Marketing Letters (2022) 33:365–377
https://doi.org/10.1007/s11002-022-09635-6

Marketing insights from text analysis

Jonah Berger[1] · Grant Packard[2] · Reihane Boghrati[3] · Ming Hsu[4] · Ashlee Humphreys[5] · Andrea Luangrath[6] · Sarah Moore[7] · Gideon Nave[1] · Christopher Olivola[8] · Matthew Rocklage[9]

Accepted: 19 May 2022 / Published online: 10 June 2022

Abstract

Language is an integral part of marketing. Consumers share word of mouth, salespeople pitch services, and advertisements try to persuade. Further, small differences in wording can have a big impact. But while it is clear that language is both frequent and important, how can we extract insight from this new form of data? This paper provides an introduction to the main approaches to automated textual analysis and how researchers can use them to extract marketing insight. We provide a brief summary of dictionaries, topic modeling, and embeddings, some examples of how each approach can be used, and some advantages and limitations inherent to each method. Further, we outline how these approaches can be used both in empirical analysis of field data as well as experiments. Finally, an appendix provides links to relevant tools and readings to help interested readers learn more. By introducing more researchers to these valuable and accessible tools, we hope to encourage their adoption in a wide variety of areas of research.

Keywords Natural language processing · Automated textual analysis · Language

✉ Jonah Berger
jberger@wharton.upenn.edu

[1] Wharton School at the University of Pennsylvania, Philadelphia, PA, USA

[2] Schulich School of Business, York University, Toronto, Canada

[3] Wharton Risk Center and Marketing Department, University of Pennsylvania, Philadelphia, PA, USA

[4] University of California, Berkeley, USA

[5] Medill School of Journalism, Media, and Integrated Marketing Communications, Northwestern University, Evanston, IL, USA

[6] Tippie College of Business, University of Iowa, Iowa City, IA, USA

[7] Alberta School of Business, University of Alberta, Edmonton, AB, Canada

[8] Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA

[9] University of Massachusetts, Boston, MA, USA

Language is an integral part of marketing. Consumers share word of mouth, salespeople pitch services, and advertisements try to convince consumers to buy. Retail employees answer questions, customer service agents try to solve problems, and movies, books, and other cultural products use language to entertain and inform. Even consumers' private thoughts are expressed using language.

Further, small differences in wording can have a big impact. The exact words used in word of mouth can shape its influence (Packard & Berger, 2017; Berger, Rocklage, and Packard 2022; Moore, 2012), the language service agents use shape customer satisfaction (Packard et al., 2018), and the words used in books, movies, and other cultural products shape their success (Berger et al., 2021).

But while it is clear that language is both frequent and important, how can we extract insight from this increasingly available form of data?

The digitization of content has created a wealth of textual information. Online reviews capture what consumers talk about and why, and social media posts shed light on brand perceptions. Customer service calls can be transcribed to understand what drives customer satisfaction, and experimental participants provide thought protocols that can be parsed for deeper insight into the mechanisms driving behavior.

But parsing this data requires the right tools: objective, scalable methods that turn text into data.

Building on recent work (e.g., Berger et al., 2020; Humphreys & Wang, 2018; Shankar & Parsana, 2022), this paper offers an accessible, hands-on introduction to three main approaches to automated textual analysis (i.e., dictionaries, topic modeling, and embeddings). We suggest these approaches can be thought of as tools to help understand the *what*, *how*, and *why* of consumer and marketing language. For those interested in what is being talked about, or the topic or themes discussed, topic modeling and embedding type approaches can be particularly useful. For those interested in how something is being talked about, or what motivations might be reflected, dictionary-based approaches can be particularly helpful.

We provide a brief summary of each approach, some examples of how it has been used, and some advantages and limitations. Further, we outline how these approaches can be used both in empirical analysis of field data as well as experiments. Finally, an appendix provides links to relevant tools and readings to help readers dive deeper.

While a detailed discussion of all the methods and uses of textual analysis is beyond the scope of this paper, we hope it provides useful pointers to places where readers can learn more.

