To appear in The Handbook of Research Methods in Consumer Psychology

Text Analysis in Consumer Research: An Overview and Tutorial

Matthew D. Rocklage and Derek D. Rucker

Kellogg School of Management, Northwestern University, 2211 Campus Drive, Evanston, IL 60208, USA

Language saturates the marketplace. We use it to convey meaning and to spread information about products, services, and brands. Marketers use it to position and sell their products and consumers use it to assess and evaluate them. Although language has long been a part of the lives of consumers, the recent emergence of behavioral traces left online has ushered forth a “big data” revolution with incredible opportunities for researchers. Amazon, TripAdvisor, and Yelp alone are repositories for over half a billion reviews and people send more than 500 million tweets on Twitter per day (Krikorian, 2013).

Researchers have taken note. Consumer behavior research now routinely makes use of data that ranges from hundreds to millions of pieces of text. For example, researchers have used language to understand what is shared and spreads (Berger & Milkman, 2012), what makes for an impactful online review (Moore, 2015), how people attempt to persuade one another (Rocklage, Rucker, & Nordgren, 2018a), and even common themes present across consumer behavior research itself (Wang, Bendle, Mai, & Cotte, 2015).

However, these opportunities also bring challenges. How do researchers quantify language in a sensible manner? How are these measures to be validated? Is one approach to linguistic analysis just as good as any other? When should researchers use one technique versus another? Indeed, despite the rapid growth of work utilizing language in the last 10 years, few resources exist to guide researchers on to how to integrate language into their projects (but see Humphreys and Wang (2018) for a recent discussion of automated text analysis).

In this chapter, we offer an overview of key methods used to quantify language and their ability to offer insights into consumer behavior. Specifically, we overview three major approaches to text analysis: manual coding, top-down dictionary-based approaches, and bottom-up data-driven approaches. We explain the basic logic behind each approach and its subdivisions, provide an introductory “how-to” for each approach, and conclude with a discussion of each approach’s pros and cons. Table 1 provides a brief summary of the approaches explored in this chapter.

Manual Coding

Overview

Manual coding often involves the use of two or more judges to rate a construct in text. Manual coding can be used for both continuous judgments (e.g., how positive a movie review is) as well as categorical judgments (e.g., whether the review references a previous movie).

As an example of manual coding for continuous judgments, Barasch and Berger (2014) examined how consumers altered their communications as a function of whether they shared information with one person (“narrowcasting”) versus a group of people (“broadcasting”). In particular, they tested whether information shared with a large group of people increased individuals’ self-presentational concerns and thus led them to reframe negative events in their day as more positive. To that end, Barasch and Berger asked two independent coders to rate each participant’s writing for the extent to which the participant reframed negative events to make themselves look less bad (1: not at all; 5: a great deal). They found that, indeed, participants who were told they would be telling a group of people about their day were more likely to reframe negative events in a more positive light compared to those who believed they would be sharing information with one person.

As an example of manual coding for categorical judgments, Moore (2015) instructed two independent raters to categorize sentences in online reviews for the kind of explanation they provided – *actions* versus *reactions*. Action explanations were defined as instances where reviewers explained why they behaved the way they did (e.g., “I bought this [non-fiction] book, because it was the sequel”). Reaction explanations were defined as instances where reviewers focused on how they responded to a product or service (e.g., “I loved this book, as I found it to be refreshing”). Moore found systematic differences in the coded responses such that reviewers naturally explained their *actions* more often for utilitarian products and explained their *reactions* more often for hedonic products.

Introductory How-To

At a general level, the steps to manually coding text are to 1) conceptualize one’s construct and specify how it will be measured, 2) train coders on how to rate the text, 3) have the coders rate the text, and 4) assess how well the coders performed based on measures of reliability. Additionally, based on the construct, researchers may also need to provide further evidence that the ratings are valid.

To elaborate, researchers must first conceptualize how their construct of interest will be measured in the text. For instance, as an initial step, researchers must decide whether the construct should be operationalized as a continuum as in Barasch and Berger (2014) or via a categorical judgment as in Moore (2015). This decision can be informed by how the past literature has conceptualized the construct, researchers’ own understanding of the construct, and the perceived value of a continuous versus categorical judgment.

Next, and perhaps most importantly, researchers must make this conceptualization concrete so that it can be communicated to coders. This step often involves detailed instructions to coders, explicit examples, and then a training period. To begin, researchers explain to coders the construct they wish to measure and provide examples coders can use to practice. For instance, after explaining the construct, researchers may use sample text from their data, ask coders to rate this text, and then assess the consistency of their ratings. If the coders are discrepant on particular items, they can be given further instructions to refine their understanding.

This training stage is important as it enhances both construct validity and the reliability of coders' ratings. In terms of construct validity, even apparently straightforward constructs might be more complicated when actual coding is discussed. Take, for example, the measurement of the positivity of a person's text. Should coders rate how positive the topic of the text is or should they rate how positive the person is in their opinion? While these two may often go together, they can also diverge – some people may be quite negative toward cuddly kittens due to allergies for instance, even though cuddly kittens are a normatively positive topic. In terms of coders' reliability, the better researchers can explicate the construct, the more reliable ratings should be, which will lead to more consistent results (Stanley & Spence, 2014).

After coders have rated the texts, researchers need to quantify the reliability of the coders' ratings to ensure they measured the construct of interest consistently. For continuous ratings, the intraclass correlation (ICC) is a typical statistic for two or more coders. Several different forms of ICCs exist and step-by-step guidance on the appropriate form for one's research objectives can be found in McGraw and Wong (1996) as well as Koo and Li (2016). Across the different forms, a coefficient of .40 or below is considered poor; .40 to .59 is fair; .60 to .74 is good; .75 to 1.00 is excellent (Cicchetti, 1994). ICCs can be calculated in standard statistical programs such as R and SPSS.

For categorical judgments with two coders, researchers often use Cohen's kappa (κ; Cohen, 1960). Researchers should aim for levels of reliability similar to those put forth for ICCs (Cicchetti, 1994). Kappa, however, can only be used with two coders and thus Krippendorff's alpha (α; Krippendorff, 2012) has become increasingly common as it can be used with any number of coders and can also be used when there are missing judgments from any given coder. Researchers may have more than two coders or missing data if, for instance, they have different coders judge subsets of the stimuli. This can occur when the set of stimuli to be rated is large or when the rating process is intensive. An alpha of .800 is recommended and .667 is considered the lowest limit for tentative conclusions (Krippendorff, 2012). Cohen's kappa can be readily calculated in R and SPSS, but Krippendorff's alpha requires an additional free macro when used with SPSS (Hayes & Krippendorff, 2007).

