Studying Political Decision Making With Automatic Text Analysis

Wouter van Atteveldt, Kasper Welbers
Department of Communication Science, VU Amsterdam

Mariken A.C.G. van der Velden
Department of Political Science, University of Zurich

Abstract. Analyzing political text can answer many pressing questions in political science, from understanding political ideology to mapping the effects of censorship in authoritarian states. This makes the study of political text and speech an important part of the political science methodological toolbox. The confluence of increasing availability of large digital text collections, plentiful computational power, and methodological innovations has lead to many researchers adopting techniques of automatic text analysis for coding and analyzing textual data. In what is sometimes termed the "text as data" approach, texts are converted to a numerical representation, and various statistical techniques such as dictionary analysis, automatic scaling, topic modeling, and machine learning are used to find patterns in and test hypotheses on these data.

These methods all make certain assumptions and need to be validated to assess their fitness for any particular task and domain. This chapter aims to contribute to the adoption and correct use of automatic text analysis techniques. We review the steps and main methods used to gather, transform, analyze, and validate texts. In particular, we describe dictionary methods and supervised and unsupervised machine learning approaches. Special attention is paid to the best practices in validating applications of each technique. We also provide an overview of applications of text analysis in political science, looking at how research in political behaviour, comparative politics, and international relations uses automatic text analysis techniques to answer important substantive questions. Finally, we look at future directions in text analysis research that are now starting to be used in political research such as crowd coding, deep learning, and word embeddings; and the reader is given starting points to further explore the relevant literature from political methodology, computational linguistics, and computer science.

Keywords: text analysis, automatic text analysis, methodology, sentiment analysis, machine learning, topic modeling, scaling

Text analysis in Political Science Research

"Words are an integral part of politics" (Wilkerson & Casas, 2017, p.530), and analyzing political text can answer many pressing questions in political science: Why do some people still reject the 'consensus view' that the Earth is getting warmer due to human activity (Boussalis & Coan, 2016)? Can the government's infringements of the rights of journalists tell us anything about its wider human rights agenda (Gohdes & Carey, 2017)? Why did the Stability and Growth Pact fail to prevent the Euro crisis (Baerg & Hallerberg, 2016)? Do simple campaign messages resonate with voters' information about parties (Bischof & Senninger, 2018)? Does a party leader represent the median preferences of the party's membership (Z. Greene & Haber, 2016)? How do autocratic leaders differ in their communication compared to democratic leaders (Barberá, Gohdes, Zeitzoff, & Iakhnis, 2018)? What is on the political agenda (D. Greene & Cross, 2017)? Do parties change their platform in anticipation of electoral losses (van der Velden, Schumacher, & Vis, 2017)?

The study of political text and speech is an important part of the political science methodological toolbox (for overviews, see Alvarez, 2016; Cardie & Wilkerson, 2008; Grimmer & Stewart, 2013; Monroe & Schrodt, 2008). The confluence of increasing availability of large digital text collections, plentiful computational power, and methodological innovations has lead to many researchers adopting techniques of automatic text analysis for coding and analyzing textual data (van Atteveldt & Peng, 2018). In what is sometimes termed the "text as data" approach, texts are converted to a numerical representation, and various statistical techniques such as dictionary analysis (e.g. Pennebaker, Francis, & Booth, 2001; Young & Soroka, 2012), automatic scaling (Benoit & Laver, 2008; Lowe & Benoit, 2013; Slapin & Proksch, 2008), topic modeling (Blei, Ng, & Jordan, 2003;

Published as:
van Atteveldt, W., Welbers, K., Van der Velden, M.A.C.G (2019) Studying Political Decision Making With Automatic Text Analysis, *Oxford Encyclopedia of Political Decision Making*. https://doi.org/10.1093/acrefore/9780190228637.013.957

Grimmer, 2010; Roberts et al., 2014), and machine learning (D'Orazio, Landis, Palmer, & Schrodt, 2014; Hillard, Purpura, & Wilkerson, 2008) are used to find patterns in and test hypotheses on these data. Most of the methods discussed in this chapter are based on word frequencies without looking at their specific context, the so-called 'bag of words' assumption. In the final section, we also discuss more sophisticated analyses, that are emerging based on recently developed techniques in Artificial Intelligence and Computational Linguistics, such as Word Embeddings and Deep Learning (Goldberg, 2017).

Despite the word 'automatic' in automatic text analysis, this does not imply little researcher effort, nor does it mean that manual coding becomes superfluous. In fact, although running an off-the-shelf topic modeling algorithm on an existing corpus can be done in minutes, it takes a lot of effort to prepare, and especially, validate the outcome of these methods (Grimmer & Stewart, 2013). For dictionary and other rule-based analyses holds the same: the dictionary needs to be either created, or, when using an off-the-shelf dictionary, it needs to be validated and often adapted to the researcher's specific goal. Manual coding is required to create validation material, and supervised machine learning approaches also require a substantial amount of coded training examples. Once this initial work is done, however, the techniques can generally scale with negligible extra effort per document, making it much cheaper to code large text collections without having to rely on samples.

This chapter aims to contribute to the adoption and correct use of automatic text analysis techniques. The next section will review the steps in automatic text analysis and the main methods used for each step. Subsequently, we provide an overview of applications of text analysis in political research. The last section discusses the main possibilities and pitfalls for automatic text analysis in political research and briefly discusses some promising techniques that are currently being developed in artificial intelligence and computational linguistics that will become important for the automatic analysis of political text in the near future.

