Automated content analysis of online political communication

Ross Petchler and Sandra González-Bailón

INTRODUCTION

Content analysis has a long tradition in the social sciences. It is central to the study of policy preferences (Budge, 2001; Laver et al., 2003), propaganda and mass media (Krippendorff, 2013 [1980]; Krippendorff and Bock, 2008), and social movements (Della Porta and Diani, 2006; Johnston and Noakes, 2005). New computational tools and the increasing availability of digitized documents promise to push forward this line of inquiry by reducing the costs of manual annotation and enabling the analysis of large-scale corpora. In particular, the automated analysis of online political communication may yield insights into political sentiment which offline opinion analysis instruments (such as polls) fail to capture. Online communication is constantly pulsating, generating data that can help us uncover the mechanisms of opinion formation - if the appropriate measurement and validity methods are developed.

Several linguistic peculiarities distinguish online political communication from traditional political texts. For a start, it is often far less formal and structured. In addition, automated content analysis techniques are not always as reliable or as valid as manual annotation, which makes measurements potentially noisy or misleading. With these challenges in mind, we provide an overview of techniques suited to two common content analysis tasks: classifying documents into known categories, and discovering unknown categories from documents (Liu, 2012; Blei, 2013). This second task is more exploratory in nature: it helps to identify topic domains when there are no clear preconceptions of the topics that are discussed in a certain communication environment. The first task, on the other hand, can help to label a large volume of text in a more efficient manner than manual annotation; for instance, when the research question requires identifying the emotional tone of political communication (as positive, negative or neutral) or its ideological slant (liberal or conservative). This chapter focuses on the application of these automated techniques to online political communication, and suggests directions for future research in this domain.

METHODS FOR AUTOMATED CONTENT ANALYSIS

The application of automated text analysis techniques requires the prior acquisition and preprocessing of data. This section discusses the logic of preprocessing texts to then provide an overview of techniques to classify documents in known categories or discover topics when no categories are known.

Acquiring and Preprocessing Online Political Texts

Political scientists have applied automated content analysis techniques to many kinds of offline political texts, including newspaper articles (Young and Soroka, 2012), presidential and legislator statements (Grimmer and King, 2011), legislature floor speeches (Quinn et al., 2010), and treaties (Spirling, 2012). Until recently, though, political texts remained relatively understudied because they were difficult to parse and process for analysis.

Acquiring online political texts is becoming simpler as more sites store and transmit them in machine-readable formals such as Extensible Markup Language (XML) and JavaScript Object Notation (JSON) or make them publicly available via application programming interfaces (APIs). When such options are unavailable, researchers familiar with statistical software or scripting languages can use new packages for automated HyperText Markup Language (HTML) scraping (Python and R, for instance, have built-in packages and libraries). Finally, when neither machine-readable nor easy-to-parse HTML data are available, researchers can crowdsource data acquisition and parsing via sites like Amazon Mechanical Turk (Berinsky et al., 2012). Overall, these new technologies enable communication scholars to access and study previously unavailable indicators of public opinion.

In order to perform automated content analysis researchers must transform texts into structured data that can be quantified (Franzosi, 2004). Prior to the advent of new computational tools this was performed by human coders using a pre-determined scheme (Krippendorf, 2013 [1980]; Neuendorf, 2001). Initially, a codebook is written guided by a research question and a theoretical context. It is iteratively improved until coders no longer notice ambiguities, at which point it is applied to the data set. Automated approaches preprocess the text to reduce the complexity of language, often using a bag-of-words model to eliminate the most frequent words and to reduce words to their morphological roots (Jurafsky and Martin, 2009; Hopkins and King, 2010; Porter, 1980). After preprocessing, documents are represented as a document-term matrix in which rows correspond to documents, sentences, or expressions (depending on then unit of analysis), and columns correspond to words or tokens. Cells can contain continuous values (representing how frequently each term occurs in each document) or binary values (representing whether each term occurs in the document).

Which of the two approaches is more appropriate (to code documents manually or to apply automated preprocessing) depends on the complexity of the research question at hand, the number of documents collected, and the tolerance for error. Although manually annotated data remain the gold standard for content analysis, the sections that follow focus mostly on cases in which data are automatically preprocessed, since this is more common when dealing with large volumes of text.

Once online political texts are converted to a structured form, several methods for automated content analysis can be applied. We divide these methods into two groups to reflect the two most common content analysis tasks: classifying documents into known categories, and discovering theoretically important categories from the content. The former task encompasses techniques such as lexicon-based classification and supervised learning. The latter task encompasses unsupervised learning and the analysis of text as networks of concepts. The following sections explain the details of these techniques and highlight their relative strengths and weaknesses to help communication researchers choose the approach best suited to their specific data and research question.