Finally, researchers may need to validate their ratings to demonstrate the ratings successfully measure the construct of interest. In the case where coders categorize sentences for whether or not they contain a noun, for example, judgments may not require further validation. However, in a number of cases, validation is likely to be required. For example, if coders are asked to rate how certain a person is via their text, these ratings might require additional validation because certainty might not be easily observable. One means to validate ratings is to correlate the ratings of the coders with other measures reported by the participant that are theoretically related to the participant's certainty (e.g., how extreme versus temperate participants' judgments are on that issue). Another means of validation would be to validate the manual coding procedure on a set of text known to differ in certainty – for example, word of mouth around a rumor versus fact – and then have coders rate the text of interest.

Pros and Cons

If researchers are interested in a one-time text analysis tool with a relatively small number of texts, manual coding can be useful as it is relatively quick, cheap, and requires little technical expertise. Moreover, even when the number of texts is large, researchers have used crowdsourcing via platforms such as Amazon.com’s Mechanical Turk (Tosti-Kharas & Conley, 2016).

Manual coding also allows for flexibility. It has the ability to be adapted for both relatively objective judgments (e.g., how many words are in the text, how many positive thoughts are present) as well as subjective judgments (e.g., readability of text, how persuasive the text is). Moreover, given that manual coding involves human coders, the coders themselves might be better able to adapt to unique text more easily compared to automated approaches. The coders also have the ability to alert the researcher who can then alter the coding instructions based on the unique text.

Even with the advent of crowdsourcing techniques, one limitation of manual coding is that it can become difficult to scale up one’s efforts when coding large numbers of texts. For instance, the consumer behavior literature is quickly becoming replete with examples of researchers using thousands, if not millions, of pieces of text (Berger & Milkman, 2012; Ghose, Ipeirotis, & Li, 2012; Spiller & Belogolova, 2017). Such coding is likely to be prohibitively time consuming and expensive.

Another obstacle with manual coding is that whereas some constructs may be easier for human coders to assess, others might be inherently difficult to judge. An example of this is construal level – or how abstractly versus concretely a person is thinking about and representing their environment (Trope & Liberman, 2003; Fujita, Henderson, Eng, Trope, & Liberman, 2006). Although manual coding schemes exist for measuring the abstractness of individuals’ language (Semin & Fiedler, 1988), it may be rather difficult to train coders on this scheme. More generally, one might find that efforts to validate a particular coding scheme fail, which might require a researcher to consider other methods.

When to Use

Manual coding is quick and cheap and presents relatively few obstacles in terms of technical expertise. As such, manual coding may be researchers' first choice when they conduct a small number of studies, use a limited number of smaller-sized texts, and are interested in a construct that is relatively easy for human coders to assess.

Top-Down Dictionary Approaches

A more automated approach to text analysis utilizes word lists or "dictionaries." This approach involves the assembly of a corpus of words that represents a construct (e.g., "amazing" and "effective" to represent positive evaluations). Researchers can then use an automated approach to search for these words in the text of interest and then quantify that text using different methods. Two prominent methods that involve dictionaries are word counts and imputation.

Word Counts

Overview. Word counts, in their simplest form, involve a count of the number of times a word of interest occurs in a piece of text. The principle behind this approach is that a given word is likely to provide a signal of a construct. For example, researchers interested in the positivity of a text might view any instance of the word "amazing" as indicative of a positive evaluation. Word counts can utilize a binary approach (whether the word is present or not), the number of times the word is used, or a frequency count (how often the word occurs out of the total number of words written). Although word counts can be conducted by human coders, the process is sufficiently easy to automate and can be conducted with spreadsheet programs (e.g., Excel) or text analysis programs such as Linguistic Inquiry and Word Count (LIWC; Pennebaker, Boyd, Jordan, & Blackburn, 2015).

As an example of a word count approach, Spiller and Belogolova (2017) examined the relationship between how often online reviewers referenced themselves (e.g., “I”) in an online review and whether these reviewers provided judgments that were less congruent – more idiosyncratic – compared with experts who judged that same product. To that end, the researchers calculated the frequency (percentage) of first-person pronouns in each review. They then used this percentage from the review to predict how discrepant that reviewer’s final star rating was from the expert consensus. Spiller and Belogolova found that the more frequently reviewers referenced themselves, the more discrepant their final judgments were from expert reviewers.

Researchers have also created dictionaries that can be applied across diverse projects. LIWC, for instance, is one of the more popular text analysis tools in the behavioral sciences and is a collection of dictionaries used to measure different constructs of interest (Pennebaker et al., 2015; Tausczik & Pennebaker, 2010). For example, to measure the positivity of a piece of text, LIWC uses a list of words that can be used to convey such positivity (e.g., “amazing,” “okay”). The more frequently these words appear in a piece of text, the more positive the person is thought to be. In the case of LIWC, the lists themselves are largely generated based on face validity as voted on by a set of two to eight external judges (Pennebaker et al., 2015). If a majority of the judges agree that a word (“amazing”) belongs to a given dictionary (positivity), then it is retained.

Introductory how-to. As overviewed, one of the benefits of a word count approach is that it is relatively easy to execute. Several ready-made dictionaries and easy-to-use software packages already exist that can be used to analyze one’s data. In some cases, researchers can simply copy their text into an Excel spreadsheet and use the “countif” function to search for words of interest in that text. An even more automated approach would be to utilize free software provided by Boyd (2016) called Recursive Inspection of Text (RIOT) Scan. RIOT Scan is preloaded with a number of existing dictionaries that rely primarily on word counts (https://riot.ryanb.cc/). As noted, LIWC is another example of an easy-to-use software package that contains a large set of existing dictionaries. LIWC offers the ability to measure anything from the presence of words related to health to temporal orientation to achievement (Pennebaker et al., 2015). Both RIOT Scan and LIWC also allow researchers to provide their own dictionaries to use for analysis as well. A drawback of LIWC, however, is that it is not free.

If researchers are interested in creating their own word list a number of steps are typically required. Specifically, researchers must 1) generate a list of candidate words that are likely to provide a signal of one's construct, 2) refine this list, and 3) validate the tool.

To elaborate, in terms of generating a list of candidate words, one typical approach researchers rely on is to create a list of words with high face validity – what can be called the "intuition" method of word generation. For instance, if researchers wish to measure positivity, they might begin by generating words that they have the intuition will signal positivity such as "amazing" and "fantastic." After generating an initial set of words, researchers can expand this initial word list via a thesaurus so as to capture as many words that signal positivity as possible.