Steps in automatic content analysis

Figure 1 illustrates the main steps in automatic text analysis. As reviewed by Wilkerson and Casas (2017), automatic text analysis generally consists of four steps: (1) *Obtaining* text; (2) *Transforming* the texts to structured data; (3) *Analyzing* the data to get substantive measurements; and (4) *Evaluating* or validating the analysis.

Obtaining text

Obtaining text is in general a domain- or task dependent process, which often follows one of the three approaches described below. Regardless of the approach, it is important to take copyright and terms-of-access into account, which can restrict the gathering as well as archiving and publishing of data. Moreover, relying on proprietary data sets fosters a reliance on (often commercial) third parties and can cause problems of representativeness and consistency that are difficult to quantify without access to the full data set (Boyd & Crawford, 2012; Lazer et al., 2009; van Atteveldt & Peng, 2018; H. Wallach, 2016)

Databases. Databases with existing collections such as Lexis Nexis or parliamentary archives provide easy access to many texts. Recently, the Harvard Dataverse, an open source software application to share, cite and archive data, has invited scholars to share their textual data collections (e.g. Rauh, de Wilde, & Schwalbach, 2017; Schoonvelde, Traber, Dahiya, & De Vries, 2016). If the data for a study is available in a collection, this is the most convenient solution.

API.. An Application Programming Interface (API) is a set of methods for communicating with a computer program. Many online platforms, such as Facebook, the New York Times, and the Dutch Royal Library, provide an API that can be used to access their data. For example, the Guardian API allows users to search news articles by sending the query (date, keywords, etc.) in a URL, and returns the data in a structured machine-readable format (title, text, author, date, etc.). If available, an API provides easy access to data, but be aware that APIs are often not intended for academic research, and access can be restricted at the discretion of the online platform (cf. Bruns et al., 2018).

Websites. To retrieve information from websites, researchers typically build a so-called scraper. A (web) scraper is an algorithm that automatically navigates a website to collect information. This makes it possible to extract lots of data from a website if no (suitable) API is available. A disadvantage is that scrapers generally need to be programmed for specific websites. Mistakes can result in incomplete data, technical artifacts, irrelevant text such as advertisement or navigational components, and problems with character encodings. This means that it is important to check the results of scraping by comparing a sample of documents to their original source, and by checking e.g. the most common words to make sure there are no boilerplate terms or incorrectly decoded characters.

Transforming text to structured data

In the second step, the digital text is transformed into a numerical representation that captures the aspects of the text that might be relevant for the analysis. In many cases, this representation is a *document-term matrix*, in which the rows represent documents, the columns represent terms (words), and the cells list the frequency of each term in each document. This is also called a 'bag of words' representation because it disregards the ordering of or relations between words. Although this seems overly simplistic, as language is much richer than the frequencies of word use, this repre-

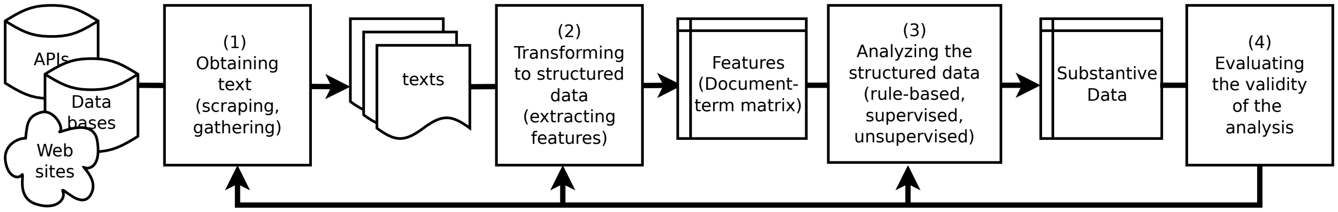

Figure 1. Steps in the Automatic Analysis of Political Text

sentation in fact contains a lot of information on topic, tone, and frames present in the text.

Creating this matrix requires several *preprocessing* steps, and there are many choices that can be made that change how texts are represented as data, and thus affect the analysis (Denny & Spirling, 2018; Z. Greene, Ceron, Fazekas, & Schumacher, 2016). First, the text is split into words or *tokens* in a process called tokenization. In western languages, tokenization into words is for the most part simple, because words are clearly delimited by whitespace and punctuation. Tokenization is more challenging, however, for languages without explicit word delimiters such as Chinese, Japanese and Korean; and for morphologically complex languages such as Arabic that require clitic tokenization (Diab, 2009; Webster & Kit, 1992). Thus, depending on the task and language, more advanced and language specific tokenizers can be required.

Words can occur in different forms with roughly the same meaning, such as different verb conjugations ('walk', 'walking', 'walk-s'), and with or without capitalization ('The', 'the'). It can be more informative and reduce computational load to count different forms of a word together as a single column. A common preprocessing step is therefore to make all text lowercase, and to reduce words to a basic form using *stemming* or *lemmatization*. Words can also be filtered out to remove noise and reduce computational load. The easiest approach is to remove predefined lists of 'stopwords' such as articles and pronouns (Griffiths & Steyvers, 2004; Yang & Pedersen, 1997), although of course it depends on the research question whether a word is meaningful or not. For example, if one looks at ingroup/outgroup discourse one might wish to keep words such as 'us' and 'them', but most predefined stopwords lists remove these words. One can also use statistical measures to remove words, for example removing words that are not informative because they are very rare or very common.