1 Dictionaries

Some of the most user-friendly methods for text analysis are top-down, dictionary-based approaches. These approaches rely on a pre-existing list—i.e., a dictionary—of words, phrases, or symbols that are counted in a piece of text. For example, if researchers want to measure how certain consumers are, they might search their text using a dictionary that contains words such as "I'm convinced," "don't know," and "absolutely" to represent the construct (Rocklage et al., 2022). If researchers are interested in measuring how self-focused consumers are, they might use a dictionary that contains words like "I," "me," and "mine" (Spiller & Belogolova, 2017). Each of these words is searched for in the target text and then summed. Texts with greater use of "me," for example, would have higher "self-focused" scores because more matches would be found from the dictionary.

This method is particularly useful for getting started with automated text analysis because dictionary software is generally easy-to-use and free, and there are many standardized dictionaries to choose from (Humphreys & Wang, 2018). Researchers can measure constructs using the Linguistic Inquiry and Word Count software (Boyd et al., 2022), sentiment/attitudes using the Evaluative Lexicon (Rocklage et al., 2018a, 2018b), and nonverbal cues using the textual paralanguage classifier (Luangrath, Xu, and Wang 2022), for just three examples (see Web Appendix for more). Each of these uses a slightly different approach to quantify language, but all rely on a dictionary to search for words of interest.

1.1 Linguistic inquiry and word count

One widely used set of dictionaries is Linguistic Inquiry and Word Count (LIWC; Boyd et al., 2022). LIWC includes a range of wordlists, many of which were developed based on psychological scales. For example, LIWC includes a wordlist for measuring positive and negative emotion based on the PANAS scale (Watson et al., 1988). LIWC includes 20 measures of linguistic features (e.g., verb tense), 60 psychological categories (e.g., emotion, cognition), and 19 substantive categories (e.g., leisure) in addition to measures of punctuation (Boyd et al., 2022). Higher-level categories can be used to summarize other subgroups in the software. For example, "clout" is a combined measure of second-person pronouns ("we"), negations, and swear words (Jordan & Pennebaker, 2015). The most recent version adds tools to build word clouds, identify language style matching, and find narrative structure. To normalize for text length (e.g., words in an online post), the LIWC software produces data in the form of percent of total words.

LIWC's dictionaries have been validated on a range of materials such as academic abstracts, English literature texts, and other spoken and written material (King & Pearce, 2010; Tausczik & Pennebaker, 2010). LIWC has been used to assess social acceptance in news media (Humphreys, 2010), emotional contagion (Berger & Milkman, 2012), attentional and social focus in tweets (Barasch & Berger, 2014), and market logics (Ertimur & Coskuner-Balli, 2015). LIWC also allows researchers to create custom dictionaries to measure other constructs (Humphreys & Wang, 2018). And although scholars have found more precise ways to measure some constructs like sentiment (Hartmann et al., 2019), LIWC remains a good place to start (www.liwc.app).

1.2 The Evaluative Lexicon

Another example of a dictionary approach is the Evaluative Lexicon (EL; Rocklage & Fazio, 2015; Rocklage et al., 2018a, 2018b). The EL is a validated measure of the valence, extremity, and emotionality of individuals’ opinions in language. To construct the dictionary, researchers used billions of words, millions of online reviews, and the judgments of a large set of external raters. Based on this data-driven approach, the EL searches only for words that provide a reliable signal of individuals’ opinions in natural language. The final dictionary includes words such as “magnificent,” “problematic,” and “flavorful.”

Rather than simply counting whether a word is present or not in a piece of text, the EL gives a score to each word in its dictionary based on validated external ratings. For example, the word “flawless” has the score of 8.24 on valence (out of 9.00), 3.74 on extremity (out of 4.50), and 3.05 on emotionality (out of 9.00). On the other hand, “elated” signals a very different opinion—one that is equally positive, but based more on emotion (scores of 8.20, 3.70, and 7.11, respectively). The EL dictionary and its scores have been extensively validated and applied across social media posts, audio transcripts, consumer reviews, and a number of other contexts (Berger, Rocklage, and Packard 2022; Rocklage & Luttrell, 2021; Rocklage et al., 2021). It is available at www.EvaluativeLexicon.com.

