches and tools that can offer substantial added value to journalism research. They could be applied selectively, but also combined or enhanced, so that they suit the demands of journalism scholars. We proposed to order techniques for ACA along a deductive–inductive continuum. By doing so, we hope to have illustrated the wide range of possible applications. It might help researchers to relate automated techniques to approaches they are familiar with. At the same time, it should have become clear that there is no such thing as "the" automated content analysis, but that ACA approaches are very versatile. While often developed outside the field of journalism studies, the techniques can be employed to answer a wide variety of journalism studies research interests: from visibility, representation, and evaluation to tone, subjectivity, and frames. We stress that using automated techniques is by no means in opposition to or in competition with manual content analysis. The entire research process, ranging from formulating research questions to making modeling decisions and interpreting results, requires a deep understanding of the data. Thus, rather than replacing humans, it is more correct to view computers as *amplifying* human abilities (Grimmer and Stewart 2013). As remarked by Günther et al. (2015, 5): "A manual topic analysis cannot simply be translated into a fully-automated approach," as "the manual analysis will yield a more specific result." Optimally then, one would combine "the best of both worlds" (Wettstein 2014, 16). Regardless of the type of approach, advantages and disadvantages have to be carefully assessed—and it is the researcher's responsibility to check whether the method actually performs well. The quality of the output of an automated system largely depends on the quality of the input data. If, for instance, the collected data do not match the expected structure (e.g., because a website changed its layout during the research period) and this remains undiscovered, this will lead to biased or inaccurate results. Finally, we want to stress that automated approaches are by no means a panacea to all the challenges that digital journalism research faces. The analysis of visual data is just one example where manual coding still outperforms automated alternatives. Yet to keep up with the ever-evolving nature of digital journalism, the ability to adapt our approaches accordingly is a virtue. Familiarizing ourselves with what the ACA toolkit has to offer is the least we can do, and we hope that this overview will serve as an inspiration to do so.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

NOTES

1. A common example is the *tf–idf* (term frequency–inverse document frequency) scheme, in which the frequency of a term in a given document is weighted by the number of documents in which it occurs.
2. The name of the program stands for Linguistic Inquiry and Word Count.

3. These are variables that are directly observable (like *number of words* or *is actor X mentioned?*), and, unlike abstract concepts as tone or frame, are not open to multiple interpretations.

REFERENCES

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Burscher, Björn, Daan Odijk, Rens Vliegenthart, Maarten de Rijke, and Claes H. de Vreese. 2014. "Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis." *Communication Methods and Measures* 8 (3): 190–206. doi:10.1080/19312458.2014.937527.

Burscher, Björn, Rens Vliegenthart, and Claes H. De Vreese. 2015. "Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?" *Annals of the American Academy of Political and Social Science* 659 (1): 122–131.

Carlson, Matt. 2014. "The Robotic Reporter: Automated Journalism and the Redefinition of Labor, Compositional Forms, and Journalistic Authority." *Digital Journalism* 3 (3): 416–431. doi:10.1080/21670811.2014.976412.

Coddington, Mark. 2014. "Clarifying Journalism's Quantitative Turn: A Typology for Evaluating Data Journalism, Computational Journalism, and Computer-assisted Reporting." *Digital Journalism* 3 (3): 331–348. doi:10.1080/21670811.2014.976400.

Conway, Mike. 2006. "The Subjective Precision of Computers: A Methodological Comparison with Human Coding in Content Analysis." *Journalism & Mass Communication Quarterly* 83 (1): 186–200.

de Fortuny, Junqué, Tom De Enric, David Martens Smedt, and Walter Daelemans. 2012. "Media Coverage in times of Political Crisis: A Text Mining Approach." *Expert Systems with Applications* 39 (14): 11616–11622. doi:10.1016/j.eswa.2012.04.013.

De Graaf, Rutger, and Robert van der Vossen. 2013. "Bits versus Brains in Content Analysis: Comparing the Advantages and Disadvantages of Manual and Automated Methods for Content Analysis." *Communications* 38 (4): 433–443. doi:10.1515/commun-2013-0025.

Fink, Katherine, and C. W. Anderson. 2014. "Data Journalism in the United States: Beyond the 'Usual Suspects'." *Journalism Studies* 16 (4): 467–481. doi:10.1080/1461670X.2014.939852.

Flaounas, Ilias, Omar Ali, Thomas Lansdall-Welfare, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. 2013. "Research Methods in the Age of Digital Journalism." *Digital Journalism* 1 (1): 102–116. doi:10.1080/21670811.2012.714928.