It is also possible to use linguistic processing to enrich text before the analysis, for example to identify the parts-of-speech of words (e.g. nouns and verbs) and named entities such as persons or organizations. It is also possible to use features that combine words, such as using n-grams (pairs or triplets of words) or to identify specific collocations such as "prime minister" or "State of the Union."

For scholars working with non-English data, the availability of tools for preprocessing can be more limited, and depending on the language care should be taken when dealing with character encoding and text direction. Also, techniques like stemming and whitespace-based tokenization can be more difficult in more morphologically complex languages such as Arabic or German (Diab, 2009; Webster & Kit, 1992). It is also possible to use automatic translation tools to translate the text to English before processing (de Vries, Schoonvelde, & Schumacher, 2017), but this is only as good as the automatic translation tool, so results should be validated thoroughly and it will probably not perform as well as developing specific analysis tools for the target language.

Analyzing the structured data

Many techniques and approaches are used to analyse the text as represented in (generally) a document term matrix. Following Boumans and Trilling (2016), these can be roughly categorized into three approaches:

Dictionaries and rule-based approaches. The oldest and possibly most transparent techniques of automatic text analysis are rule based methods. In a rule based method the researcher explicitly instructs the computer to look for certain words or patterns, and interprets the resulting outcome as a measurement. These techniques range from simple keyword searches (e.g. search how often a party or person name is mentioned) to elaborate dictionaries with complex contextual or syntactic rules to identify issues or frames. Although dictionaries require no manual coding except for validation, it can take significant effort to develop a high-quality dictionary. Dictionary or keyword analysis is also often used to select documents for analysis. King, Lam, and Roberts (2017), however, caution against bias in such approaches since researchers often overlook frequently used synonyms of the concepts they intend to sample on.

Supervised machine learning. Where in rule-based approaches the researcher supplies the rules to determine the output given the input, in supervised machine learning the researcher supplies coded training examples from which the relation between input and output are automatically 'learned' by an algorithm such as *naive bayes* or *support vector machines* (Michalski, Carbonell, & Mitchell, 2013). In other words, manually annotated (coded) training data is used to build a statistical model, which is then used to 'predict' the values (codes) for unannotated data using word frequencies (or other features) as predictors (independent variables).

The crucial step for applying supervised machine learning is acquiring good annotated data for training and testing the model. To guarantee that the model and the validity tests are representative for a specific corpus, a sufficiently large sample of the corpus could be manually annotated. A cheaper alternative is to use an existing annotated corpus. For example, there are large corpora of product reviews, in which users have given both verbal evaluation and a quantitative (star) rating. Since these corpora contain both text and rating, they can be used to train a sentiment classifier (Aue & Gamon, 2005). Of course, if the domain or task of the researcher is different from that of the original data source, the classifier might not perform as well and in any case the trained model should be validated on texts from the actual research domain before using.

Unsupervised approaches. In contrast to the deductive nature of dictionary and supervised methods, unsupervised text analysis techniques are more closely related to an inductive or bottom-up approach. An algorithm is used to train a model or calculate certain statistics about a corpus without human supervision. Choices in the model that is used and parameter settings can determine what type of output is generated, but there is no a priori defined set of concepts (Denny & Spirling, 2018). The data is thus approached with a more or less blank slate, with the purpose of finding new patterns that could be meaningful. By investigating and interpreting the patterns, the analyst can acquire new insights into the data, and by validating and labeling patterns the data can be coded bottom-up.

The most commonly used family of unsupervised text analysis methods are topic models (Blei et al., 2003). The basic LDA topic model assumes that each document can contain multiple topics, and each word can also occur in multiple topics. Given the desired number of topics, the algorithm then estimates the most likely distribution of words over topics and topics over documents based on the corpus. For example, Jacobi, Van Atteveldt, and Welbers (2015) automatically estimate the subtopics (or frames) in the nuclear technology discourse since 1945 using an LDA topic model. While it is possible to estimate the number of topics by running multiple models and comparing model fit (see e.g., Griffiths & Steyvers, 2004), which leads to good predictive models, this does not necessarily lead to the most useful topics for human interpretation (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009). There are many different extensions on the basic LDA topic model, such as *correlated topic models* (Blei & Lafferty, 2005), *hierarchical topic models* (Griffiths, Jordan, Tenenbaum, & Blei, 2004), *dynamic topic models* (Blei & Lafferty, 2006), and *structural topic models* (Roberts et al., 2014). These extended models enable the direct modeling of statistical relations between topics and between topics and covariates such as time or author.

In the end, unsupervised approaches are simply tools for finding patterns of word use, that can only correlate with meaningful concepts such as topics and ideological dimensions. It is still a human task to make sense of these patterns. Thus, while unsupervised approaches require minimal assumptions and are fast and cheap to deploy (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010), using them effectively still requires extensive and systematic interpretation through some form of manual coding (Grimmer & Stewart, 2013; Maier et al., 2018).

Evaluating the validity of the analysis

The importance of validation for automatic text analysis cannot be emphasized enough. When used incorrectly, automatic text analysis techniques can easily lead to wrong or biased results. Where possible, established procedures for testing validity should be followed to prove that a method accurately measures what the researcher claims to measure, within a reasonable margin of error. Even if an existing and previously validated method is used the researcher needs to establish that it also works on the current data (Grimmer & Stewart, 2013). Next to the method of analysis itself, choices in data-sampling and pre-processing strategies can also have large implications for the overall results (Barberá et al., 2018; Denny & Spirling, 2018; Z. Greene et al., 2016).