Classifying Documents into Known Categories

The goal of supervised content analysis techniques is to classify documents into a number of known categories. For example, news articles may have left-leaning or right-leaning ideological

biases (Gentzkow and Shapiro, 2010) or have positive or negative coverage (Eshbaugh-Soha, 2010). This section offers an overview of the techniques that allow that sort of classification. There are two main methods. The first is a lexicon-based approach, which uses relative keyword frequencies to measure the prevalence of each category in a document. The second is supervised learning, which uses a training data set of manually annotated documents to classify new, unlabeled documents.

Lexicon-based classification

The lexicon (or dictionary)-based approach to document classification is the simplest automated content analysis technique (Liu, 2012). It is based on a list of words and phrases and their associated category labels. For example, a lexicon for classifying micro-blog posts according to sentiment may map the words 'good' and 'beautiful' to the positive category and the words 'bad' and 'ugly' to the negative category. A lexicon for classifying blog posts according to ideological subject may map the words 'healthcare' and 'environment' to the left-leaning category and 'foreign policy' and 'taxes' to the right-leaning category.

Off-the-shelf lexicons include the Linguistic Inquiry and Word Count, or LIWC (Pennebaker et al., 2001), and the General Inquirer (Wilson et al., 2005). Not all lexicons are based on binary categories. Some sentiment lexicons have positive, neutral, and negative terms, measured on a several points scale. The Affective Norms for English Words (ANEW) lexicon, for instance, labels words and phrases according to psychometric categories which rate words on three emotional dimensions: valence, arousal, and dominance (Bradley and Lang, 1999; Osgood et al., 1957). This lexicon helps to analyze documents by counting the relative frequency with which words appear and averaging the scores associated to each word in each dimension, from 0 to 9. This approach has been applied effectively to extract sentiment measures from a number of online data sources (Dodds and Danforth, 2009; Dodds et al., 2011).

The success of a lexicon-based content analysis relies on the quality of the lexicon; that is, how appropriate it is in the context of the specific research question and data being analyzed (Gonzalez-Bailón and Paltoglou, 2015). Using ‘off-the-shelf’ lexicons compiled with generic research goals may produce poor results when applied to specific types of political communication (Loughran and McDonald, 2011). It is always best to generate lexicons specific to a research question, and there are three main approaches for doing so. The first is to manually annotate the sentiment of all adjectives in a dictionary of all the words in a corpus, in line with the information domain under scrutiny (that is, 'warming' can be labeled differently if used in environmental policy or foreign affairs communication). This is time-consuming but tunes the lexicon to specific communication contexts. Researchers concerned with efficiency as well as accuracy have used online crowdsourcing platforms such as CrowdFlower, Amazon Mechanical Turk, and Tasken to quickly and accurately label large sentiment lexicons. For example, Dodds et al. (2011) created a lexicon of 10222 words by merging the 5000 most frequently occurring words in a Tweet corpus, Google Books, music lyrics, and the *New York Times;* they then used Amazon Mechanical Turk to obtain 50 sentiment ratings of each word on a nine-point scale from negative to positive. They found that the sentiment lexicon labeled by crowdsourcing workers was highly correlated with the ANEW lexicon.

The second way to generate a sentiment lexicon is dictionary-based. The general approach is to manually label the sentiment of a small set of seed words, and then search a dictionary (the most frequently used is WordNet; see Miller et al., 1990) for their synonyms and antonyms; these snowballed terms are then labeled with the same or opposite sentiment as the corresponding seed word and then are added to the set of seed words. The process is iterated until no words remain unlabeled. For example, the seed word 'excellent' is labeled positive; synonyms such as 'beautiful', 'fabulous', and 'marvelous' are labeled as positive as well; while antonyms such as 'awful', 'rotten', and 'terrible' are labeled as negative. An example of a lexicon generated using the dictionary-based approach is Sentiment Lexicon, constructed by Hu and Liu (2004). This dictionary-based approach quickly generates a large list of labeled sentiment words, but requires manual cleaning and ignores ambiguity due to context, which is particularly important in the analysis of political communication.

The third way to generate a sentiment lexicon is corpus-based. The general approach is to manually label the sentiment of a small set of seed words and then define linguistic rules to identify similar or dissimilar sentiment words. A seed word may be 'beautiful' and its label 'positive'; linguistic rules based on connective words (such as 'and' or 'but') help to assign labels to subsequent words. For instance, if a document in a corpus contains the phrase 'The car is beautiful and spacious' then the term 'spacious' could be assigned the label 'positive' based on the connective word 'and'. Conversely, if a document in a corpus contains the phrase 'The car is spacious but difficult to drive' then the term 'spacious' could be assigned the label 'negative' based on the connective word 'but'. This methodology requires clearly defined linguistic rules in order to achieve good results; and linguistic rules assume sentiment consistency across documents, which is not necessarily the case for most empirical domains: the same word can express opposite sentiments in different communication contexts (Liu, 2012). Overall, though, the corpus-based methodology to lexicon generation is useful in two cases: to discover other sentiment words and their orientations on the basis of a hand-made seed list; and to adapt a general-purpose lexicon to a specific communication domain. The corpus-based approach is less useful for building a general-purpose sentiment lexicon than the dictionary-based approach because dictionaries encompass more words.