The next step is to refine the word list by removing those words that are less likely to provide a clear signal of one's construct. As a continuation of the intuition approach, this is often accomplished by asking a set of outside raters – sometimes lay individuals and sometimes trained experts – to judge whether the word should be included or not. If a majority agrees that the word signals the construct, then the word is retained; otherwise, it may be removed. Take, for instance, the word "like." Whereas this word has the ability to signal a positive evaluation ("I like the product”), it can also be used as a comparison word (“This product is like another I’ve used”). Thus, the word “like” may be left out of the final word list. While the intuition approach to generation and refinement has been widely used, it relies on researchers’ and judges’ ability to imagine all the various ways in which a word will be used in real-world text. This is often a difficult task. Moreover, this approach will miss words that are not immediately obvious signals of one’s construct, but are nevertheless important. For example, using a data-driven approach, Schwartz and colleagues (2013) found that individuals who reference “sports” and “vacation” more often in their Facebook posts tend to be higher in emotional stability – a novel finding.

To forgo relying on intuition and simple face validity, recent work combines the word generation and refinement steps into a single step that utilizes data-driven methods. For example, Rocklage and colleagues (2018b) were interested in identifying evaluative words – words specifically used by people to express favorable or unfavorable opinions and attitudes. Rocklage and colleagues used a “pruning” approach to generate and refine a list of evaluative words by first taking a very large list of words and narrowing that word list down to words likely to signal an evaluation. Specifically, they began with 6.2 million words derived from a set of five sources (e.g., Amazon, TripAdvisor, and Yelp reviews) and then narrowed this list down to approximately 1,550 words that were most indicative of an opinion being expressed in text. For example, to prune their word list they used a series of regression equations to examine whether a word was a consistent predictor of positive versus negative online reviews via the reviewer’s self-reported final star rating. Words that failed to consistently predict reviewers’ final star rating were removed as they did not provide a reliable signal of reviewers’ evaluation (e.g., the word “fair” could be a positive evaluation of a person, but a lukewarm reaction to food and therefore would be removed).

The final step in the creation of the word list is validation. One approach to validating a word list is to use a series of texts that researchers have strong a priori reasons to believe will systematically differ in regard to their construct of interest. For example, Kacewicz and colleagues (2014) attempted to identify a text-based signal of social power. Among other possibilities, they reasoned that individuals with less social power would tend to use more tentative language (e.g., “maybe,” “perhaps”) compared to those with greater social power. To investigate this possibility, they examined texts where they knew the social power of the actor and examined whether differences in language emerged. Another approach to validating a word list would be to correlate self-reported levels of the construct with text that has been analyzed with one’s word list. For example, one could examine whether self-reported power covaries with the quantified text (but see Boyd and Pennebaker (2017) for limitations of this approach).

Pros and cons. Word count approaches offer several advantages. First, they tend to be relatively easy to automate and implement. Once words are identified, it is a rather straightforward task to count the number of times a word appears in a piece of text and this can even be conducted quite easily using existing spreadsheet software (e.g., using Excel and its “countif” function).

Another advantage of word count approaches is that, given their simplicity, they often lend themselves to high face validity. The words that are used and the reasoning for their use are often clear to readers and reviewers based on the construct that researchers are interested in measuring. For example, if one is interested in measuring how positive a person is, searching for the word “amazing” has strong face validity. Of course, as with manual coding, diligence and care is required to validate one’s tool.

Another advantage of word counts, particularly in comparison to manual coding, is that they are easier to reproduce, both for one's own analyses as well as for other researchers' data. Indeed, given a static word list, it is easy to recode one's own data and reproduce the analyses as well as to pass along the word list for others to use. In this way, variability on the side of coding is eliminated, whereas this could present an obstacle with manual coding where different coders from different regions, cultures, education levels, and life experiences might view the same text differently.

A final advantage of word count approaches is that they are already prevalent and included with a number of ready-made text analysis tools. Given a large set of existing text analysis tools rely on word counts, they are rather easy to pick up and use.

Despite these advantages, word counts have limitations. Although they have high face validity, word count approaches may perform poorly when applied to real-world text if adequate care is not taken. For example, a construct can be signaled in different ways across different topics. Although the word "horrifying" is very often used to signal a negative opinion and thus may be included on a word list for measuring negativity, it can also be used in a positive manner when individuals refer to a horror film they delight in. Thus, researchers may potentially arrive at incorrect conclusions based on their analyses. To mitigate this issue, the word list can be constructed and then validated across a wide range of topics. Similarly, it is possible that people may *negate* the words in real-world text – a person may say that a movie was "*not* amazing." It is currently ambiguous exactly how to best code negations; for example, is "not amazing" considered negative or just less positive than "amazing"? Researchers have taken varied approaches to this issue. One conservative approach is to ignore any word that is preceded by common negations (e.g., "not," "isn't," "wasn't," etc). To summarize the broader point, however, it is important that researchers not rely on the face validity of their measure and to directly validate it for use in natural text.

Another drawback of word count approaches is how the absence of a construct is coded in a piece of text. Word counts assign a value of 0 to pieces of text that do not contain the construct. In essence, missing data do not exist when using word count approaches because the *absence* of a word is thought to be indicative of a low degree of that construct. However, it is not obvious that the absence of a word implying positivity, for example, indicates that the person is not positive in his/her assessment. A piece of text might be coded as 0 on positivity for any number of reasons: it could be because the person is truly low on that construct (e.g., low in positivity), one's dictionary is missing a given word, the text being analyzed is not relevant to the construct (e.g., analyzing text on how positive the person is when the text contains only factual information), and so on. In essence, a score of 0 confounds a truly low level of a construct – the person is not very positive or is even negative – with any number of alternative reasons for why the text may have scored a 0. Thus, researchers may conclude that low positivity is predictive of their variable of interest when, in fact, the signal is simply insufficient to be picked up in the dataset.

As a consequence of assigning 0 to a piece of text in these cases, this approach can also lead to statistical problems. For instance, a sizeable number of cases can be assigned to 0 due to low baserates of the construct or due to short pieces of text. Thus, there can be decreased variability and a skew in one's data (see Table 3 in Pennebaker et al. (2015) for examples of categories with low baserates; e.g., anxiety, anger, sadness). Although a given word count approach may be extremely accurate, the data may lack a strong signal of the construct of interest. Such a skew presents a challenge for statistical techniques that rely on the assumption of normally distributed data. In these situations, the mean of the data is a poor representation of the central tendency.

Another limitation of word count approaches is that they treat each word as equally indicative of a construct. Using positivity as an example, a word count approach would treat the words “okay” and “amazing” as equally indicative of how positive a person was, even if “amazing” is diagnostic of a more positive evaluation. In essence, word count approaches risk the loss of a great deal of information that can be communicated via language and may conflate mild levels of a construct with extreme levels.