For dictionaries and rule-based methods, the most straightforward procedure is to compare the results to a set of texts for which the *correct* outcomes are known. This set, also called the *gold standard*, is ideally created by conducting a manual content analysis, following established best practices for safeguarding reliability (Krippendorff, 2012), to ensure that the outcomes reflect what the researcher wants to measure as accurately as possible. To compare the results, common measures are precision and recall for classification tasks or correlation for continuous outcomes (Gilbert, 2014; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011).

The use of a gold standard and these measures is also common practice for evaluating supervised machine learning models. However, the approach is different, because the (manually) annotated data that is used to test the validity is generally drawn from the same data that is used to train the model. For a fair test, it is then important not to use the same cases for both training and testing at the same time. One way is to achieve this is by splitting the annotated data into a *training set* and a *test set* (also called *held-out set*). A variation of this approach is *k-fold cross-validation*, in which the annotated data is split into a given number (k) subsets. The model is then trained on all but one subset, and the remaining (held-out) subset is used for testing. This is repeated k times, so that all subsets have been used as both training and test sets, but without overlapping training and test cases.

Validation of unsupervised approaches is less straightforward (Maier et al., 2018). Since the concepts of interest are not defined a priori, it is not possible to develop a gold standard. Instead, the researcher can interpret the patterns that are found to determine whether they correlate with meaningful concepts. However, this is prone to error because humans excel at finding meaning in patterns, even if these patterns are partially or largely random. See Chang et al. (2009) for a discussion of techniques to quantitatively assess the robustness and coherence of topic models.

Applications

This section reviews applications of the general steps and methods for automatic text analysis as described in the previous section. In this review, we discuss how automatic text analysis over the three main subfields of political science is used to the tackle important challenges and answer old and new theoretical questions.

Political Behavior

Investigating whether political candidates or political parties are able to mobilize voters with their campaigns, and if so, on which issues they do so, which tone of voice they use, and how that affects voters and other candidates has been a key topic in political science. Scholars of political behavior have used various methodological approaches—survey research, interviews, and experiments in the lab or field—to determine how and when campaigns matter for political outcomes (for overview, see Jacobson, 2015).

Automatic sentiment analysis is the general name for tools to automatically classify the tone of text, and is an active subfield of computational linguistics (Liu, 2012; Pang & Lee, 2008; Wiebe, Wilson, Bruce, Bell, & Martin, 2004). Although there are earlier studies of automatic sentiment analysis, progress was limited until the emergence of large digital corpora of subjective texts, especially online reviews, in the mid-2000's (Mäntylä, Graziotin, & Kuutila, 2018). As reviewed by Liu (2012), the two main current approaches to automatic sentiment analysis are using dictionaries and supervised machine learning. Some words, like 'horrible', have an obviously negative valence while other words (such as 'fantastic') are strongly positive. Thus, it makes sense to construct a dictionary of such negative and positive terms, and such a lexical approach to sentiment analysis has been used since at least the General Inquirer (Stone, Bayles, Namerwirth, & Ogilvie, 1962), an early tool for dictionary-based automatic text analysis. For applications of these dictionaries to understand campaign rhetoric, see e.g. Young and Soroka (2012), Brundidge, Scott, Choi, and Muddiman (2014), and Jones et al. (2018). Dictionaries do not always yield valid results, however. Especially for sentiment analysis, the results of different dictionaries do not correlate well with each other or with human coding (González-Bailón & Paltoglou, 2015; Soroka, Young, & Balmas, 2015). Barberá, Boydstun, Linn, McMahon, and Nagler (2016) show that machine learning significantly outperforms existing dictionaries for estimating the tone of news coverage of the economy.

Next to the tone of the campaign, parties or candidates can also vary the content of their message by either taking a position on a political issue or by selectively emphasizing some issues over others. Besides expert surveys (e.g. Bakker et al., 2015), the most common way to measure the policy position of political parties, is the analysis of manifestos and other documents in which ideological preferences are expressed. There is a long tradition of manual analysis in the Comparative Manifesto Project (Budge, Klingemann, et al., 2001; Klingemann, Volkens, McDonald, Budge, & Bara, 2006), and scholars are also using *unsupervised* and *supervised* scaling methods to automatically determine ideological positions from such texts. *Wordscores* (Laver, Benoit, & Garry, 2003) is a supervised scaling technique developed to replaces the hand-coding of texts with computerized coding schemes. By using reference documents to 'anchor' the scales, for example placing manifestos of left-wing and right-wing parties from a previous election to place new manifestos on a left-right scale, the algorithm uses word frequencies to estimate similarities. Slapin and Proksch (2008) developed this algorithm further into *Wordfish*, an unsupervised scaling technique to estimate positions based on word frequencies in political texts without using reference texts. These scaling methods have been used to estimate party and government positioning to explain party competition and political representation (e.g. Debus, 2008; Hakhverdian, 2009; Laver, Benoit, & Sauger, 2006). As reviewed by Lowe (2008, 2013), however, word scores and other scaling algorithms depend on fairly strong assumptions about the distribution of words and texts that we know are violated in most cases. Moreover, the inability to distinguish ideologically centrist terms from terms without ideological value mean that the scaled texts are strongly biased towards the center, and they need to be rescaled before being compared with the reference texts. Scholars have therefore recently employed crowd-sourced approaches to estimate positions from text (Benoit, Conway, Lauderdale, Laver, & Mikhaylov, 2016; Lehman & Zobel, 2017). Next to positions, scholars have recently also turned to automated text analysis approaches to explain the salience of the issues pronounced by parties (see e.g. van der Velden et al., 2017 and Sagarzazu & Klüver, 2017).