Flew, Terry, Christina Spurgeon, Anna Daniel, and Adam Swift. 2012. "The Promise of Computational Journalism." *Journalism Practice* 6 (2): 157–171. doi: 10.1080/17512786.2011.616655.

Freelon, Deen. 2014. "On the Interpretation of Digital Trace Data in Communication and Social Computing Research." *Journal of Broadcasting & Electronic Media* 58 (1): 59–75. doi:10.1080/08838151.2013.875018.

Gayo-Avello, Daniel. 2013. "A Meta-analysis of State-of-the-art Electoral Prediction from Twitter Data." *Social Science Computer Review* 31 (6): 649–679. doi:10.1177/0894439313493979.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297. doi:10.1093/pan/mps028.

Günther, Elisabeth, and Michael Scharkow. 2014. "Recycled Media: An Automated Evaluation of News Outlets in the Twenty-first Century." *Digital Journalism* 2 (4): 524–541.

Günther, Elisbabeth, Ines Engelmann, Christoph Neuberger and Thorsten Quandt. 2015. "From Text to Topics: A Comparison of a Manual and an Automated Content Analysis." Paper presented at Re-inventing Journalism, Winterthur, Switzerland. http://www.amiando.com/eventResources/h/v/fSMcdpvvmsUWee/Presentations_C2_Elisabeth_Guenther.pdf

Hellsten, Iina, James Dawson, and Loet Leydesdorff. 2009. "Implicit Media Frames: Automated Analysis of Public Debate on Artificial Sweeteners." *Public Understanding of Science* 19 (5): 590–608. doi:10.1177/0963662509343136.

Hopkins, Daniel J., and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–247.

Houston, Melanie J., Jeremy T. Bruskotter, and David Fan. 2010. "Attitudes toward Wolves in the United States and Canada: A Content Analysis of the Print News Media, 1999-2008." *Human Dimensions of Wildlife* 15 (5): 389–403.

Huang, Yen-Pei, Tiong Goh and Chern Li Liew. 2007. "Hunting Suicide Notes in Web 2.0: Preliminary Findings." Ninth IEEE International Symposium on Multimedia Workshops, 517–521. doi: 10.1109/ISM.Workshops.2007.92.

Husby, Stephanie D. and Denilson Barbosa. 2012. "Topic Classification of Blog Posts Using Distant Supervision." In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 28–36. France: Avignon.

Illia, Laura, Karan Sonpar, and Martin W. Bauer. 2014. "Applying Co-Occurrence Text Analysis with ALCESTE to Studies of Impression Management." *British Journal of Management* 25 (2): 352–372. doi:10.1111/j.1467-8551.2012.00842.x.

Jonkman, Jeroen, and Piet Verhoeven. 2013. "From Risk to Safety: Implicit Frames of Third-party Airport Risk in Dutch Quality Newspapers between 1992 and 2009." *Safety Science* 58: 1–10. doi:10.1016/j.ssci.2013.03.012.

Jurka, Timothy P., Loren Collingwood, Amber Boydstun, Emiliano Grossman, and Wouter van Atteveldt. 2013. "RTextTools : A Supervised Learning Package for Text Classification." *The R Journal* 5 (1): 6–12.

Kleinnijenhuis, Jan, Friederike Schultz, Dirk Oegema, and Wouter van Atteveldt. 2013. "Financial News and Market Panics in the Age of High-frequency Sentiment Trading Algorithms." *Journalism* 14 (2): 271–291. doi:10.1177/1464884912468375.

Krennmayr, Tina. 2014. "What Corpus Linguistics Can Tell Us about Metaphor Use in Newspaper Texts." *Journalism Studies* 16 (4): 530–546. doi: 10.1080/1461670x.2014.937155.

Krestel, Ralf, and Bhaskar Mehta. 2010. "Learning the Importance of Latent Topics to Discover Highly Influential News Items." In *KI 2010: Advances in Artificial Intelligence*, edited by Rüdiger Dillmann, Jürgen Beyrer, Uwe D. Hanebeck and Tanja Schultz, 211–218. Berlin, Germany: Springer.

Ksiazek, Thomas B., Limor Peer, and Andrew Zivic. 2014. "Discussing the News: Civility and Hostility in User Comments." *Digital Journalism*. doi: 10.1080/21670811.2014.972079.

Lewis, Seth C., Rodrigo Zamith and Alfred Hermida. 2013. "Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods." *Journal of Broadcasting & Electronic Media* 57 (1): 34–52. doi:10.1080/08838151.2012.761702.