These three techniques are based on different assumptions that affect the results they produce. None of these sentiment lexicons is perfect because they are too general to suit the specific needs of different communication domains. In addition, certain words and phrases in online political communication are too informal, specific, or novel (and therefore infrequent) to be contained in existing lexicons. A corpus-based technique can capture and label these distinct words; for instance, Brody and Diakopoulos (2011) find that lengthened words in microblog posts (for example, 'looove') are strongly associated with subjectivity and sentiment; and Derks et al. (2007) find that emoticons (for example, ':)') strengthen the intensity of online communication. Researchers have already incorporated the peculiarities of online communication into their sentiment models (Paltoglou et al., 2010; Paltoglou and Thelwall, 2012), but often additional manual labeling is needed to add other novel words to the seed list. These limitations make validation a crucial component of automated content analysis (Grimmer and Stewart, 2013). Having the appropriate validation strategies in place is necessary to increase confidence in measurement.

Supervised learning

The second main approach to document classification using pre-existing categories is supervised learning. Supervised algorithms learn from a training data set of manually annotated documents how to classify new, unlabeled documents. The supervised learning approach has three steps. First, it constructs a training data set. Second, it applies an automated algorithm to determine the relationships between features of the training data set and the categories that are used to classify documents. And third, it predicts (or assigns) categories for unlabeled documents and validates that classification. The remainder of this section reviews these three steps in turn.

The first step in supervised learning is to construct a training data set. As described above, this involves transforming unstructured textual data into structured quantitative data. In addition to preprocessing, it is common for researchers to manually code documents for features that the bag-of-words model ignores: for instance, they may add features accounting for the source or the author of a document. The larger the training data set, the more information supervised learning algorithms have with which to make predictions, but scaling up can be computationally costly. The specific research question and data source inform the balance between the need for a large training data set and the costs of compiling training data.

The second step in supervised learning is to apply an algorithm that will associate text features to each category in the classification scheme. There are many different algorithms and the field of machine learning and natural language processing is quickly growing in this area; Hastie et al. (2009) offer a good overview of the techniques available. Each model has specific characteristics and parameters, which makes a general discussion difficult, but popular algorithms include (multinomial) logistic regression, the naive Bayes classifier (Maron and Kuhns, 1960), random forests (Breiman, 2001), support vector machines (Cortes and Vapnik, 1995), and neural networks (Bishop, 1995). Each of these algorithms uses the information gathered from the training data to assign new examples of text into the classification categories.

This assignment takes place in the third and final step, where supervised approaches predict the categories for unlabeled documents and ·validate the results. A model that performs well will replicate the results of manual coding, which still offers the gold standard; a model that performs poorly will fail to replicate these results. The standard method to validate models is cross-validation. This entails splitting the labeled documents into equally sized groups (usually about ten) and then predicting the categories of the observations in each group using the pooled observations in the+other groups. This method avoids overfitting to data because it focuses on out-of-sample prediction. Overall, the supervised approach systematically performs better than unsupervised approaches in the analysis of online communication because it is able to capture more accurately the contextual features of the text and language used (Gonzalez-Bailon and Paltoglou, 2015).

Discovering Categories and Topics from Documents

Unsupervised learning

In contrast to supervised approaches, unsupervised techniques do no require manually annotated training data; consequently, they are much less costly to implement. They are good exploratory techniques but their results can be difficult to evaluate: concepts such as validity and consistency

compared to human labeling do not immediately apply because these techniques are used, in part, to overcome the lack of predefined labels or categories - hence their exploratory nature. This section briefly discusses three categories of unsupervised techniques: cluster analysis, dimensionality reduction, and topic modeling.

The goal of cluster analysis is to partition a corpus of documents into groups of similar documents, where 'similar' is measured in terms of word frequency distributions. The most widely used clustering algorithm is *k*-means (MacQueen, 1967), which partitions documents into *k* disjoint groups by minimizing the sum of the squared Euclidean distances within clusters; distance is measured as the number of words that any two documents share. Other clustering algorithms use different distance metrics or objective functions (which are used to optimize or find the best clustering classification out of all possible classifications). Given that few papers provide guidance on which similarity metrics, objective functions, or optimization algorithms to choose, Grimmer and Stewart (2013) caution social scientists from importing clustering methods developed in other, more technical fields like machine learning. The computer-assisted cluster analysis technique suggested by Grimmer and King (2011) offers a more intuitive tool for the task of fully automated cluster analysis.