1.3 Textual paralanguage classifier

The textual paralanguage classifier (PARA) identifies nonverbal communication cues in text (Luangrath, Xu, and Wang 2022). In contrast to other tools that rely predominantly on words themselves, PARA takes an alternative approach and focuses on nonverbal parts of speech. This tool detects 19 different auditory, tactile, and visual features of text (Luangrath et al., 2017). For example, vocal aspects of text speech convey stress with CAPS (e.g., GREAT), emphasis (e.g., !!!!), tempo (e.g., amazinggggggg, in this case denoted with “stretchable words”), vocalizations (e.g., ugh or ahh), body language–like emojis (e.g., 😘), or emoticon facial expressions (e.g.,:-D), among others. The detection of these linguistic markers influences perceptions of sentiment valence and intensity, and improves prediction accuracy of consumer engagement on social media (Luangrath, Xu, and Wang 2022). PARA software operates using a panel of five sub-dictionaries and rule-based algorithms. PARA is particularly helpful for text that is more informal such as social media data, customer service chats, email, blogs, comments, text created in apps, or any content generated via mobile device as these often contain textual paralanguage. PARA is less helpful when analyzing formal text (e.g., shareholder reports). PARA can be found at www.textualparalanguage.com.

1.4 When to use

Dictionary approaches are useful when text can be specified in relatively precise or finite ways that can be easily represented by word presence or absence. For that reason, they excel at measuring individual and cultural focus (i.e., what is being attended to) or emphasis on a single particular subject or construct (Humphreys & Wang, 2018; Tausczik & Pennebaker, 2010). Because words are specified a priori, dictionary methods also perform well when researchers have a firm idea of the operationalization of constructs in the text. And because there are many well-validated dictionaries available, the approach allows for concurrent and construct validity when working across studies.

When it is difficult to specify the operationalization of constructs, however, or when measuring the construct requires studying sentence structure or inter-relation of words within a sentence, other methods may be helpful. Similarly, while dictionaries can be used broadly across contexts, classifiers or other machine learning approaches designed for prediction may perform better in very specific contexts. The meaning of words like "we" and "our," for example, may be quite different in conversation than in academic papers.

2 Topic modeling

Beyond the individual words companies, consumers, or employees use, what broader topics or themes are they talking about? Do hotel reviews tend to talk about the room, the service, or the food? Should retail employees focus on customer needs or the products offered? And in an experimental context, does a manipulation impact what topics study participants focus on?

Topic models can answer these questions and more. Rather than focusing on top-down, pre-determined constructs, as is often the case with dictionaries, topic modeling is usually bottom-up, using words that co-occur within and across texts to determine the latent topics that appear. Based on this, the method outputs different topics and the words associated with them. This, in turn, can be used to identify how much of a given text is about each latent topic. For a travel review, for example, 51% of the review might be about a hotel room, 25% about the front desk, and 24% about the restaurant.

One common topic modeling approach is latent Dirichlet allocation (LDA; Blei et al., 2003), although a variety of options are available (Vayansky & Kumar, 2020). While some of these lean to more complex neural network approaches, basic topic modeling can be performed in a straightforward manner using R or Python, and less technical users can simply upload a text file at a website to generate LDA results (see www.textanalyzer.org).

Marketing researchers have used topic models in a variety of novel ways. Tirunillai and Tellis (2014) explore dimensions linked to quality, how they change over time, and how that relates to competitive brand positioning. Li and Ma (2020) show how marketers can use topic modeling of consumer search terms to identify where consumers are in the decision-making process. The approach was used to find spoilers in movie reviews, which surprisingly help, rather than hurt, ticket sales (Ryoo et al., 2021). Topic modeling can also be used to find and examine specific psychological constructs relevant to marketing. Zhang, Li, and Ng (2021) performed guided LDA by training it on an initial set of words associated with warmth and competence, and then scored thousands of brands appearing in Yelp reviews according to those perceptual dimensions. Chung and colleagues (2022) used the approach to uncover the motivations (e.g., intrinsic vs. financial) of people who rent their properties on Airbnb. Topic model results can also be useful as control variables, such as accounting for different topics that might arise in customer service conversations (Packard & Berger, 2021).