Loughran, Tim, and Bill McDonald. 2011. "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *Journal of Finance* 66 (1): 35–65.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press.

Matthes, Jörg, and Matthias Kohring. 2008. "The Content Analysis of Media Frames: Toward Improving Reliability and Validity." *Journal of Communication* 58 (2): 258–279. doi:10.1111/j.1460-2466.2008.00384.x.

Mishne, Gilad A. (2007). "*Applied Text Analytics for Blogs*." PhD dissertation, University of Amsterdam. http://hdl.handle.net/11245/2.47196.

Mohammed, Saif M. and Peter D. Turney. 2010. "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon." *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 26–34. Stroudsburg, PA: ACL. https://www.aclweb.org/anthology/W/W10/W10-02.pdf.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16 (4): 372–403. doi:10.1093/pan/mpn018.

Morgan, Jonathan Scott, M. Zubair Shafiq, and Cliff Lampe. 2013. "Is News Sharing on Twitter Ideologically Biased?" In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 887–897. New York: ACM.

Mostafa, Mohamed M. 2013. "More than Words: Social Networks' Text Mining for Consumer Brand Sentiments." *Expert Systems with Applications* 40 (10): 4241–4251. doi:10.1016/j.eswa.2013.01.019.

Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis. 2015. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Chichester, UK: Wiley.

Odijk, Daan, Björn Burscher, Rens Vliegenthart, and Maarten de Rijke. 2013. "Automatic Thematic Content Analysis: Finding Frames in News. *Social Informatics*." *Lecture Notes in Computer Science* 8238: 333–345. doi:10.1007/978-3-319-03260-3_29.

Pennebaker, James W., Roger J. Booth and Martha E. Francis. 2007. *Linguistic Inquiry and Word Count: LIWC*. Austin; TX: LIWC.net.

Pestian, John P., Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, K. Jan Wiebe, Bretonnel Cohen, John Hurdle, and Christopher Brew. 2012. "Sentiment Analysis of Suicide Notes: A Shared Task." *Biomedical Informatics Insights* 5: 3–16. doi:10.4137/BII.S9042.

Rayson, Paul and Roger Garside. 2000. "Comparing Corpora Using Frequency Profiling." NAACL-ANLP 2000 Workshop: Syntactic and Semantic Complexity in Natural Language Processing Systems. http://acl.ldc.upenn.edu/W/W00/W00-0901.pdf

Řehůřek, Radim and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Roback, Andrew and Libby Hemphill. 2013. "How Constituents Lobby Members of Congress on Twitter." *In Annual Meeting of the American Political Science Association*. http://ssrn.com/abstract=2301133

Scharkow, Michael. 2011. "Thematic Content Analysis Using Supervised Machine Learning: An Empirical Evaluation Using German Online News." *Quality & Quantity* 47 (2): 761–773. doi:10.1007/s11135-011-9545-7.

Schneider, Gerold. 2014. "Automated Media Content Analysis from the Perspective of Computational Linguistics." In *Automatisierung in Der Inhaltsanalyse*, edited by Katharina Sommer, Martin Wettstein, Werner Wirth and Jörg Matthes, 40–54. Cologne, Germany: Herbert von Halem.

Schuck, Andreas R. T., Georgios Xezonakis, Matthijs Elenbaas, Susan A. Banducci, and Claes H. de Vreese. 2011. "Party Contestation and Europe on the News Agenda: The 2009 European Parliamentary Elections." *Electoral Studies* 30 (1): 41–52. doi:10.1016/j.electstud.2010.09.021.

Sheafer, Tamir, Shaul R. Shenhav, Janet Takens, and Wouter van Atteveldt. 2014. "Relative Political and Value Proximity in Mediated Public Diplomacy: The Effect of State-Level Homophily on International Frame Building." *Political Communication* 31 (1): 149–167. doi:10.1080/10584609.2013.799107.

Simon, Adam F. 2001. "A Unified Method for Analyzing Media Framing." In *Communication in U.S. Elections: New Agendas*, edited by Roderick P. Hart and Daron R. Shaw, 75–89. Lanham, MD: Rowman and Littlefield.

Sjøvaag, Helle and Eirik Stavelin. 2012. "Web Media and the Quantitative Content Analysis: Methodological Challenges in Measuring Online News Content." *Convergence: The International Journal of Research into New Media Technologies*, 18 (2), 215–229. doi:10.1177/1354856511429641.

Stieglitz, Stefan and Linh Dang-Xuan. 2012. "Political Communication and Influence through Microblogging: An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior." *2012 45th Hawaii International Conference on System Sciences*, 3500–3509. doi: 10.1109/HICSS.2012.476.