Besides substantive messages, parties and candidates try to make their messages persuasive by varying their language in different ways. The comparative work of Aalberg, Esser, Reinemann, Strömbäck, and De Vreese (2016) demonstrates that right-wing populist have a distinct style of communication—that is using more simplistic language and more emotive language. These findings have been investigated using automated textual approaches. That is, a dictionary approach for the level of emotionality (e.g. see van der Velden, van Atteveldt, & Vis, 2018—for other approaches of emotions in text, see e.g. Valentino & Vanderbroek, 2014 and Valentino, Neuner, & Vandenbroek, 2018) and the Flesch–Kincaid scale—a readability scale designed to indicate how difficult a passage in English is to understand based on the length of the words and sentences used—to determine whether populists differentiate the level of difficulty of their speech (Bischof & Senninger, 2018).

Comparative Politics

Asking what is on the political agenda, how legislators try to influence the agenda, the effect of the agenda on other domains, such as the economy or the environment, and whether the answers differ over various political contexts are central questions in comparative politics that can be answered with (automatic) text analysis methods.

The selective emphasis of certain topics is an important part of the political agenda as expressed by an actor. Texts about a topic like foreign affairs will in general use different words from texts about the environment, making a list of keywords (dictionary or keyword analysis) a plausible method for determining the topic of a text. For example, Atkinson, Lovett, and Baumgartner (2014) created a list of 90 keyword searches spanning the 19 major topic areas of the Comparative Agendas Project codebook, and this has also been done for other languages (Zoizner, Sheafer, & Walgrave, 2017). Strong advantages of this approach are transparency and ease of use, but since the dictionary must be created manually it can be hard to find all possible synonyms of a topic without also including false positives. To overcome the difficulties of manually creating and validating a dictionary for coding topics, scholars have used *supervised* techniques. In the Comparative Agendas Project, this has been used for automatically determining the topic of political communication, both for legislative speech and for news articles (Burscher, Vliegenthart, & De Vreese, 2015; Hillard et al., 2008).

Scholars have also used *unsupervised* techniques such as topic modeling (Grimmer, 2010) to investigate the political agenda. For example, D. Greene and Cross (2017) used a dynamic topic model to analyze how the political agenda of the EU parliament evolved over time. Quinn et al. (2010) applied unsupervised text clustering to investigate attention to political topics in the US Senate, while Ceron (2015) used similar methods to look at factions in a party and how this influences the political agenda. Using the unsupervised scaling method called *Wordfish*, Proksch and Slapin (2010) show that parties may actively seek to prevent some members from taking the floor while promoting opportunities for others to control the message that their partisans convey in parliament.

Moreover, not only legislative actors have agenda's that are of interest to political scientists. Scholars of political economy have utilized automated text approaches to investigate the preferences of economic actors, such as the European Central Bank, by combining topic and scaling models (Baerg & Lowe, 2018). Using machine learning methods, Baerg and Hallerberg (2016) demonstrates that European Member States with Euroskeptic populations are more successful in weakening the European Commission, and thereby contributed to the failure of the Stability and Growth pact. Students interested in lobby groups, moreover, also apply automated textual approached to estimate the effect of lobbying on changing the political agenda, oftentimes combining topic models with supervised machine learning approaches (for example, see Boussalis, Coan, & Holman, 2018, Boussalis & Coan, 2016, and Peiffer & Boussalis, 2010).

International Relations

In international relations, automatic analysis of news sources has been used for studying and forecasting conflict for over 30 years (Schrodt, 2015; Schrodt & Gerner, 1994). The KEDS system (later named TABARI) used an extensive set of word-order rules and a dictionary to automatically extract conflict events. Its successor, PETRARCH, uses a more linguistically sophisticated approach, using syntactic analysis of sentences to distinguish between e.g. attacker and attacked. van Atteveldt and Peng (2018) use a similar approach for automatic *clause analysis*, extracting clauses consisting of a source, agent (who acts), and predicate (what is done) from newspaper articles.

Of particular interest to scholars of international relations is how the development of Internet technologies influences conflict and state repression. We know that authoritarian governments around the world develop sophisticated technologies for controlling information, for example by blocking access to social media (Gohdes, 2018; Roberts, 2018). Roberts (2018) demonstrates that even censorship that is easy to circumvent can still be enormously effective. Taking advantage of digital data harvested from the Chinese Internet and leaks from China's Propaganda Department, she combines various automated text techniques and sheds light on how and when censorship influences the Chinese public. Gohdes (2015) demonstrates how to apply various supervised machine learning approaches to measure the number of undocumented conflict fatalities prior to and during network blackouts (see also Gohdes & Carey, 2017). Using all social media posts (Facebook and Twitter) published by any head of state or government in all U.N.-member countries, Barberá et al. (2018) test the diversionary theory of foreign policy and the relationship between regime type and responsiveness to domestic publics using automated translation and machine learning methods.