The goal of dimensionality reduction is to shorten the number of terms in the term-document space while maintaining the structure of the corpus. One dimensionality reduction technique is principal component analysis, which transforms a document-term matrix into linearly uncorrelated variables that correspond to the latent semantic topics in the data set. The technique is not different from more conventional uses in multivariate modeling where a subset of variables are selected to represent a larger data set (Dunteman, 1989). A related dimensionality reduction technique is multidimensional scaling, which projects a corpus of documents into *N*-dimensional space such that the distances between documents correspond to dissimilarities between them. These methods provide good intuition of the topics that characterize a corpus of text but are best used as exploratory techniques; principal component analysis, in particular, is a typical data reduction step performed prior to subsequent, more substantive analysis.

Finally, the goal of topic modeling is to represent each document as a mixture of topics. Each topic is a probability mass function over words that reflect a distinct information domain. For instance, the topic 'foreign policy' may assign high probabilities to words such as 'war', 'treaty', and 'Iraq'; while the topic 'economy' may assign high probabilities to words such as 'unemployment', 'GDP', and 'labor'. The most widely used topic model is called latent Dirichlet allocation (LDA) (Blei et al., 2003). This technique has recently been applied to the analysis of newspaper content to dissect the framing of policies (DiMaggio et al., 2013). The method provides a new computational lens into the structure of texts and, as the authors state:

> finding the right lens is different than evaluating a statistical model based on a population sample. The point is not to correctly estimate population parameters, but to identify the lens through which one can see the data more clearly. Just as different lenses may be more appropriate for long-distance or middlerc1ngc vision, different models may be more appropriate depending on theanalyst's substantive focus. (ibid.: 20)

Again, the crucial step in the analysis comes with validation; that is, with the substantive interpretation of the themes identified.

Network Representations of Text

As the sections above have illustrated, content analysis is essentially a relational exercise: words that relate to the same topic are associated by co-appearing frequently in documents and they tend to cluster; likewise, positive words tend to be connected to other positive words, and as shown above, language connectors might change the affective tone of words by setting them in a different linguistic context. Networks offer a mathematical representation of the relational nature of language, and provide yet another tool for the analysis of its structure. Networks have been used to model narratives, and to analyze identity formation (Bearman and Stovel, 2000); to represent mental models (Carley and Palmquist, 1992); and to map semantic associations (Borge-Holthoefer and Arenas, 2010). A network approach has also been used with Twitter data to identify entities by looking at the co-occurrence of words and the clusters that emerge from those connections (Mathiesen et al., 2012). The nodes in these networks are words; what changes depending on the approach is the definition of the links that connect those words: co-occurrence is one of the options, but links can also be used to track the temporal evolution of narratives, as when political movements change their framing or candidates change their positions during an election campaign. These networks can be constructed and visualized using standard network analysis tools.

One of the by-products of generating a dictionary-based lexicon (discussed above) is that the method also creates a network of words that researchers can use to label the strength as well as the sign of the sentiment expressed. For example, Kamps et al. (2004) determined the strength and sentiment of words according to their distances in WordNet from labeled seed words; in this case, two words are linked if they are synonyms, and distance is measured as the number of links that need to be crossed to go from one word to another. Blair-Goldensohn et al. (2008) also used WordNet to construct a network of positive, negative, and neutral sentiment words, and then labeled the strength of the words using a propagation algorithm: starting from a seed word, its sign (positive, negative, or neutral) is propagated to all its neighboring words in the network (its synonyms); following a majority rule, that sign is further propagated to the neighbors of the neighbors, and so on, recursively, until all words have a sign assigned - the valence of which gets weaker the further apart the word is from its seed. These network-based techniques help to extend the dictionary-based approach by suggesting measures of sentiment strength.

APPLICATIONS TO THE ANALYSIS OF ONLINE POLITICAL COMMUNICATION

Sentiment in Online Political Talk

When applying sentiment analysis to political communication, it is important to remember that different methods inherit different assumptions from psychological theories of emotions The ANEW lexicon, for instance, derives from now classic psychological research suggesting that three dimensions account for variance in the expression of emotion: valence (which ranges from pleasant to unpleasant), arousal (which ranges from calm to excited), and dominance (which ranges from domination to control; Osgood et al., 1957). Neurological research, on the other hand, suggests that five emotional dimensions underlie most brain activity: fear, disgust, anger,

happiness, and sadness (Murphy et al., 2003). Reducing the breadth of human emotions to just a few dimensions is arguably a crude simplification, but necessary to make problems tractable; however, it also introduces measurement error that has to be taken into consideration when operationalizing research questions about the affective tone of political communication.

Sentiment analysis of online political communication must take into account not only measurement error but also sampling bias. Internet users, and in particular those present in social media, are typically not representative of the population: they tend to be female, young, and urban (Duggan and Brenner, 2013); in addition, the bias might be more or less important depending on the context and subject of communication. For some dimensions of public opinion, the bias might not matter, but for others it can be crucial. Again, it is only through validity tests that the measures of public opinion extracted from online communication can be relied upon (Grimmer and Stewart, 2013). The increasing number of Internet users who join social media sites and discuss politics means that the volume of online political communication is growing, and the profile of users involved is changing. Analyses of how on line sentiment changes over time must therefore account for these non-stationary characteristics, typically by comparing short, adjacent periods of online communication rather than the entire history of communication on a given site.