In addition to field data, topic modeling can also be used on the language produced in experiments. Researchers could analyze thought listings after a manipulation, for example, to see if thoughts differ across conditions in conceptually or substantively meaningful ways. This approach might be especially useful when self-report scales are not available, when participants have less insight into their own attitudes, or when response bias may lead to inaccurate responding.

2.1 When to use

Like any method, topic modeling has limitations. While fit statistics such as coherence or perplexity can help, interpretation of each topic's theme or meaning is ultimately up to the researcher, leaving considerable degrees of freedom if topic meaning is important. Independent judges can be used to score the topics to help in such cases. What's more, topic modeling does not account for the proximity of words within texts. Even if "river" and "bank" appear several sentences or paragraphs away from each other, topic modeling might think they are related. Embeddings can help complement this shortcoming, as can embedded topic modeling that combines aspects of both approaches (Dieng et al., 2020).

3 Embeddings

Word embedding models have emerged as a popular way to capture semantic information contained in text without labor-intensive manual labeling. These models rely on statistical algorithms to learn semantic representations from word co-occurrence patterns in natural language (e.g., Bullinaria & Levy, 2012; Landauer & Dumais, 1997; Lenci, 2018). They examine the appearance of a word across different contexts (i.e., surrounding words) and represent it as a dense numerical vector—often with tens or hundreds of dimensions—in a vector space. This allows for performing mathematical operations on text, such as calculating how different words, paragraphs, or entire texts are related (e.g., using measures like cosine similarity).

Importantly, the mapping of words to vector representations is based not only on co-occurrence and frequency, but also on context. Consequently, words used in similar ways have similar vector representations. "Dog" and "cat" (i.e., pets) may appear close together in vector space, for example, as might "banana" and "blueberry" (i.e., fruits), but "dog" and "blueberry" should appear farther apart. Moreover, such vector spaces can capture analogies between words. Subtracting the vector representation of "men" from that of "king," for example, yields a vector that is equivalent to the one obtained by subtracting "women" from "queen" (Mikolov et al., 2013). That is, given the analogy to solve: "king" is to "men" as "queen" is to "___", these vector spaces can correctly predict that "women" should be the answer.

Word embedding models thus quantify the semantic relations between different words, such that their degree of contextual overlap indicates their semantic relatedness (Boleda, 2020; Harris, 1970; Lenci, 2018). Importantly, it also extends to higher-order relationships beyond direct co-occurrence. For example, synonyms rarely co-occur, as usually only one is used in a given context, yet their closeness in meaning is reliably captured by word embedding models (Bullinaria & Levy, 2012). As such, word embedding models trained on large text corpora would have access not only to semantic relationships between, for example, product categories and brands (e.g., "fast food" and "McDonald's"), but also to their relationships with shared concepts (e.g., "cheap", "hamburger", and "drive-through").

Because of their relative novelty and data requirements, word embedding techniques have seen fewer applications within marketing, at least so far. Nonetheless, several recent papers demonstrate embeddings' potential to address a variety of important questions. Gabel et al. (2019) utilize a "product embeddings" technique (P2V-MAP) on market basket data from a grocery retailer to quantify latent, attribute-level similarities between products, and thereby map market structures (e.g., product complementarities vs. competitions). They find, for example, that wines form distinct clusters along price ranges, likely reflecting consumer loyalty to specific price tiers for wine (Jarvis & Goodman, 2005). Timoshenko and Hauser (2019) demonstrate how word embeddings can identify customer needs from product reviews, while offering important advantages over more conventional techniques (e.g., interviews and focus groups). Toubia, Berger, and Eliashberg (2021) use embeddings to quantify the speed, volume, and circuitousness of texts, demonstrating that these features help explain whether books, movies, academic papers, and other cultural products succeed or fail (also see Laurino Dos Santos & Berger, 2022). Bhatia and Olivola (2018, 2021) showed that word embeddings can be used to predict the subjective dimensions of brand perception (e.g., brand personality traits, Aaker, 1997) for hundreds of brands and evaluation dimensions. Then, they were able to quantify and map the associations between brands and a rich variety of concepts. Such semantic maps, in turn, can serve as a foundation to study many interesting questions. For example, Aka et al. (2020) relied on this approach to link the perceptions of brands to the personality traits of consumers who "like" them on Facebook and tested whether consumers prefer brands that "fit" their own psychological tendencies (see Nave et al., 2020, for a similar approach). Finally, Zhang et al. (2018) showed that word embedding models trained on large text corpora can be used to predict consumer brand recall, without having to rely on collecting additional survey data.