Conclusion and Future Direction

Automatic text analysis has become an established methodological approach in political science research, and there is a broad and diverse range of techniques that can be used to go from text to data that can be used to answer interesting political science questions. The goal of this chapter is to contribute to the knowledge and correct use these techniques, of which we covered a variety of applications and approaches. Perhaps most importantly, the chapter shows that the various techniques that are used serve different purposes and have many trade-offs and limitations. Ultimately, whether or not an analysis has been performed correctly comes down to proper validity testing. The measurements obtained via the automated analysis need to correspond to what the researcher is trying to measure. This holds for supervised learning techniques, where validation is often an integral step of applying the method; but it holds just as much for unsupervised learning and dictionary techniques, which are (too) often presented at face validity.

In all cases, the choice of method should depend on the substantive research question. This implies that a researcher needs to be knowledgeable about various techniques and trade-offs for different stages of a project, from selecting the method, to preparing the data, applying the right tools, and interpreting the results. This requires a certain level of computational skills, which are not always integrated in the political science curriculum. Fortunately, modern languages such as R and Python make it much easier to learn the computational skills needed for performing cutting-edge techniques (Heiberger & Riebling, 2016; Welbers, Van Atteveldt, & Benoit, 2017).

The field of automatic text analysis is evolving fast. This chapter focused on common techniques in the recent political science literature, but there are many interesting development on the horizon that are worth pursuing. There are two relatively new approaches for computer-assisted manual text annotation. First, there is *crowd coding*, where a simple coding task is uploaded to a platform (e.g., MTurk, Figure Eight), and a crowd of coders can apply to perform the task. Although untrained crowd coders have higher variability than trained (undergraduate) students (Barberá et al., 2016), their much lower cost and greater convenience allows multiple codings per unit, which can for some tasks increase overall validity and give an estimate of spread as well as average tone (Barberá et al., 2016; Benoit et al., 2016; Weber et al., 2018). Second, there is *active learning*, where machine learning and human annotation are combined into an iterative workflow (Hillard et al., 2008). A machine learning model is first trained using a (small) set of training data. The active learning algorithm then selects which cases need to be annotated next to maximize the learning rate (Sculley, 2007; Smailović, Grčar, Lavrač, & Žnidaršič, 2014). The human annotator annotates these cases, the model is updated, and the cycle repeats. This allows the researcher to achieve accurate classification with fewer manually coded examples, which is especially relevant in cases where some categories are underrepresented (Wiedemann, 2018).

In the field of supervised machine learning, new models are constantly being developed to achieve better predictions. In recent years, *neural networks* that can be trained on very large data sets have been used to get highly accurate results on a variety of tasks (Goldberg, 2017). In automatic text analysis, convolutional and recurrent deep neural networks for instance show improvements for sentiment classification (Nakov, Ritter, Rosenthal, Sebastiani, & Stoyanov, 2016; Socher et al., 2013; Tang, Qin, & Liu, 2015) and detecting political ideology (Iyyer, Enns, Boyd-Graber, & Resnik, 2014).

A related development is the use of word embeddings as a technique for converting text to data. Word embeddings map words into a lower dimensional vector space, where words with semantically similar meanings are closer together. The widely used *word2vec* approach (Mikolov, Chen, Corrado, & Dean, 2013) creates these embeddings by training a shallow neural network to predict a word based on its context of previous and next words (the continuous bag of words model), or by predicting the context based on the word (the skip-gram model). This creates "a surprisingly rich semantic encoding of relations and analogies" (Szegedy et al., 2013, 1), that can for instance be used to assist dictionary creation (Rice & Zorn, 2013), or as a pre-trained word embeddings layer in a neural network (Goldberg, 2017; Nakov et al., 2016).

The future of automatic text analysis in political science and related fields will also be shaped by developments at an institutional level (van Atteveldt & Peng, 2018). In particular, the development and adaptation of techniques depends on institutional incentives for developing and sharing data and tools and for computational skills training. Development and sharing of tools and data is a crucial part of methodological innovation in automatic text analysis and should be accepted as independent scientific contributions (Crosas, King, Honaker, & Sweeney, 2015; Lazer et al., 2009; H. Wallach, 2016). As a discipline we should make sure that the tools and data that we depend on remain accessible, since relying on proprietary data and closed tools endangers reproducibility and replicability (Boyd & Crawford, 2012; Lazer et al., 2009)

Further reading

For a general discussion of automatic text analysis methodology, Grimmer and Stewart (2013) is the best starting point. When using topic models, H. M. Wallach, Mimno, and McCallum (2009) gives an overview of evaluation methods and challenges. For applying these methods using R, see Welbers et al. (2017) for a description and tutorial of text analysis methods and Roberts et al. (2014) for structural topic models. If needed, Wickham and Grolemund (2016) gives a general overview of data handling in R.

To learn more about natural language processing, see Manning and Schütze (1999), Jurafsky and Martin (2009) and Bender (2013). See Michalski et al. (2013) for a general introduction to machine learning, and Goldberg (2017) for an introduction to deep learning for natural language processing. See Liu (2012) for an overview of sentiment analysis methods.

References

Aalberg, T., Esser, F., Reinemann, C., Strömbäck, J., & De Vreese, C. H. (2016). *Populist political communication in europe*. Routledge.

Alvarez, M. R. (2016). *Computational social science*. Cambridge University Press.

Atkinson, M. L., Lovett, J., & Baumgartner, F. R. (2014). Measuring the media agenda. *Political Communication*, *31*(2), 355–380.

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (ranlp)* (Vol. 1, pp. 2–1).

Baerg, N. R., & Hallerberg, M. (2016). Explaining instability in the stability and growth pact: The contribution of member state power and euroskepticism to the euro crisis. *Comparative Political Studies*, *49*(7), 968–1009.