The assumptions made by automated methods about emotional mechanisms and the nature of the samples analyzed demand a thoughtful research design when studying on line communication. In many cases basic methods produce useful results that rival more sophisticated approaches; in particular, simple word frequencies and analysis of how the volume of communication fluctuates over time often yield good insights while preserving efficiency. Carvalho et al. (2011), for instance, found that in some cases these basic descriptive statistics predict sentiment as accurately as more advanced statistical techniques. This suggests that exploratory analysis can be crucial to avoid rushing into the implementation of more complex solutions when a simpler, more intuitive approach would perform as well.

In addition to the lexicons introduced above, a number of alternative approaches have also been developed to facilitate the study of online communication. These include OpinionFinder (OF), which rates expressions as strongly or weakly subjective (Wilson et al., 2005); and the Profile of Mood States (POMS) questionnaire (Lorr et al., 2003), in which respondents rate each of 65 adjectives on a five-point scale. The questionnaire produces emotion scores in six dimensions: Tension-Anxiety, Anger-Hostility, Fatigue-Inertia, Depression-Dejection, Vigor-Activity, and Confusion-Bewilderment. Like ANEW, the POMS lexicon is suited for analyzing more complex emotions in online communication; the OF lexicon, like LIWC, is used for simpler tasks such as the identification of polarity in sentiment analysis. Other prominent lexicons optimized for the analysis of online communication include SentiWordNet (Adrea and Sebastiani, 2006) and SentiStrength (Thelwall et al., 2010). SentiStrength is particularly useful for online political communication because it includes misspellings and emoticons which abound in online talk.

Recent empirical applications of these approaches include Connor et al. (2010), Bollen et al. (2011) and Castillo et al. (2013). Connor et al. (2010) derive sentiment valence from Twitter posts using a subjectivity lexicon based on a two-step polarity classification. They compare Twitter sentiment to consumer confidence and election polling data. They find high correlations

(between 0.7 and 0.8) and evidence that smoothed Twitter sentiment predicts consumer confidence (but not election) poll results with relatively high accuracy. However, Bollen et al. (2011) find that the intersection of a tweet corpus and their subjectivity lexicon is not a good leading indicator of the direction of shifts in the Dow Jones Industrial Average. This highlights how sentiment analysis of online communication may not work in all contexts: some lexicons are better suited to particular problem domains, such as consumer confidence, but not financial markets. Finally, Castillo et al. (2013) apply the SentiStrength lexicon to measure sentiment in cable news coverage; although this is traditional media content, the data were accessed through a software company that develops applications for smartphones and tablets that display extra information about TV shows, including captions of content.

Unsupervised Learning Applications

As explained above, many unsupervised learning methods are used as exploratory tools rather than testing techniques, and thus are less common in published literature on online communication. Nevertheless, a few prominent examples exist, although many are still peripheral to the core research questions of political communication.

Turney (2002), for instance, classifies online reviews as positive or negative by estimating the semantic orientation of sentences containing adjectives or adverbs. Specifically, the paper makes use of the pointwise mutual information-information retrieval (PMI-IR) algorithm to measure the number of co-occurrences between words and the seed words 'excellent' and 'poor' on Alta Vista search engine results. This co-occurrence frequency determines the semantic orientation of words, and thus can be used to rate online reviews as positive or negative.

Quinn et al. (2010) use a technique similar to LDA in order to analyze the daily legislative attention given to various topics in 118000 United States Senate floor speeches from 1997 to 2004. They found 42 topics, the most prominent being legislative procedures, armed forces, social welfare, environment, and commercial infrastructure. Yano et al. (2009) use LDA in order to model topics in political blog posts and their corresponding comments sections. They found five topics: religion, (election) primary, Iraq War, energy, and domestic policy. Associated with each topic are a set of words that appeared in blog posts and a set of words that appeared in comments. Additionally, the authors predict which users are likely to comment on which blogs. Finally, another recent example applies the same method to the analysis of issue salience in the Russian blogosphere (Kolstova and Koltcov, 2013). The authors use the method to identify a shift in topics during the political protests that took place during the parliamentary and presidential elections in late 2011 and early 2012.

In sum, unsupervised methods are less frequently used because they are exploratory techniques employed to charter communication domains that lack predefined boundaries. They are good for estimating the structure of a corpus of text when no a priori classifications exist, but they still require a posteriori theoretical and subjective labeling of categories. This stands in contrast to supervised techniques: whereas manual annotation is the starting point for supervised techniques, it is the ending point for the unsupervised approach.