3.1 When to use

While embeddings are quite useful, they are not without limitations. Given the key assumption that related words tend to appear in similar contexts, word representations depend on the properties of the text corpus used to learn them. In some cases, researchers will want to utilize word embeddings trained on text corpora tailored to their research questions (e.g., using a time-stamped corpora of tweets to study the evolution of brand perceptions on social media). In practice, however, training and validating such models requires access to very large text corpora with millions or even billions of words. Consequently, off-the-shelf embedding representations, learned from large and rich text corpora (e.g., Google News, Twitter, and Wikipedia), are often used (e.g., https://code.google.com/archive/p/word2vec/).

Technical challenges also remain. One is imprecision due to semantic ambiguity. In a Twitter corpus, for example, the word "apple" will sometimes refer to the brand, and in other cases to the fruit, yielding imprecise embedding representations of "apple." Embedding representations can also differ depending on the type of documents used to learn them. A brand or product will likely be represented differently in a model trained using a corpus of financial reports, for example, versus one trained using a corpus of consumer reviews.

4 Using text in experiments

Most of the text analysis examples discussed so far used field data, but these tools can also be used in experiments. Indeed, papers on language-focused topics frequently employ mixed methods, incorporating text analysis in both field data and experiments (e.g., Packard et al., 2018). In experiments, text can be used as an independent variable, dependent variable, or mediator—and text analysis tools can assess text used in each of these ways.

Manipulating text as an independent variable is useful for researchers studying how senders are affected by producing certain language, or how receivers are affected by hearing certain language. To manipulate language production, researchers can give participants general instructions to follow (e.g., be persuasive; Rocklage et al., 2018a, 2018b). Alternately, participants can be asked to complete a controlled, pre-scripted text that varies across conditions. For example, some participants complete sentences with explanations, while others complete sentences without (e.g., Moore, 2012). To manipulate language that receivers are exposed to, researchers can construct texts that vary in specific ways and measure their impact on participants' attitudes and behaviors (Lafreniere, Moore, and Fisher 2022; Rocklage & Fazio, 2020). For example, participants could read researcher-created reviews for material versus experiential purchases to see which they rely upon more (Dai et al., 2020).

Examining text as a dependent variable is useful for exploring how language use or preferences vary under different conditions. For example, researchers have used text analysis tools to test how audience size (e.g., small vs. large), device type (e.g., mobile vs. desktop), or goals (e.g., persuasion) alter participant-generated text in terms of sentiment or emotionality (Barasch and Berger 2014; Melumad et al., 2019; Rocklage et al., 2018a, 2018b). Alternately, participants—as senders or receivers—may choose from researcher-created text that varies in controlled ways (e.g., Moore & McFerran, 2017; Schellekens et al., 2010). For example, senders might choose which of several sentences they would use when writing a review, while readers might choose which sentences would be more helpful when reading a review (Moore, 2015).

The tools described above can be used to assess text in experiments, whether it is used as an independent or a dependent variable. When text is manipulated as an independent variable, these tools can be employed to conduct manipulation checks.

For example, participant-generated text can be checked to ensure that it varies as expected across experimental conditions (e.g., more positive emotion words when participants are assigned to write about a positive vs. negative purchase). Further, when researcher-created text is used, as either an independent or a dependent variable, tools can be used to ensure that these texts vary in terms of the language of interest (e.g., pronouns), but do not vary in other ways (e.g., sentiment; Moore & McFerran, 2017; Packard et al., 2018).