Baerg, N. R., & Lowe, W. (2018). *A textual taylor rule: Estimating central bank preferences combining topic and scaling methods.* Working paper.

Bakker, R., De Vries, C., Edwards, E., Hooghe, L., Jolly, S., Marks, G., … Vachudova, M. A. (2015). Measuring party positions in europe: The chapel hill expert survey trend file, 1999–2010. *Party Politics*, *21*(1), 143–152.

Barberá, P., Boydstun, A., Linn, S., McMahon, R., & Nagler, J. (2016). Methodological challenges in estimating tone: Application to news coverage of the us economy. In *Meeting of the midwest political science association, chicago, il.*

Barberá, P., Gohdes, A. R., Zeitzoff, T., & Iakhnis, E. (2018). *From riot police to tweets: How world leaders use social media during contentious politics.* Working paper.

Bender, E. M. (2013). Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. *Synthesis Lectures on Human Language Technologies*, *6*(3), 1–184.

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295.

Benoit, K., & Laver, M. (2008). Compared to what? a comment on "a robust transformation procedure for interpreting political text" by martin and vanberg. *Political Analysis*, *16*(1), 101–111.

Bischof, D., & Senninger, R. (2018). Simple politics for the people? complexity in campaign messages and political knowledge. *European Journal of Political Research*, *47*(2), 473–495.

Blei, D. M., & Lafferty, J. D. (2005). Correlated topic models. In *Proceedings of the 18th international conference on neural information processing systems* (pp. 147–154).

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *the Journal of machine Learning research*, *3*, 993–1022.

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

Boussalis, C., & Coan, T. G. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change*, *36*, 89–100.

Boussalis, C., Coan, T. G., & Holman, M. R. (2018). Climate change communication from cities in the usa. *Climatic Change*, *Forthcoming*, 1–15.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662–679.

Brundidge, J., Scott, A. R., Choi, S., & Muddiman, A. (2014). The "deliberative digital divide": Opinion leadership and integrative complexity in the u.s. political blogosphere. *Political Psychology*, *35*(6), 741–755.

Bruns, A., Bechmann, A., Burgess, J., Chadwick, A., Clark, L. S., Dutton, W. H., … others (2018). Facebook shuts the gate after the horse has bolted, and hurts real research in the process. *Internet Policy Review*.

Budge, I., Klingemann, H.-D., et al. (2001). *Mapping policy preferences: estimates for parties, electors, and governments, 1945-1998* (Vol. 1). Oxford University Press on Demand.

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *The Annals of the American Academy of Political and Social Science*, *659*(1), 122–131.

Cardie, C., & Wilkerson, J. (2008). Text annotation for political science research. *Journal of Information Technology & Politics*, *5*(1), 1-6.

Ceron, A. (2015). Brave rebels stay home: Assessing the effect of intra-party ideological heterogeneity and party whip on roll-call votes. *Party Politics*, *21*(2), 246–258.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei,

D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Neural information processing systems* (pp. 288–296).

Crosas, M., King, G., Honaker, J., & Sweeney, L. (2015). Automating open science for big data. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 260–273.

Debus, M. (2008). Party competition and government formation in multilevel settings: Evidence from germany. *Government and Opposition*, *43*(4), 505–830.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis*, 1–22.

de Vries, E., Schoonvelde, M., & Schumacher, G. (2017). *No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications.* Working paper.

Diab, M. (2009). Second generation amira tools for arabic processing: Fast and robust tokenization, pos tagging, and base phrase chunking. In *2nd international conference on arabic language resources and tools* (Vol. 110).

D'Orazio, V., Landis, S. T., Palmer, G., & Schrodt, P. (2014). Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *Political Analysis*, mpt030.

Gilbert, C. H. E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international conference on weblogs and social media (icwsm-14).*

Gohdes, A. R. (2015). Pulling the plug: Network disruptions and violence in civil conflict. *Journal of Peace Research*, *52*(3), 352–367.

Gohdes, A. R. (2018). Studying the internet and violent conflict. *onflict Management and Peace Science*, *35*(1), 89–106.

Gohdes, A. R., & Carey, S. C. (2017). Canaries in a coalmine? what the killings of journalists tell us about future repression. *Journal of Peace Research*, *54*(2), 157–174.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1–309.

González-Bailón, 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*, *659*(1), 95–107.

Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, *25*(1), 77–94.

Greene, Z., Ceron, A., Fazekas, Z., & Schumacher, G. (2016). The nuts and bolts of automated text analysis. comparing different document pre-processing techniques in four countries. In *Open science framework* (pp. 1–23).

Greene, Z., & Haber, M. (2016). Leadership competition and disagreement at party national congresses. *British Journal of Political Science*, *46*(3), 611–632.

Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., & Blei, D. M. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems* (pp. 17–24).

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, *18*(1), 1–35.

Grimmer, J., & Stewart, B. M. (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

Hakhverdian, A. (2009). Capturing government policy on the left–right scale: Evidence from the united kingdom, 1956–2006. *Political Studies*, *57*(4), 720–745.

Heiberger, R. H., & Riebling, J. R. (2016). Installing computational social science: Facing the challenges of new information and communication technologies in social science. *Methodological Innovations*, *9*, 2059799115622763.

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, *4*(4), 31–46.

Iyyer, M., Enns, P., Boyd-Graber, J., & Resnik, P. (2014). Political ideology detection using recursive neural networks. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (Vol. 1, pp. 1113–1122).