Relatedly, word count approaches rest on the assumption that a greater use of a set of words signals a more extreme degree of a construct. For example, when measuring the positivity of a person, if a person used the word “okay” three times, that person would be considered more positive than someone who used the word “amazing” twice. However, using the word “okay” three times does not necessarily mean the person is more extremely positive. Thus, it can be problematic to assume that a greater frequency of a set of words indicates that a person is more extreme on that dimension.

When to use. Word count approaches are often used when the word list already exists and thus can easily be applied to one’s own data. They are also very useful when researchers anticipate analyzing a large set of data either in their current study or analyzing text across a number of future studies. As such, word count approaches are particularly advantageous for automatizing the analysis process.

Moreover, if the word list was originally generated and refined with a data-driven approach, it also has the potential to enhance measurement accuracy. For instance, as mentioned previously, Schwartz and colleagues (2013) found that emotionally-stable individuals were more likely to reference a more active lifestyle, which served as a novel indicator of that personality trait. Given the novelty of this indicator, human coders may have missed this signal and provided worse estimates of emotional stability.

Imputation

Overview. Imputation represents an alternative dictionary-based approach to analyze text. The aim of imputation is to essentially replace each occurrence of a word with a numerical value. By assigning values to each word, the imputation approach moves beyond treating language as a binary indicator via the presence or absence of a word. Instead, an imputation approach proposes that language, and the individual words that it consists of, can provide a deeper and more nuanced understanding of the *level* of a construct. For example, whereas “effective” and “amazing” both signal positivity and would contribute equally in a word count approach, the imputation approach would assign positivity values to each of these words and thereby allows them to signal different levels of positivity.

As a concrete example of an imputation tool, consider the Evaluative Lexicon (EL; Rocklage & Fazio, 2015; Rocklage et al., 2018b). The EL uses an imputation approach to measure the emotionality, extremity, and valence of individuals’ evaluations via natural language. To develop the EL, the researchers accumulated a large list of words that provided a signal of individuals’ opinions (e.g., “amazing,” “pleasant,” “mediocre,” “abhorrent”). They then asked a large set of external participants to judge the extent to which each word implied a negative to positive evaluation (0: very negative; 9: very positive) and a separate set of participants to judge the extent to which each word implied an evaluation based on emotion (0: not at all emotional; 9: very emotional). As a result, the normative valence, extremity, and emotionality of a word can be represented. For example, the word “effective” scores a 7.10 out

9.00 on positivity and a 3.10 out of 9.00 on emotionality, whereas “amazing” scores a 7.97 on positivity and 6.60 on emotionality. These normative values can then be put in place of that word – i.e., imputed – each time that word is used. They can then be averaged together, for example, for each piece of text to form an index of each facet for that text.

Rocklage, Rucker, and Nordgren (2018a) demonstrated the utility of the imputation approach and the EL to examine how the intent to persuade a person changes individuals' language. Specifically, when individuals were instructed or incentivized to persuade people via a review, individuals used words associated with greater positivity and emotionality. Of note, Rocklage et al. tended to observe these differences above-and-beyond the word count of positive words. As such, this work represents an example were imputation provides additional precision around the measurement of a construct (i.e., positivity) not available via pure word counts. For those interested, the EL software for analyzing natural text is freely available at www.evaluativelexicon.com.

Imputation has also been used successfully to measure other constructs such as how concrete versus abstract a word is (Brysbaert, Warriner, & Kuperman, 2013; Snefjella & Kuperman, 2015), how much warmth and competence is conveyed by a word (Holoien & Fiske, 2013, Studies 1a and 1b), and how rare a word is (Brysbaert & New, 2009).

Introductory how-to. As with word counts, the creation of one's own word list for imputation requires a series of steps. The steps are similar to word count approaches with a notable exception in that the words must be quantified in some fashion. Thus, for imputation methods, the general sequence is to 1) generate a list of candidate words that are likely to provide a signal of one's construct, 2) refine this list, 3) quantify the words themselves, and then 4) validate the tool. Because these procedures parallel the word count method, we only take note of the core difference, which is the quantification of the words themselves.

The quantification of the words involves obtaining a numerical value that should be associated with each word. Quantification can be accomplished through the use of either manual coding or derived from existing data. When measuring positivity, for instance, researchers could ask a large set of external judges to rate the extent to which each word implies a positive or negative evaluation (Rocklage et al., 2018b). Alternatively, researchers could use existing data to quantify the words. For instance, Brysbaert and New (2009) quantified how rare a word was by how often it occurred in a large sample of text (times used per million words). Of course, regardless of which method is used, the results must be validated to assess whether the words capture the actual construct of interest.

Pros and cons. Given that the imputation approach uses a dictionary-based approach, it shares some of the advantages of a word count approach. For instance, it can also be high in face validity and allows researchers to easily reproduce their own and others' results.

Perhaps the most salient benefit of an imputation approach, which is an advantage over a word count approach, is that imputation can capture a greater degree of information conveyed by language. As illustrated in the example of the EL, instead of treating each word as signaling equal levels of a construct, imputation allows words to signal different levels of the construct of interest. As such, imputation can offer a more accurate estimate of the construct of interest and, moreover, is less likely to conflate mild and extreme levels of that construct. Similarly, unlike a word count approach, the imputation approach does not rely on the assumption that a greater use of a set of words indicates more extreme levels of that construct.

Moreover, given that the imputation approach allows even a single word to provide a nuanced degree of information – the word "amazing" provides a specific level of valence and emotionality – a stronger signal can often be extracted from even short pieces of text. The implication of this is that whereas word count approaches can result in a skew in the analyzed text with relatively little variance, the imputation approach has the possibility of providing values that are more evenly distributed given that each word has its own value. As such, the imputation approach might be a more useful technique for analyzing short pieces of text. For example, on Twitter, most tweets might contain a single evaluative word, which would create a skew in the data. However, imputation methods might still detect differences in such tweets. For consumer researchers, imputation methods might be particularly useful for short texts that occur on social media.

Imputation approaches also share drawbacks with word count approaches. Specifically, researchers using imputation must consider how well their dictionary generalizes to different topics. For example, the word "horrifying" might be quantified as highly negative in the context of a restaurant review, but highly positive in the context of horror movies. Similarly, researchers must consider how negations affect their results.

An additional drawback of the imputation method is that it is generally more intensive to develop and use. In terms of development, researchers need to both create a list of words and obtain the values underlying the words. Moreover, researchers must validate both the words used as well as their underlying values. In terms of use, fewer ready-made tools exist to facilitate imputation, though Rocklage and colleagues' EL software (www.evaluativelexicon.com) can be modified to allow researchers to supply their own words and associated values.