FUTURE LINES OF WORK

This chapter has given an overview of techniques for the automated analysis of large-scale texts, especially as they are generated in online communication. Although this is a massive area of research, and is fast evolving, a few facts have already been established. One is the consistent evidence that the effectiveness of automated classifiers is not independent from the communication domain being analyzed: the meaning of words or their emotional load varies with the context in which they are used. More work is required to build tailored dictionaries that can capture the nuances of political communication as it takes place in different information contexts; for this, supervised-learning approaches offer the most accurate (and promising) solutions. Likewise, more work is needed to consolidate validation strategies, for instance by measuring the strength to which online measures of public opinion are correlated with more traditional measures, such as polls and surveys. A more systematic account of the efficiency and robustness of different algorithms is also needed: some corpora of text are better analyzed by certain techniques than others. Supervised methods, for instance, are more appropriate for content expressed in Twitter messages, whereas for longer communication, such as blog entries, unsupervised methods might be more appropriate. More research is needed to assess the robustness of each method for different data sources, as facilitated by online communication. In any case, the appropriateness of each technique has to be assessed in the light of each particular research question.

Validation is a crucial step in the application of automated content analysis, and this implies finding ways of assessing the accuracy, precision, and reliability of automated classifiers as compared to human coding. For instance, researchers who choose a lexicon-based approach face several design considerations. The first is what type of lexicon to generate or adapt. Some sentiment lexicons have binary categories (positive and negative), some have ternary categories (positive, negative, and neutral), and some have ordinal categories (-5 to +5, for example). A second design consideration is what word features to include in a lexicon. Some lexicons simply have word valence (ranging from positive to negative), while others have additional features such as arousal (ranging from calm to excited). The type of sentiment lexicon a researcher chooses should be based on the features a lexicon offers and the specific research question they seek to answer.

Researchers who choose a lexicon-based approach also face several implementation considerations. Most of these have to do with how to detect and resolve the complexities of text. The algorithm that implements a lexicon-based approach should often not just naively match words but also be sensitive to their local context. For instance, it should be aware of negating words (such as 'no', 'not', and 'none') and strengthening punctuation (such as exclamation marks, question marks, and ellipses). Some of the lexicons revised in this chapter, such as SentiStrength, already take these language modifiers and intensifiers into account. Good lexicon-based approaches to sentiment detection do not just rely on word matching: they are also sensitive to how the local context of each word affects the overall sentiment.

The advantage of automated content analysis is that it helps to scale up the amount of text analyzed by lowering the costs of coding and the efficiency of document classification; but it still needs to be reliable. Many sentiment lexicons are based on psychological theories of language

use but it is still unclear whether these psychometric instruments work for written communication and large-scale text analysis. In addition, these techniques are still not very good at capturing essential features of political talk, such as sarcasm. A document may contain many strong sentiment words but the author might actually have intended the opposite sentiment to that captured by the automated approach. This means that automated methods might be more appropriate when applied to text in which sarcasm and figurative language are rarely used, for instance news reports; communication through social media, on the other hand, might be more vulnerable to measurement error. As the tools for automated content analysis become more prevalent in communication research, more unified standards for evaluation and assessment will have to be consolidated. The advantages of automated methods are, overall, too great to dismiss.

ACKNOWLEDGEMENTS
We would like to thank the participants of the workshop 'Extracting Public Opinion Indicators from Online Communication', sponsored by the Oxford John Fell Fund under project 113/365 while the authors were based at the University of Oxford. We are especially grateful to Scott Blinder, Javier Borge-Holthoefer, Andreas Kaltenbrunner, Patrick McSharry, Karo Moilanen, and Georgios Paltoglou for insightful discussions.

LEARNING MORE
Methods for automated content analysis are fast evolving, and any list of available resources is likely to be soon outdated. What follows are a few recommendations on where to start to learn more about the methods and applications of automated tools. Rather than an exhaustive list, these references offer entry points to what is a vast and quickly expanding area of research.

FURTHER READING
Krippendorff (2013 [1980]). Now in its third edition, this book is a classic in content analysis, a long-standing reference that precedes the explosion of automated methods for the analysis of large-scale data. Even though the book does not consider emerging methods, the discussion on validity and reliability still applies.

Liu (2012). This monograph is one of the most up-to-date reviews of opinion mining methods. It offers an accessible discussion of state-of-the art tools for automated content analysis, and it defines basic terminology as well as research standards.

Dilubler et al. (2012). This research note offers an interesting comparison of the validity of automated versus human coding in identifying basic units of text analysis. The discussion considers how automated methods offer an improvement to human coding schemes without loss of validity

Grimmer and Stewart (2013). This article offers an interesting overview of methods that analyze text at the document level In addition to discussing in an accessible way the basic features of different approaches, the article also emphasizes the need to develop new validation methods.

Dodds and Danforth (2009). One of the first examples that used unsupervised methods to extrapolate opinion measures from large-scale communication. It offers a good schematic

example of how unsupervised methods work, and how it can be applied to several data sets to track aggregated sentiment dynamics.