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modeling. *Digital Journalism*, *4*(1), 89—106.

Jacobson, G. C. (2015). How do campaigns matter? *Annual Review of Political Science*, *18*, 31–47.

Jones, K. L., Noorbaloochi, S., Jost, J. J., Bonneau, R., Nagler, J., & Tucker, J. A. (2018). Liberal and conservative values: What we can learn from congressional tweets. *Political Psychology*, *39*(2), 423–443.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing (2nd edition).* Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

King, G., Lam, P., & Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, *61*(4), 971–988.

Klingemann, H.-D., Volkens, A., McDonald, M. D., Budge, I., & Bara, J. (2006). *Mapping policy preferences ii: estimates for parties, electors, and governments in eastern europe, european union, and oecd 1990-2003* (Vol. 2). Oxford University Press on Demand.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

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

Laver, M., Benoit, K., & Sauger, N. (2006). Policy competition in the 2002 french legislative and presidential elections. *European Journal of Political Research*, *45*(4), 667–697.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721.

Lehman, P., & Zobel, M. (2017). Positions and saliency of immigration in party manifestos: A novel dataset using crowd coding. *European Journal of Political Research*, *Forthcoming*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1–167.

Lowe, W. (2008). Understanding wordscores. *Political Analysis*, *16*(4), 356–371.

Lowe, W. (2013). *Putting it all on the line: Some unified theory for text scaling.* Presented at the American Political Science Association meeting, Chicago.

Lowe, W., & Benoit, K. (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, mpt002.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... Adam, S. (2018). Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2-3), 93–118.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16–32.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Monroe, B. L., & Schrodt, P. (2008). Introduction to the special issue: The statistical analysis of political text. *Political Analysis*, *16*(4), 351–355.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)* (pp. 1–18).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135.

Peiffer, C. A., & Boussalis, C. (2010). Foreign assistance and the struggle against hiv/aids in the developing world. *The Journal of Development Studies*, *46*(3), 556–573.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Proksch, S.-O., & Slapin, J. B. (2010). Position taking in european parliament speeches. *British Journal of Political Science*, *40*(3), 587–611.

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

Rauh, C., de Wilde, P., & Schwalbach, J. (2017). *The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states.* Working paper.

Rice, D. R., & Zorn, C. (2013). Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Proceedings of NDATAD*, 98–115.

Roberts, M. E. (2018). *Censored. distraction and diversion inside china's great firewall*. Princeton University Press.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Sagarzazu, I., & Klüver, H. (2017). Coalition governments and party competition: Political communication strategies of coalition parties. *Political Science Research and Methods*, *5*(2), 333–349.

Schoonvelde, M., Traber, D., Dahiya, T., & De Vries, E. (2016). Euspeech: a new dataset of eu elite speeches. In *Proceedings of the international conference on the advances in computational analysis of political text* (pp. 75–80).

Schrodt, P. (2015). *Comparing methods for generating large scale political event data sets.* Paper presented at the Text as Data meetings, New York University, 16-17 October 2015.

Schrodt, P., & Gerner, D. J. (1994). Validity assessment of a machine-coded event data set for the Middle East, 1982–1992. *American Journal of Political Science*, *38(3)*, 825–854.

Sculley, D. (2007). Online active learning methods for fast label-efficient spam filtering. In *Ceas* (Vol. 7, p. 143).

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705–722.

Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information sciences*, *285*, 181–203.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642).

Soroka, S., Young, L., & Balmas, M. (2015). Bad news or mad news? sentiment scoring of negativity, fear, and anger in news content. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 108–121.

Stone, P. J., Bayles, R. F., Namerwirth, J. Z., & Ogilvie, D. M. (1962). The General Inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267–307.

Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1422–1432).

Valentino, N. A., Neuner, F. G., & Vandenbroek, L. M. (2018). The changing norms of racial political rhetoric and the end of racial priming. *The Journal of Politics*, *80*(3), 000–000.

Valentino, N. A., & Vanderbroek, L. M. (2014). Political communication, information processing, and social groups. In *The oxford handbook of political communication.* Oxford University Press.

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2-3), 81–92.

van der Velden, M. A., Schumacher, G., & Vis, B. (2017). Living in the past or living in the future? analyzing parties' platform change in between elections, the netherlands 1997–2014. *Political Communication*, *Forthcoming*, 1–20.

van der Velden, M. A., van Atteveldt, W., & Vis, B. (2018). *Tabloidization of the Political Discourse?* Working paper.

Wallach, H. (2016). Computational social science: Towards a collaborative future. In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction* (p. 307). Cambridge University Press.

Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking lda: Why priors matter. In *Advances in neural information processing systems* (pp. 1973–1981).

Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., … Tamborini, R. (2018). Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures*, *12*(2-3), 119–139.

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in nlp. In *Proceedings of the 14th conference on computational linguistics-volume 4* (pp. 1106–1110).

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in r. *Communication Methods and Measures*, *11*(4), 245–265.

Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".

Wiebe, J., Wilson, T., Bruce, R. F., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics*, *30(3)*, 277–308.

Wiedemann, G. (2018). Proportional classification revisited: Automatic content analysis of political manifestos using active learning. *Social Science Computer Review*, 0894439318758389.

Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, *20*, 529–544.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, pp. 412–420).

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*(2), 205–231.

Zoizner, A., Sheafer, T., & Walgrave, S. (2017). How politicians' attitudes and goals moderate political agenda setting by the media. *The International Journal of Press/Politics*, *22*(4), 431-449.