Finally, when using text as a mediator, dictionary-based tools can be applied to participant-generated text designed to capture a hypothesized process. For example, Wu and colleagues (Wu et al., 2019) conceptualized the proportion of other-focused pronouns (e.g., they, she) in participants' open-ended responses as a reflection of attention to others and used this proportion as a mediating variable.

5 Conclusion

Language is part of almost every marketing interaction. Brands, consumers, and employees use language to communicate, persuade, and offer assistance. Consequently, by quantifying the insights hidden in language, automated textual analysis opens up a range of interesting research questions.

In this paper, we offer an accessible introduction to the three main approaches to automated text analysis, discussing how they can be used to extract meaning from text (i.e., *how*, *why*, and *what*), and how these approaches might complement each other.

Dictionaries, for example, could be used to study why some products or services get talked about for longer than others (e.g., because more concrete words or emotional language is used), or how technology shapes communication. Topic modeling of online reviews could be used to explore drivers of consumer motivations, and reasons for product or service failure. And embeddings models could be used to study cultural differences (e.g., in gender bias or discrimination), or how brands evolve (e.g. how the representation of "Tesla" has changed over the past decade) across different markets given their ability to account for context.

These three approaches can also be used to provide insight into the two key functions of language (Berger et al. 2020). First, language *impacts* the audiences that consume it. The words used by consumers, salespeople, or advertisements shape the attitudes and actions of the people that hear or read them. Packard and Berger (2021), for example, used a concreteness dictionary to show that when customer service agents use more concrete language it boosts customer satisfaction; Berger and Packard (2018) used topic modeling to test ideas about the impact of atypicality or novelty on product success; and Wang, He, and Curry (2021) used word embeddings to figure out which product attributes most impact consumer attitudes as they are expressed in online reviews.

Second, language also *reflects* things about the consumer, company, or culture that created it. What someone said, for example, provides insight into their personality and demographic characteristics, and company language sheds light on everything from attitudes towards customers to things like gender bias and discrimination.

Proserpio, Troncoso, and Valsesia (2021), for example, used dictionaries to test whether responses from hotel management reflect a gender bias. Boghrati and Berger (2022) use embeddings and a quarter of a million songs over 60 years to explore whether gender bias has changed over time. And a combination of dictionaries, topic modeling, and embeddings was used to reveal how reviewers' expressed attitudes reflect the reviewer's personal motivation to share their opinion (Chakraborty et al., 2022).

Overall, the three approaches outlined can help researchers study how text both impacts audiences and reflects things about language producers. Language can be used to both understand (and predict) consumer behavior and other marketing outcomes, as well as gain insight into people and culture more generally. Hopefully the tools outlined here will help more researchers explore this exciting area.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11002-022-09635-6.

Declarations

Conflict of Interest There is no conflict of interest or funding. No human subjects were collected, so there is no ethical approval needed.

References

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of Marketing Research, 34*, 347–356.

Aka, A., Olivola, C., Bhatia, S., & Nave, G. (2020). Computational consumer segmentation and brand management. *Advances in Consumer Research, 48*, 825–830.

Barasch, A., & Berger, J. (2014). Broadcasting and narrowcasting: How audience size affects what people share. *Journal of Marketing Research, 5*, 286–299.

Berger J, Barasch A (2015) Posting posed, choosing candid: Photo posters mispredict audience preferences. ACR North American Advances

Berger, J., Kim, Y. D., & Meyer, R. (2021). What makes content engaging? How emotional dynamics shape success. *Journal of Consumer Research, 48*, 235–250.

Berger, J., & Milkman, K. (2012). What makes online content viral? *Journal of Marketing Research, 49*, 192–205.

Berger, J., & Packard, G. (2018). Are atypical things more popular? *Psychological Science, 29*, 1178–1184.

Berger J, Rocklage MD, Packard G (2022) Expression modalities: How speaking versus writing shapes word of mouth. Journal of Consumer Research

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing, 84*(1), 1–25.