Tools for Content Analysis

• R packages:
- ReadMe: http://gking.harvard.edu/readme
- TextMining: http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf
- LDA Topic Modeling: http://www.cs.princeton.edu/-blei/topicmodeling.html
- TextTools: http://www.rtexttools.com/

• Other software:
- LexiCoder: http://www.lexicoder.com
- SentiStrength: http://sentistrength.wlv.ac. uk
- LIWC: http://www.liwc.net.

REFERENCES

Adrea, E. and Sebastiani, F. (2006). SentiWordNet: a publicly available lexical resource for opinion mining. Paper presented at the 5th Conference on Language Resources and Evaluation, Genoa, Italy.

Bearman, P. and Stovel, K. (2000). Becoming a Nazi: a model for narrative networks, *Poetics*, 27(1), 69-90.

Berinsky, A.J., Huber, G.A. and Lenz, G.S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3), 351-368. doi: 10.1093/pan/mpr057.

Bishop, C.M, (1995). *Neural Networks for Pattern Recognition.* New York: Oxford University Press.

Blair-Goldensohn, S., Hannan, K., McDonald, R ., Neylon, T., Reis, G.A. and Reynar, J. (2008). Building a sentiment summarizer for local service reviews. WWW Workshop on NLP in the Information Explosion Era, Beijing.

Blei, D. (2013). Topic modeling and digital humanities. *Journal of Digital Humanities*, 2(1).

Blei, D., Ng, A. and Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning and Research*, 3, 993-1022.

Bollen, J., Mao, H., and Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.

Borge-Holthoefer, J. and Arenas, A. (2010). Semantic networks: structure and dynamics. *Entropy*, 12(5), 1264-1302.

Bradley, M.M. and Lang, PJ. (1999). *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings.* Gainesville, FL.

Breiman, L. (200 I). Random forests. *Machine learning,* 45(1), 5-32. doi: 10.1023/A:1010933404324.

Brody, S. and Diakopoulos, N. (2011). Cooooooooooooooollllllllllllllllllllllllllll: using word lengthening to detect sentiment in microblogs. Paper presented at the Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA.

Budge, I. (ed.). (2001). *Mapping Policy Preferences. Estimates for Parties, Electors and Governments 1945-1998.* Oxford: Oxford University Press.

Carley, K. and Palmquist, M. (1992). Extracting, representing, and analyzing mental models. *Social Forces,* 70(3), 601-637. doi: 10.2307/2579746.

Carvalho, P., Sarmento, L., Teixeira, J. and Silva, M.J. (2011). Liars and saviors in a sentiment annotated corpus of comments to political debates. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.* Stroudsburg, PA: Association for Computational Linguistics.

Castillo, C., de Francisci Morales, G., Mendoza, M. and Khan, N. (2013). Says who? Automatic text-based content analysis of television news. Paper presented at the MNLP Workshop, ACM International Conference on Information and Knowledge Management, San Francisco, CA.

Connor, B.O., Balasubramanyan, R., Routledge, B.R. and Smith, N.A. (2010). From Tweets to polls: linking text sentiment to public opinion time series. *Fourth International AAAI Conference of Weblogs and Social Media.* Menlo Park, CA: AAAI.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning,* 20(3) 273-297. doi: 10.1023/A:1022627411411.

Däubler, T., Benoit, K., Mikhaylov, S. and Laver, M. (2012). Natural sentences as valid units for coded political texts. *British Journal of Political Science.* 42(04), 937-951.

Della Porta, D. and Diani, M. (2006). *Social Movements: An Introduction,* 2nd edn. London; Wiley-Blackwell

Derks, D., Bos, A.E.R. and von Grumbkow, J. (2007). Emoticons and online message interpretation. *Social Science Computer Review,* 26, 379·-388. doi: 10.1177/0894439307311611.

DiMaggio, P., Nag, M. and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of government arts funding in the US. *Poetics,* 41(6), 570-606.

Dodds, P.S. and Danforth, C.M. (2009). Measuring the happiness or large-scale written expression: songs, blogs, and presidents. *Journal of Happiness Studies,* DOI: 10.1007/ s10902-009-9150-9.

Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A. and Danforth, C.M. (2011). Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. *PloS ONE.* 6(12), e26752. doi: 10.1371/journal.pone.0026752.

Duggan, M. and Brenner, J. (2013). The demographics or social media users. http://www.pewinternet.org/Reports/2013/Social-media-users.aspx.

Dunteman, G.H. (1989). *Principal Components Analysis.* London: Sage.

Eshbaugh-Soha, M. (2010). The tone of local presidential news coverage. *Political Communication,* 27(2). 121-140. doi: 10.1080/10584600903502623.

Franzosi, R. (2004). *From Words to Numbers. Narrative, Data, and Social Sciences.* Cambridge: Cambridge University Press.

Gentzkow, M. and Shapiro, J.M. (2010). What drives media slant? Evidence from US daily newspapers. *Econometrica,* 78(1), 35-71.