A final noteworthy limitation of imputation is that it can result in more missing data compared to word counts. Whereas a word count approach assigns a 0 to all text that does not contain one of its words, imputation approaches often code such pieces of text as unable to be included in analyses. Take for instance, the scenario where a person makes the statement, “John had cereal.” Whereas word count approaches would assign this piece of text a 0 on positivity (the text would score low on positivity), an imputation approach would make no judgment on this sentence. Indeed, for imputation methods, because no valence is expressed it is unclear whether this statement should be analyzed for its valence at all. From a different perspective, the imputation approach emphasizes accuracy in measurement above absolute coverage of cases. Researchers should consider this coverage/accuracy trade-off when analyzing their data.

When to use. Similar to word counts, imputation is useful when datasets are large or when conducting a number of projects based around a given construct. In addition, relative to word counts, imputation methods can be of particular value for added precision and accuracy in the assessment of a construct.

Bottom-Up Data-Driven Approaches

The approaches we have overviewed thus far often provide a very principled, construct-oriented method to analyzing one’s text. Put simply, researchers start with a construct of interest (e.g., positivity) and consider how to extract such information from the text (i.e., manual coding, word counts, or imputation). For consumer behavior researchers, these approaches resonate with the field’s emphasis on the advancement of theory (Janiszewski, Labroo, & Rucker, 2016) as it relates to constructs of importance. However, cases also exist where researchers may be interested in deriving insights from text in a fully bottom-up manner and without preconceptions. This might be the case if researchers seek to understand the most prominent facets of a given product category based on online product reviews (Tirunillai & Tellis, 2014) or desire to automatically extract common themes across a wide range of text (Wang et al., 2015).

In these cases, a number of bottom-up, data-driven approaches are available, which have largely grown out of linguistics and computer science. We overview two prominent approaches: topic modeling and word representations. Topic modeling is used to extract common topics and themes from one's text, whereas word representations are often used to quantify how related words are to one another. Both approaches are based on the assumption that words that occur in similar contexts are related. That is, "you shall know a word by the company it keeps" (Firth, 1957, p. 11). This assumption is called the distributional hypothesis in linguistics – words that occur in similar contexts have similar meanings (e.g., Harris, 1954) – and it underlies much of the current advances in bottom-up text analysis.

Topic Modeling

Overview. Topic modeling provides a method for researchers who wish to derive topics from their text in a bottom-up, data-driven fashion. Topic modeling can be used, for instance, when a researcher desires to extract common themes across pieces of text. In essence, topic modeling techniques are similar to a factor analysis of one's texts as they distill a large set of text data into more concise categories. Moreover, some topic modeling approaches provide researchers the extent to which any piece of text contains each topic. So, a topic model might indicate that a document contains 50% Topic A, 30% Topic B, etc.

Although more detailed explanations are available elsewhere (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Landauer, Foltz, & Laham, 1998; Blei, Ng, & Jordan, 2003), at a general level, topic modeling relies on the idea that words that co-occur within a document as well as co-occur across documents (e.g., "durable," "sturdy," and "strong") are likely to be representative of some latent topic (e.g., product durability). Different topic modeling approaches use different methods to assess co-occurrence and enhance the coherence of each topic that is extracted. Currently, latent Dirichlet allocation (LDA; Blei et al., 2003) is the most popular topic modeling approach and will be the primary focus of this section.

The technique LDA uses to extract and refine topics is rather complex. The basic framework, however, is that LDA begins by semi-randomly assigning each word in a piece of text to a topic and then iteratively adjusting each topic (e.g., giving the word a stronger or weaker weight within the topic) in order to maximize the topic's internal consistency based largely on the co-occurrences of the words both within and across the documents that are analyzed.

As an example of LDA in use, Berger and Packard (2018) were interested in the role of distinctiveness for why some things catch on and become popular. To explore this idea, they analyzed the lyrics of approximately 2,000 songs over a three-year period and extracted 10 primary topics that represented those songs. For example, a topic they labeled as "uncertain love" was represented by words such as "ain't," "can't," and "love" whereas a topic they labeled "spiritual" was represented by words such as "believe," "grace," and "lord." Given that LDA provides the extent to which each document contains each topic (e.g., 50% of the lyrics for Song 1 are represented by Topic A), they quantified the extent to which a given song deviated from the normal topics for that genre. They found evidence that the more the song's lyrics deviated from its genre – that is, the more distinct they were – the more popular it was in terms of online downloads.

As another example, researchers interested in the discussion of health-related topics in online restaurant reviews used LDA along with pre-specified seed words (e.g., "calorie," "fat," "health") to extract a health-related topic from these reviews (Puranam, Narayan, & Kadiyali, 2017). They measured each restaurant review for the percent of this health topic present, tracked the degree to which the topic was present across time, and examined the extent to which an external event – a new health law – predicted a change in how prevalent this topic was. They found an uptick in the mention of their health-related topic across reviews after the government instituted the mandatory posting of caloric information on menus.

Introductory how-to. The common steps for LDA are to 1) preprocess the text, 2) specify the number of topics to extract, and 3) use LDA to extract the topics. Across these steps, however, there is a great deal of flexibility and LDA is considered to have an "art" to it in order to optimize results. There are also no strong rules regarding minimum sample sizes and researchers are often encouraged to empirically assess the appropriateness of LDA for their data via trial and error (see Tang, Meng, Nguyen, Mei, & Zhang, 2014 for a discussion).

In terms of preprocessing, it is common to remove "stop words" (e.g., "a," "an," "the") given that they provide little content information, to remove punctuation, and to change all words to lowercase so that upper- versus lowercase instances of a word become unimportant. Beyond such steps, however, consensus has yet to emerge for when to take additional steps for preprocessing (see Schofield & Mimno, 2016; Schofield, Magnusson, & Mimno, 2017).

In terms of specifying the number of topics (*k*) to extract, researchers often base this number on their intuition, which can then be updated once the model is run and researchers attempt to interpret the extracted topics (but see Teh et al. (2006) for an automated approach). From these results, researchers may then decide to extract greater or fewer topics depending on how interpretable the topics are, for instance.

One of the most popular ways to analyze one's data using LDA is via the MALLET program (McCallum, 2002), which is available to researchers for free as standalone software (http://mallet.cs.umass.edu) as well as via Python (using the 'gensim' toolkit) and R (using the 'mallet' package). See Graham and colleagues' (2012) tutorial for getting started with MALLET.