Stieglitz, Stefan, Linh Dang-Xuan, Aexel Bruns, and Christoph Neuberger. 2014. "Social Media Analytics." *Business & Information Systems Engineering* 6 (2): 89–96. doi:10.1007/s12599-014-0315-7.

Taboada, Maite, Julian Brooke and Manfred Stede. 2009. "Genre-based Paragraph Classification for Sentiment Analysis." In *Proceedings of the SIGDIAL 2009 Conference on the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 62–70. Morristown, NJ: ACL. doi:10.3115/1708376.1708385.

Tausczik, Yla R., and James W. Pennebaker. 2009. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29 (1): 24–54. doi:10.1177/0261927X09351676.

Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. "Sentiment Strength Detection in Short Informal Text." *Journal of the American Society for Information Science and Technology* 61 (12): 2544–2558. doi:10.1002/asi.21416.

Trilling, Damian. 2015. "Two Different Debates? Investigating the Relationship between a Political Debate on TV and Simultaneous Comments on Twitter." *Social Science Computer Review* 33 (3): 259–276. doi:10.1177/0894439314537886.

Uhl, Matthias W. 2014. "Reuters Sentiment and Stock Returns." *Journal of Behavioral Finance* 15 (4): 287–298.

Van Atteveldt, Wouter. 2008. *Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content*. Charleston, SC: BookSurge.

Van Atteveldt, Wouter, Jan Kleinnijenhuis, and Nel Ruigrok. 2008. "Parsing, Semantic Networks, and Political Authority Using Syntactic Analysis to Extract Semantic Relations from Dutch Newspaper Articles." *Political Analysis* 16 (4): 428–446. doi:10.1093/pan/mpn006.

Van Atteveldt, Wouter, Jan Kleinnijenhuis, Nel Ruigrok, and Stefan Schlobach. 2008. "Good News or Bad News? Conducting Sentiment Analysis on Dutch Text to Distinguish between Positive and Negative Relations." *Journal of Information Technology & Politics* 5 (1): 73–94. doi:10.1080/19331680802154145.

Van der Meer, G. L. A. Toni, Piet Verhoeven, Hans Beentjes, and Rens Vliegenthart. 2014. "When Frames Align: The Interplay between PR, News Media, and the Public in times of Crisis." *Public Relations Review* 40 (5): 751–761. doi:10.1016/j.pubrev.2014.07.008.

Vargo, Chris J., Lei Guo, Maxwell McCombs, and Donald L. Shaw. 2014. "Network Issue Agendas on Twitter during the 2012 U.S. Presidential Election." *Journal of Communication* 64: 296–316. doi:10.1111/jcom.12089.

Vis, Farida. 2013. "A Critical Reflection on Big Data: Considering APIs, Researchers and Tools as Data Makers." *First Monday* 18 (10): 1–16. doi:10.5210/fm.v18i10.4878.

Vliegenthart, Rens, Hajo G. Boomgaarden and Jelle W. Boumans. 2011. "Changes in Political News Coverage: Personalisation, Conflict and Negativity in British and Dutch Newspapers." In *Challenging the Primacy of Politics*, edited by Kees Brants and Karin Voltmer, 92–110. London, UK: Palgrave Macmillan.

Vlieger, Esther, and Loet Leydesdorff. 2012. "Content Analysis and the Measurement of Meaning: The Visualization of Frames in Collections of Messages." In *Research Methodologies, Innovations and Philosophies in Systems Engineering and Information Systems*, edited by Manuel Mora, Ovsei Gelman, Anette Steenkamp and Manesh S. Raisinghani, 322–340. Hershey, PA: Information Science Reference.

Wettstein, Martin. 2014. "'Best of Both Worlds': Die Halbautomatische Inhaltsanalyse [Best of Both Worlds: The Semi-automated Content Analysis]." In *Automatisierung in Der Inhaltsanalyse*, edited by Katharina Sommer, Martin Wettstein, Werner Wirth and Jörg Matthes, 16–39. Cologne, Germany: Herbert von Halem.

Young, Lori, and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29 (2): 205–231. doi:10.1080/10584609.2012.671234.

Jelle W. Boumans (author to whom correspondence should be addressed), Department of Communication Science (Corporate Communication), University of Amsterdam, The Netherlands; Corresponding author. E-mail: j.w.boumans@uva.nl

Damian Trilling, Department of Communication Science (Political Communication), University of Amsterdam, The Netherlands; E-mail: d.c.trilling@uva.nl.