Gonzalez-Bailon, S., Banchs, R.E. and Kaltenbrunner, A. (2012). Emotions, public opinion and US presidential approval rates: a 5-year analysis of online political discussions. *Human Communication Research,* 38, 121-143.

Gonzalez-Bailon, S., Paltoglou, G. (2015). Signals of public opinion in online communication: a comparison of methods and data sources, *The Annals of the American Academy of Political and Social Science,* in press.

Grimmer, J. and King G., (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences,* 108(7), 2643-2650.

Grimmer, J. and Stewart, B. (2013), Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysts*, 21(3), 267-297.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* 2nd edn. New York: Springer.

Hopkins, Daniel and King, Gary (2010). Extracting systematic social science meaning from text. *American Journal of Political Science,* 54(1), 229-247.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD- 2014).* New York City: ACM Press.

Johnston, H. and Noakes, J.A. (eds) (2005). *Frames of Protest. Social Movements and the Framing Perspective.* Lanham, MD: Rowman & Littlefield.

Jurafsky, Dan and Martin, James (2009), *Speech and Natural Language Processing: An Introduction to Natural Language Processing, Computational linguistics, and Speech Recognition.* Upper Saddle River, NJ: Prentice Hall,

Kamps, J., Marx, M., Mokken, R.J. and de Rijke, M. (2004). Using WordNet to measure semantic orientations of adjectives. Paper presented at the LREC. http://dblp.uni-trier.de/db/conf/lrec/lrec2004.html#KampsMMR04.

Kolstova, O. and Koltcov, S. (2013). Mapping the public agenda with topic modelling: the case of the Russian LiveJournal. *Policy and Internet,* 5(2), 207-227.

Krippendorff, K. (2013 [1980]). *Content Analysis. An Introduction to its Methodology,* Los Angeles, CA: Sage.

Krippendorff, K. and Bock, M.A. (eds) (2008). *The Content Analysis Reader.* Thousand Oaks, CA: Sage.

Laver, M., Benoit, K. and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review,* 97(2), 311-331.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Chicago, IL: Morgan & Claypool.

Lorr, M., McNair, D.M., Heuchert, J.W.P. and Droppleman, L.F, (2003), *POMS: Profile of Mood States.* Toronto: MHS.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance,* 66(1), 35-65.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* 1: 281-297. London: Cambridge University Press.

Maron, M.E. and Kuhns, J.L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM,* 7(3), 216-244. doi: 10.1145/321033.321035.

Mathiesen, J., Yde, P. and Jensen, M.H. (2012). Modular networks of word correlations on Twitter. *Scientific Reports,* 2.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J. (1990). Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography,* 3(4), 235-244.

Murphy, F.C., Nimmo-Smith, I. and Lawrence, A.D. (2003). Functional neuroanatomy of emotions: a meta-analysis. *Cognitive, Affective, and Behavioral Neuroscience,* 3, 207-233.

Neuendorf, K. (2001). *The Content Analysis Guidebook.* London: Sage.

Osgood, C.E., Suci, G.J. and Tannenbaum, P.H. (1957). *The Measurement of Meaning*, Vol. 47. Urbana, IL, University of Illinois Press.

Paltoglou, G., Gobron, S., Skowron, M., Thelwall, M. and Thalmann, D. (2010). Sentiment analysis of informal textual communication in cyberspace. *Proc. ENGAGE*, 13-23.

Paltoglou, G. and Thelwall, M. (2012). Twitter, MySpace, Digg: unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):66: 1-66: 19.

Pennebaker, J.W., Booth, R.J. and Francis, M.E. (2001). *Linguistic Inquiry and Word Count: LIWC.* Mahwah, NJ: Erlbaum Publishers.

Porter, M. (1980). An algorithm for suffix stripping. *Program,* 14(3), 130-137.

Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H. and Radev, D.R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science,* 54(1), 209-228.

Spirling, A. (2012). US treaty making with American indians: institutional change and relative power. 1784-1991. *American Journal of Political Science,* 56(1), 84-97. doi: 10.1111/j.1540-5907.2011.00558.x.

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

Turney, P.D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Paper presented at the *Proceedings of the 40th Annual Meeting of Association for Computational Linguistics*, Philadelphia, PA.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe,J., Choi, Y., ... Patwardhan, S. (2005). OpinionFinder: a system for subjectivity analysis. *Proceedings of HLT/EMNLP on Interactive Demonstrations*, Vancouver, British Columbia, Canada.

Yano, T., Cohen, W.W. and Smith, N.A. (2009). Predicting response to political blog posts with topic models. *Human language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL* (pp. 477-485). Association for Computational Linguistics.

Young, L. and Soroka, S. (2012). Affective news: the automated coding of sentiment in political texts. *Political Communication,* 29, 205-231. doi: 10.1080/10584609.2012.671234.