Pros and cons. The strengths of topic modeling are derived from the fact that it allows for a data-driven approach to understand common topics and themes in one's texts. It can potentially reveal novel insights into, for example, common aspects of a product category that people tend to focus on (Tirunillai & Tellis, 2014), which can then provide researchers with ideas for follow-up studies. As such, topic modeling can be quite powerful for exploratory analyses. However, topic modeling need not be limited to exploratory analyses. For example, in the work of Berger and Packard (2018), the construct of interest was not at the level of the topics themselves, but rather the *distinctiveness* of the topics in comparison to what was common for the genre. Thus, topic modeling was a way to operationalize commonality and to then measure deviation from this commonality.

One drawback to the topic modeling approach is that it can require a great deal of expertise, both in terms of technical computing skills as well as in the execution, interpretation, and refinement of the topic models. For example, in terms of refinement, researchers must often specify how many topics they want a topic model to extract. Given the bottom-up nature of this tool, this number is often based on trial-and-error (e.g., the number of topics where the topics are dissimilar enough to one another, but that are all also interpretable).

Another limitation of topic modeling is that it provides a list of words that are most indicative of a given topic, but the topic itself is not labeled. Researchers must still provide a label to each topic and thus different researchers may have fairly diverse interpretations of the same topic and therefore the meaning of the documents themselves. For example, whereas a topic derived from online restaurant reviews that contains the words "burrito," "rice," "beans," "salsa," and "chicken" is likely to represent Mexican restaurants, a topic that contains the words "dinner," "appetizers," "good," "main," and "drinks" appears more diffuse and is therefore more open to interpretation (Puranam et al., 2017).

When to use. Topic modeling is a powerful approach to derive insights into the content of one's text. As such, it is often used in an exploratory fashion. As mentioned previously, however, it can also be appropriated to test relevant hypotheses if conceptualized thoughtfully.

Word Representations

Overview. Another bottom-up, data-driven approach to text analysis is called word representation, whereby a word is quantified and represented as a "coordinate," or vector, in multidimensional space. Word representations are often used to gain insight into how related words are as based on their use in natural text. For example, the words "sturdy" and "durable" are likely to be closer in multidimensional space than the words "sturdy" and "efficient."

To calculate this relatedness, this approach uses an initial set of text and places each word from that text in a large vector space. The vectors themselves are largely derived from how often words co-occur together both within and across documents or throughout the text within some smaller window of text (e.g., how often the word "bread" occurs within five words of the word "butter"). Different calculations can then be made using these vectors (e.g., cosine similarity) to assess how related words are to one another.

Given that the vector space is constructed based on text written by humans, researchers have argued that the construction of word representations and their subsequent relations reflect a plausible model of the acquisition and representation of human knowledge (Landauer & Dumais, 1997). As such, word representations have been used to model the associations individuals hold and to test theories of cognition (Landauer & Dumais, 1997; Landauer et al., 1998).

Different methods exist to calculate word vectors and these different methods are often categorized as either count-based or prediction-based approaches (Levy, Goldberg, & Dagan, 2015). In terms of count-based approaches, the most popular is latent semantic analysis (LSA; Deerwester et al., 1990), which begins with the creation of a matrix of how often each word in a piece of text occurs both within that piece of text as well as across one's different documents of interest. These co-occurrences can then be reduced to a smaller set of dimensions ("topics") with a technique called singular value decomposition (SVD), which is similar to a principal components or factor analysis.

Prediction-based approaches are relatively new and are represented by such techniques as word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and, though it is sometimes categorized as its own approach, Global Vectors (GloVe; Pennington, Socher, & Manning, 2014). Word2vec, the most popular prediction-based word representation technique, is considered a prediction-based model as it is specifically trained to maximize its ability to predict nearby words. To do so, word2vec creates word vectors to represent each word based in part on how often that word (e.g., "bread") tends to appear within, for example, five words of another word (e.g., "butter"). Thus, word2vec operates by optimizing the ability to predict nearby words in a piece of text ("butter") based on a given word ("bread"). Research is ongoing with respect to how word representation models differ both conceptually and in performance (Levy & Goldberg, 2014b; Levy et al., 2015).

As an example of the word representation approach, Bhatia (2017) used LSA to create word vectors to measure the extent to which stereotypically African American versus white American names (e.g., Darnell vs. Brandon) co-occurred with pleasant (e.g., "cheer," "honor," "laughter") versus unpleasant words (e.g., "jail," "grief," "ugly") using a large sample of newspaper articles. In line with previous research in psychology (Greenwald, McGhee, & Schwartz, 1998), African American names tended to show greater relatedness to unpleasant words whereas white American names tended to show greater relatedness to pleasant words. As argued by Bhatia, these distances may reflect broader associations that lay individuals hold.

A second example comes from research on valence which tested whether positive information is more similar to other positive information compared to how similar negative information is to other negative information. In other words, is "good" more alike than "bad"? Using LSA, Koch and colleagues (2016) found that, indeed, 20 common positive topics (e.g., "flowers" and "chocolate") were more closely related than 20 common negative topics (e.g., "disease" and "crime") as quantified by their distance in a data-derived vector space.

Introductory how-to. To begin, researchers can utilize existing word vector databases and thus need not necessarily create their own. For example, the University of Colorado at Boulder (http://lsa.colorado.edu) provides word representations that have been derived via LSA on different texts (e.g., a large collection of textbooks). Similarly, word2vec has been applied to 300 billion words from the Google News database ("word2vec," n.d.). To utilize the existing word2vec database, researchers can use the gensim toolkit for Python. A tutorial for using word2vec is available on the gensim homepage ("gensim," n.d.). In sum, a very large number of words has already been placed in a vector space and calculations can be made from these existing datasets.

Researchers can also create their own word representations. This might be desirable if, for instance, researchers want to compare how related the word "durable" is with a brand name for two different sets of individuals in online reviews (e.g., experts versus novices). Is "durable" more closely related to the brand name for one set of individuals versus another? Similarly, training one’s own word vectors may also be important if one is interested in changes in relatedness across time. For an example of similar approaches in political science, see Li, Schloss, and Follmer (2017).

We provide an example of using word2vec to create word representations given its current popularity. Word2vec models can be created in the gensim toolkit. Although researchers can use the default settings provided by gensim, they can also modify these parameters. At present, however, consensus does not exist regarding the optimal parameters to use and in which situations. Instead, researchers experiment with different parameters and examine how sensible the results are (e.g., whether words that are intuitively related are also closely related via word2vec; see Levy & Goldberg, 2014a for a discussion). When creating word representations, it is considered desirable to have around 30,000 unique words, but smaller samples have also been used (particularly when using LSA; Altszyler, Sigman, Ribeiro, & Slezak, 2016). See the gensim homepage for more information on creating one’s own word vectors (“gensim,” n.d.).

