The Meaning Extraction Method: An Approach to Evaluate Content Patterns From Large-Scale Language Data

David M. Markowitz *

School of Journalism and Communication, University of Oregon, Eugene, OR, United States

Qualitative content analyses often rely on a top-down approach to understand themes in a collection of texts. A codebook prescribes how humans should judge if a text fits a theme based on rules and judgment criteria. Qualitative approaches are challenging because they require many resources (e.g., coders, training, rounds of coding), can be affected by researcher or coder bias, may miss meaningful patterns that deviate from the codebook, and often use a subsample of the data. A complementary, bottom-up approach—the Meaning Extraction Method—has been popular in social psychology but rarely applied to communication research. This paper outlines the value of the Meaning Extraction Method, concluding with a guide to conduct analyses of content and themes from massive and complete datasets, quantitatively. The Meaning Extraction Method is performed on a public and published archive of pet adoption profiles to demonstrate the approach. Considerations for communication research are offered.

Keywords: meaning extraction, thematic extraction, themes, automated text analysis, language

INTRODUCTION

A cornerstone of communication research is the evaluation of what people say and mean with content (Lovejoy et al., 2014). Traditional approaches to understand meaning from text data use content analysis to evaluate whether themes in a corpus are consistent or inconsistent with theory. Consider an example from Dixon et al. (2008), who evaluated the impact of race on police-civilian interactions for traffic stops in Cincinnati, Ohio. The authors used Communication Accommodation Theory (Giles and Smith, 1979) to evaluate how conflict among community members and police might be represented by accommodation patterns (e.g., a convergence process of adapting to another person’s speech). They predicted that different-race interactions would contain less accommodation than same-race interactions and used communication quality of the officers, among other outcomes, to evaluate this prediction (Dixon et al., 2008). Ratings by independent coders supported their hypothesis, suggesting that accommodation is related to the composition of the officer-civilian relationship.

The prior example typifies the content analysis process. Generally, researchers begin with theory, develop a codebook based on the theory to forecast themes that might emerge from the data, evaluate the reliability and validity of their codes against the data, and assess how well the theory approximated the
data based on the findings (Berelson, 1952; Riffe et al., 1998; Neendorf, 2002; Krippendorff, 2018). Content analysis is fundamental to understand the meaning of texts, but there are pragmatic constraints with this approach.

Suppose a communication researcher collected a large sample of texts (e.g., newspaper articles) with thousands of cases. How could they evaluate the texts in their sample for themes? This task would be difficult for humans to qualitatively code in an accurate and timely manner. Therefore, the researcher might select a random and more manageable (smaller) sample to content-analyze, using a codebook that draws on theory to assess if certain texts fit expected themes. This approach does not resolve a key limitation of completeness, however, since many cases would be excluded in the qualitative review (Lacy et al., 2015). Suppose a different communication researcher collected a sample of texts where themes were not easily predictable from theory because the phenomenon was understudied (e.g., transcribed confession tapes from prisoners). How could they still evaluate this corpus for themes? Here, the researcher might use grounded theory to iteratively review and analyze texts to better understand meaning (Glaser and Strauss, 1967). Grounded theory helps to elicit one’s lived experience and highlights the interests of the participant rather than the researcher (Hsieh and Shannon, 2005).

Certainly, there are tradeoffs associated with any research method, and with an increased urgency to include large sample sizes and replicable approaches in communication research (Dienlin et al., 2020; Keating and Totzkay, 2019), it is crucial that communication scholars consider additional content analysis approaches that can handle large-scale data and facilitate replications. The current article presents an automated text analysis method to understand meaning from communication data, specifically language patterns, when researchers have a large number of cases in a sample. While there is a rich history of using and developing automated approaches to understand what people mean with words (Reinert, 1990; Landauer and Dumais, 1997; Burgess et al., 1998; Beaudouin, 2016; Sbalchiero, 2018), few have been explicated for communication researchers in a transparent and holistic manner. This paper provides a tutorial and guide for the Meaning Extraction Method (Chung and Pennebaker, 2008), which is popular in social psychology but rarely applied to communication research. In fact, a search of papers from all International Communication Association journals, National Communication Association journals, and top communication journals indexed by the ISI Web of Science Journal Citation Report (Song et al., 2020) revealed only four articles that mentioned or used the method (LeFebvre et al., 2015; Kim et al., 2016; Stanton et al., 2017; Tong et al., 2020). Half of these papers were written by an author who helped to popularize the method in psychology as well.