Bhatia, S., & Olivola, C. (2018). Data-driven computational brand perception. *Advances in Consumer Research, 46*, 204–208.

Bhatia S, Olivola CY (2021) Computational brand perception: Fine-tuned word embedding techniques for predicting consumer brand-trait associations. Working Paper.

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

Bogharti R, Berger JA (2022) Quantifying gender bias in consumer culture. Available at SSRN 4004777

Boleda, G (2020) Distributional semantics and linguistic theory. Annual Review of Linguistics

Boyd RL, Ashokkumar A, Seraj S,Pennebaker JW (2022) The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin. https://www.liwc.app

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods, 44*, 890–907.

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science, 35*(6), 953–975.

Chakraborty, I., Kim, M., & Sudhir, K. (2022). Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes. *Journal of Marketing Research*. https://doi.org/10.1177/00222437211052500

Chung, J., Johar, G. V., Yanyan, L., Netzer, O., & Pearson, M. (2022). Mining consumer minds: Downstream consequences of host motivations for home-sharing platforms. *Journal of Consumer Research, 48*, 817–838.

Dai, H., Chan, C., & Mogilner, C. (2020). People rely less on consumer reviews for experiential than material purchases. *Journal of Consumer Research, 46*, 1052–1075.

Dieng, A. B., Ruiz, J. R. F., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics, 8*, 439–453.

Ertimur, B., & Coskuner-Balli, G. (2015). Navigating the institutional logics of markets: Implications for strategic brand management. *Journal of Marketing, 79*, 40–61.

Gabel, S., Guhl, D., & Klapper, D. (2019). P2V-MAP: Mapping market structures for large retail assortments. *Journal of Marketing Research, 56*, 557–580.

Harris, Z (1970) Distributional structure. In: Papers in Structural and Transformational Linguistics, pp. 775–794

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing, 36*, 20–38.

Humphreys, A. (2010). Megamarketing: The creation of markets as a social process. *Journal of Marketing, 74*, 1–19.

Humphreys, A., & Wang, R. (2018). Automated text analysis for consumer research. *Journal of Consumer Research, 44*, 1274–1306.

Jarvis, W., & Goodman, S. (2005). Effective marketing of small brands: Niche positions, attribute loyalty and direct mar- keting. *Journal of Product & Brand Management, 14*(5), 292–299.

Jordan K, Pennebaker JW (2015) Seeking rewards, avoiding risks, and taking the middle ground: A language-based approach to identifying reward- vs risk-oriented thinking. https://wordwatchers.wordpress.com/tag/rubio/ Accessed February 2022

Jorge-Botana, G., Olmos, R., & Luzón, J. M. (2020). Bridging the theoretical gap between semantic representation models without the pressure of a ranking: Some lessons learnt from LSA. *Cognitive Processing, 21*, 1–21.

King, B. G., & Pearce, N. A. (2010). The contentiousness of markets: Politics, social movements, and institutional change in markets. *Annual Review of Sociology, 36*, 249–267.

Kutuzov A, Øvrelid L, Szymanski T, Velldal E (2018) Diachronic word embeddings and semantic shifts: A survey. arXiv:1806.03537

Lafreniere KC, Moore SG, Fisher RJ (2022) The power of profanity: The meaning and impact of swearwords in word-of-mouth. Forthcoming, Journal of Marketing Research

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*, 211–240.

Laurino Dos Santos H, Berger J (2022) The speed of stories: Semantic progression and narrative success. Journal of Experimental Psychology: General.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics, 4*, 151–171.

Li, A. H., & Ma, L. (2020). Charting the path to purchase using topic models. *Journal of Marketing Research, 57*, 1019–1036.

Luangrath, A. W., Peck, J., & Barger, V. A. (2017). Textual paralanguage and its implications for marketing communications. *Journal of Consumer Psychology, 27*, 98–107.

Luangrath AW, Xu Y, Wang T (2022) Paralanguage Classifier (PARA): An algorithm for automatic coding of paralinguistic nonverbal parts of speech in text. Working Paper

Melumad, S., Inman, J. J., & Pham, M. T. (2019). Selectively emotional: How smartphone use changes user-generated content. *Journal of Marketing Research, 56*, 259–275.