Once the word vectors have either been obtained from others or trained by oneself, the most common approach is to calculate how related any given word is to another by using a cosine similarity measure (available in gensim). The resulting values can then be used in further analyses by, for example, correlating them with other measures of interest. Word vectors could also be used to measure the relatedness of the language between two individuals conversing about a given product, service, or brand (Duran, Paxton, & Fusaroli, 2018; Ta, Babcock, & Ickes, 2017).

Pros and cons. As with topic modeling, the power of the word representation approach is that it can provide a data-driven quantification of how related words are to one another. These distances can then be used to examine hypotheses of interest. For consumer behavior research, for example, it would be possible to assess the extent to which different brands are related to positive versus negative words or even with particular dimensions of a product (e.g., durability).

Another strength of this approach is that it can be used with relatively little expertise by utilizing existing word vectors. For example, as detailed previously, word2vec has already been applied to 300 billion words from the Google News database ("word2vec," n.d.). Thus, a very large number of words have already been placed in a vector space and calculations can be made with this existing dataset.

As with topic modeling, however, word representations are derived in an atheoretical manner and thus may be problematic for disciplines that emphasize theory-driven approaches. More specifically, given its data-driven nature, it is not entirely clear exactly what word vectors represent. When words are close to one another in vector space, what exactly does this mean? Although they are related in some way, it is not always clear how. For instance, the words "good" and "great" are relatively close in vector space using the Google News database (.73 cosine similarity), but so are the words "good" and "bad" (.72). Thus, it appears that word vectors and their cosine similarity represent an abstract understanding of some kind of *relatedness* ("good," "great," and "bad" are all evaluative), but not necessarily *similarity* (Hill, Reichart, & Korhonen, 2015). Efforts to improve similarity estimates are ongoing within computer science (e.g., Mrkšić et al., 2016).

When to use. As referenced previously, researchers have argued that word vectors and their associations are representations of human knowledge (Landauer & Dumais, 1997). Thus, a primary reason to use the word representation approach is to assess associations between words in a data-driven manner. For consumer behavior researchers, such associations can be used to test hypotheses of interest.

Another interesting use of word vectors would be to examine how “aligned” or “in sync” two individuals are in their language. When two consumers are talking about a product, for instance, the relatedness of their language can be calculated and used to predict different outcomes specific either to that dyad or to perceptions of the product itself (see Duran et al., 2018 for more on this general approach).

Summary

The explosion of available text has opened up a number of opportunities for researchers. To help researchers make use of these data, the current chapter provides an overview and discussion of important text analysis approaches available. These approaches range from the relatively straightforward manual coding of text to more automated methods that rely on validated dictionaries to fully data-driven approaches that provide insights from language in a bottom-up fashion. As we have noted, each of these approaches has its advantages and disadvantages, but, taken together, they provide researchers a large toolkit to analyze language with methods not previously possible. It is our hope that the current chapter both informs consumer behavior researchers regarding the cutting-edge text analysis methods available and encourages them to pursue these analyses for their own work.

References

Altszyler, E., Sigman, M., Ribeiro, S., & Slezak, D. F. (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. *ArXiv:1610.01520 [Cs]*. Retrieved from http://arxiv.org/abs/1610.01520

Barasch, A., & Berger, J. (2014). Broadcasting and Narrowcasting: How Audience Size Affects What People Share. *Journal of Marketing Research*, *51*(3), 286–299. https://doi.org/10.1509/jmr.13.0238

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205. https://doi.org/10.1509/jmr.10.0353

Berger, J., & Packard, G. (2018). Are Atypical Things More Popular? *Psychological Science*. https://doi.org/10.1177/0956797618759465

Bhatia, S. (2017). The semantic representation of prejudice and stereotypes. *Cognition*, *164*, 46–60. https://doi.org/10.1016/j.cognition.2017.03.016

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

Boyd, R. L. (2016). *RIOT Scan: Recursive Inspection of Text Scanner*. Retrieved from http://riot.ryanb.cc

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, *18*, 63–68. https://doi.org/10.1016/j.cobeha.2017.07.017

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4), 284–290.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, *41*(6), 391–407.

Duran, N., Paxton, A., & Fusaroli, R. (2018). ALIGN: Analyzing Linguistic Interactions with Generalizable techNiques - a Python Library. *PsyArXiv*. https://doi.org/10.17605/OSF.IO/A5YH9

Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930-1955. In *Studies in Linguistic Analysis* (pp. 1–32). Oxford: Blackwell.

Fujita, K., Henderson, M. D., Eng, J., Trope, Y., & Liberman, N. (2006). Spatial distance and mental construal of social events. *Psychological Science*, *17*(4), 278–282. https://doi.org/10.1111/j.1467-9280.2006.01698.x

gensim: topic modelling for humans. (n.d.). Retrieved February 20, 2018, from https://radimrehurek.com/gensim/models/word2vec.html

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science*, *31*(3), 493–520. https://doi.org/10.1287/mksc.1110.0700

Graham, S., Weingart, S., & Milligan, I. (2012, September 2). Getting Started with Topic Modeling and MALLET. Retrieved February 19, 2018, from https://programminghistorian.org/lessons/topic-modeling-and-mallet

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*, 146–162.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, *41*(4), 665–695. https://doi.org/10.1162/COLI_a_00237

Holoien, D. S., & Fiske, S. T. (2013). Downplaying positive impressions: Compensation between warmth and competence in impression management. *Journal of Experimental Social Psychology*, *49*(1), 33–41. https://doi.org/10.1016/j.jesp.2012.09.001

Humphreys, A., & Wang, R. J. H. (2018). Automated Text Analysis for Consumer Research. *Journal of Consumer Research*, *44*(6), 1274–1306. https://doi.org/10.1093/jcr/ucx104

Janiszewski, C., Labroo, A. A., & Rucker, D. D. (2016). A Tutorial in Consumer Research: Knowledge Creation and Knowledge Appreciation in Deductive-Conceptual Consumer

Research. *Journal of Consumer Research*, *43*(2), 200–209. https://doi.org/10.1093/jcr/ucw023

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun Use Reflects Standings in Social Hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125–143. https://doi.org/10.1177/0261927X13502654

Koch, A., Alves, H., Krüger, T., & Unkelbach, C. (2016). A general valence asymmetry in similarity: Good is more alike than bad. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1171–1192. https://doi.org/10.1037/xlm0000243

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krikorian, R. (2013). New Tweets per second record, and how! Retrieved April 3, 2018, from https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how.html

Krippendorff, K. (2012). *Content Analysis: An Introduction to Its Methodology* (3rd ed.). Thousand Oaks, CA: SAGE.

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Levy, O., & Goldberg, Y. (2014a). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 2).

Levy, O., & Goldberg, Y. (2014b). Neural Word Embedding As Implicit Matrix Factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (pp. 2177–2185). Cambridge, MA, USA: MIT Press. Retrieved from http://dl.acm.org/citation.cfm?id=2969033.2969070

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, *3*(0), 211–225.

Li, P., Schloss, B., & Follmer, D. J. (2017). Speaking two “Languages” in America: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. *Behavior Research Methods*, *49*(5), 1668–1685. https://doi.org/10.3758/s13428-017-0931-5

McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. Retrieved from http://mallet.cs.umass.edu

McGraw, K. O., & Wong, S. P. (1996). Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods*, *1*(1), 30–46.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (pp. 3111–3119). USA: Curran Associates Inc. Retrieved from http://dl.acm.org/citation.cfm?id=2999792.2999959

Moore, S. G. (2015). Attitude predictability and helpfulness in online reviews: The role of explained actions and reactions. *Journal of Consumer Research*, *42*(1), 30–44.

Mrkšić, N., Séaghdha, D. Ó., Thomson, B., Gašić, M., Rojas-Barahona, L., Su, P.-H., … Young, S. (2016). Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL* (pp. 142–148). Retrieved from http://arxiv.org/abs/1603.00892

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Retrieved from https://repositories.lib.utexas.edu/handle/2152/31333

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the Empirical Methods in Natural Language Processing* (pp. 1532–1543). Doha, Qatar.

Puranam, D., Narayan, V., & Kadiyali, V. (2017). The Effect of Calorie Posting Regulation on Consumer Opinion: A Flexible Latent Dirichlet Allocation Model with Informative Priors. *Marketing Science*, *36*(5), 726–746. https://doi.org/10.1287/mksc.2017.1048

Rocklage, M. D., & Fazio, R. H. (2015). The Evaluative Lexicon: Adjective use as a means of assessing and distinguishing attitude valence, extremity, and emotionality. *Journal of Experimental Social Psychology*, *56*, 214–227. https://doi.org/10.1016/j.jesp.2014.10.005

Rocklage, M. D., Rucker, D. D., & Nordgren, L. F. (2018a). Persuasion, Emotion, and Language: The Intent to Persuade Transforms Language via Emotionality. *Psychological Science*, *29*(5), 749–760. https://doi.org/10.1177/0956797617744797

Rocklage, M. D., Rucker, D. D., & Nordgren, L. F. (2018b). The Evaluative Lexicon 2.0: The measurement of emotionality, extremity, and valence in language. *Behavior Research Methods*, *50*(4), 1327–1344. https://doi.org/10.3758/s13428-017-0975-6

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling Out the Stops: Rethinking Stopword Removal for Topic Models. In *Proceedings of the 15th Conference of the*

European Chapter of the Association for Computational Linguistics (Vol. 2, pp. 432–436). Valencia, Spain.

Schofield, A., & Mimno, D. (2016). Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, *4*, 287–300.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Semin, G. R., & Fiedler, K. (1988). The Cognitive Functions of Linguistic Categories in Describing Persons: Social Cognition and Language. *Journal of Personality and Social Psychology*, *54*(4), 558–568.

Snefjella, B., & Kuperman, V. (2015). Concreteness and Psychological Distance in Natural Language Use. *Psychological Science*, 0956797615591771. https://doi.org/10.1177/0956797615591771

Spiller, S. A., & Belogolova, L. (2017). On Consumer Beliefs about Quality and Taste. *Journal of Consumer Research*, *43*(6), 970–991. https://doi.org/10.1093/jcr/ucw065

Stanley, D. J., & Spence, J. R. (2014). Expectations for Replications: Are Yours Realistic? *Perspectives on Psychological Science*, *9*(3), 305–318. https://doi.org/10.1177/1745691614528518

Ta, V. P., Babcock, M. J., & Ickes, W. (2017). Developing Latent Semantic Similarity in Initial, Unstructured Interactions: The Words May Be All You Need. *Journal of Language and Social Psychology*, *36*(2), 143–166. https://doi.org/10.1177/0261927X16638386

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32* (pp. I–190–I–198). Beijing, China: JMLR.org. Retrieved from http://dl.acm.org/citation.cfm?id=3044805.3044828

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. https://doi.org/10.1198/016214506000000302

Tirunillai, S., & Tellis, G. J. (2014). Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*, *51*(4), 463–479. https://doi.org/10.1509/jmr.12.0106

Tosti-Kharas, J., & Conley, C. (2016). Coding Psychological Constructs in Text Using Mechanical Turk: A Reliable, Accurate, and Efficient Alternative. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00741

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review*, *110*(3), 403–421.

Wang, X. (Shane), Bendle, N. T., Mai, F., & Cotte, J. (2015). The Journal of Consumer Research at 40: A Historical Analysis. *Journal of Consumer Research*, *42*(1), 5–18. https://doi.org/10.1093/jcr/ucv009

word2vec. (n.d.). Retrieved January 28, 2018, from https://code.google.com/archive/p/word2vec/

Table 1

Text Analysis Techniques with a Short Summary of Each Technique, its Pros and Cons, and Representative Examples

Method	Summary	Pros	Cons	Examples
Manual coding	Uses two or more human coders to rate text on a predefined construct	Quick; cheap; requires little technical expertise; flexible	Difficult for use with large datasets or over a series of projects; human coders may have difficulty judging construct	Barasch & Berger, 2014; Moore, 2015
Top-down dictionary approaches				
Word counts	Uses a predefined dictionary to count the presence or absence, number of instances, or frequency of words to measure construct of interest	Easy to automate and implement; high face validity; reproducible; prevalent in literature	Cannot rely on face validity alone; can result in skewed data which presents interpretation and statistical obstacles; less nuanced; assumes that greater frequency represents more extreme levels of a construct	Spiller & Belogolova, 2017; Pennebaker et al., 2015
Imputation	Uses a predefined dictionary along with underlying values for each word to measure construct of interest	Similar advantages to word counts; more nuanced measurement; works well with shorter pieces of text	More intensive to create; can result in missing data (emphasizes accuracy over coverage)	Rocklage, Rucker, & Nordgren, 2018a; Brysbaert, Warriner, & Kuperman, 2013
Bottom-up data-driven approaches				

Topic modeling	Analyzes text to extract common themes/topics	Extracts topics without preconceptions which can reveal novel insights	Requires technical expertise; topics must be labeled by researcher and thus open to interpretation	Berger & Packard, 2018; Puranam, Narayan, & Kadiyali, 2017
Word representations	Quantifies words to provide estimates of how related words are to one another	Provides data-driven estimates of how related words are to one another; existing databases decrease need for technical expertise	Ambiguous exactly what “relatedness” represents	Bhatia, 2017; Koch, Alves, Krüger, & Unkelbach, 2016