Mikolov T, Yih WT, Zweig, G (2013) Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 746–751

Moore, S. G. (2012). Some things are better left unsaid: How word of mouth influences the storyteller. *Journal of Consumer Research, 38*, 1140–1154.

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

Moore, S. G., & McFerran, B. (2017). She said, she said: Differential interpersonal similarities predict unique linguistic mimicry in online word of mouth. *Journal of the Association for Consumer Research, 2*, 229–245.

Nave, G., Rentfrow, J., & Bhatia, S. (2020). We are what we watch: Movies contents predicts the personality of their social media fans. *Advances in Consumer Research, 48*, 825–830.

Packard, G., & Berger, J. (2017). How language shapes word of mouth's impact. *Journal of Marketing Research, 54*, 572–588.

Packard, G., & Berger, J. (2021). How concrete language shapes customer satisfaction. *Journal of Consumer Research, 47*, 787–806.

Packard, G., Moore, S. G., & McFerran, B. (2018). (I'm) happy to help (you): The impact of personal pronoun use in customer–firm interactions. *Journal of Marketing Research, 55*, 541–555.

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.

Rocklage, M. D., & Fazio, R. H. (2020). The enhancing versus backfiring effects of positive emotion in consumer reviews. *Journal of Marketing Research, 57*, 332–352.

Rocklage MD, He S, Rucker DD, Nordgren LF (2022) Beyond sentiment: The value and measurement of consumer certainty in language. under review

Rocklage, M. D., & Luttrell, A. (2021). Attitudes based on feelings: Fixed or fleeting? *Psychological Science, 32*, 364–380.

Rocklage, M. D., Rucker, D. D., & Nordgren, L. F. (2018a). The Evaluative Lexicon 2.0: The measurement of emotionality, extremity, and valence in language. *Behavior Research Methods, 50*, 1327–1344.

Rocklage, M. D., Rucker, D. D., & Nordgren, L. F. (2018b). Persuasion, emotion, and language: The intent to persuade transforms language via emotionality. *Psychological Science, 29*, 749–760.

Rocklage, M. D., Rucker, D. D., & Nordgren, L. F. (2021). Mass-scale emotionality reveals human behaviour and marketplace success. *Nature Human Behaviour, 5*, 1323–1329.

Ryoo, J. H., Wang, X., & Lu, S. (2021). Do spoilers really spoil? Using topic modeling to measure the effect of spoiler reviews on box office revenue. *Journal of Marketing, 85*, 70–88.

Schellekens, G. A., Verlegh, P. W., & Smidts, A. (2010). Language abstraction in word of mouth. *Journal of Consumer Research, 37*, 207–223.

Shankar, V, Parsana, S (2022) An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing. Journal of the Academy of Marketing Science, 1–27.

Spiller, S. A., & Belogolova, L. (2017). On consumer beliefs about quality and taste. *Journal of Consumer Research, 43*, 970–991.

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*, 24–54.

Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science, 38*, 1–20.

Tirunillai, S, Tellis, GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation 463–479.

Toubia O, Berger J, Eliashberg J (2021) How quantifying the shape of stories predicts their success. Proceedings of the National Academy of Sciences 118, no. 26

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems, 94*, 101582.

Wang X(S), He J, Curry DJ, Ryoo JH 2021 Attribute embedding: Learning heirarchical representations of product attributes from consumer reviews Journal of Marketing https://doi.org/10.1177/00222429211047822

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063–1070.

Wu, E. C., Moore, S. G., & Fitzsimons, G. J. (2019). Wine for the table: Self-construal, group size, and choice for self and others. *Journal of Consumer Research, 46*, 508–527.

Zhang, Z., Nrusimha, A., & Hsu, M. (2018). Predicting consumer brand recall and choice using large-scale text corpora. *Advances in Consumer Research, 46*, 204–208.

Zhang K, Shaobo L, Ng S (2022) Sizes are gendered: The effect of size cues in brand names on brand stereotyping. Journal of Consumer Research