Together, the Meaning Extraction Method originates from fields outside of communication research but is primed to contribute an additional, replicable way to analyze content from large-scale language data. There are many instances when such an additional approach might be useful. First, an additional quantitative approach can be confirmatory for qualitative analyses. If the qualitative and quantitative analyses align or are concurrent, this would increase the reliability and validity of the theory. Second, an additional approach would be useful to facilitate exploratory analyses. Theory provides a set of top-down, testable propositions that can be observed in data, but perhaps there are patterns that can also inform the theory that were unaware to researchers at the onset of the study. An additional bottom-up approach allows for themes to emerge from the data, similar to grounded theory, but in an automated manner. Altogether, the Meaning Extraction Method is not a substitute for content analysis as humans are irreplaceable for deep interpretation. Instead, it should be an additional tool in the toolkit of researchers who seek to understand what people say and mean with text.

THE MEANING EXTRACTION METHOD: ORIGINS AND CHARACTERISTICS

The idea that researchers can automatically extract themes from text is rooted in early analytic techniques by Benzecri and colleagues, who popularized the application of correspondence analysis to textual data (Benzecri, 1980). Correspondence analysis is a statistical approach that extracts dimensions from qualitative data matrices, allowing the identification of themes based on the co-occurrence of words in the text. Since this pioneering work, more recent techniques have automated the process of transforming language patterns into quantitative variables that can be used in approaches such as the Meaning Extraction Method.

The original paper introducing the Meaning Extraction Method evaluated how verbal content (e.g., adjectives, nouns) associated with personality traits. Chung and Pennebaker (2008) had over 1,100 participants fill out demographic questionnaires and a personality inventory (John and Srivastava, 1999), then presented them with a writing task. Participants were instructed “to think about who you are, who you have been in the past, and who you would like to be” and then write about who they are for 20 min.

The writing content was remarkably stable across participants and formed seven themes (words in italics are verbal descriptors from each theme): sociability (e.g., quiet, shy reserved), evaluation (e.g., ugly, fat, attractive), negativity (e.g., mad, hurt, bad), self-acceptance (e.g., wonderful, loving, lost), fitting in (e.g., interesting, funny, crazy), psychological stability (e.g., aware, healthy, emotional), and maturity (e.g., mature, successful, caring). Crucially, many of these extracted themes also correlated with personality dimensions from The Big Five. Other studies—ranging from evaluations of food-related

---

2The approach reviewed in this paper is certainly amenable to small datasets, but it offers particular advantages when human coding is impossible, implausible, resource-intensive, or if the analysis is entirely exploratory. As noted below, the size of the dataset will require researchers to adjust their input thresholds (e.g., the number of words to retain in an analysis) to create interpretable themes.

3A query of the phrase “meaning extraction method” was performed on each journal website to complete this analysis.
themes on Reddit in American cities (Blackburn et al., 2018) to how liars communicate differently than truth-tellers (Markowitz and Griffin, 2020)—have used the Meaning Extraction Method to evaluate how content matters to reveal social and psychological phenomena, demonstrating the versatility of the approach.

The framework established by Chung and Pennebaker (2008) is statistical and operates on a key assumption that different words that belong to a specific theme tend to be used together and, consequently, will co-occur in the same text segment. This approach follows conventional dimension reduction or factor analysis techniques. Note, some familiarity with statistics and Principal Component Analysis^4 (PCA) is beneficial to perform and interpret outputs from the Meaning Extraction Method, but the technique is approachable even with basic or introductory statistics training. This paper offers a guide on how to build, conceptualize, and interpret a PCA using language data (for others, see Boyd, 2017).

Language Data Compared to Questionnaire Data
A note on the characteristics of language-as-data is important before embarking on the Meaning Extraction Method. Advice by Chung and Pennebaker (2008) is especially helpful for those familiar with PCA or other dimension reduction techniques because the analysis of language data is unique relative to questionnaire items. The most notable difference between the two is the frequency of words in a communication act. Inconsistent with Likert-type scales, most words have a “modal use of zero,” which suggests that the most common value for each word is zero (Chung and Pennebaker, 2008, p. 106). Therefore, instead of counting the prevalence of a word as a raw frequency or percentage, words are assigned a binary score (1 = the word is present in a text, 0 = the word is absent from a text). This idea is further detailed in the method section below, along with other analytic conventions associated with language data.

Demonstrating the Meaning Extraction Method
The rest of this article demonstrates the process of extracting themes from large text samples. Below, public data from published research were submitted to the Meaning Extraction Method. Markowitz (2020) collected 115,318 unadopted and 560,686 adopted pet adoption profiles from Petfinder and observed that these profile types had different linguistic signatures (Study 2). The paper drew on persuasion theory (Pettig and Cacioppo, 1986) and used a dictionary-based approach with Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) to evaluate language features of influence that separated adopted from unadopted pet profiles. The data suggest adopted pet profiles contained a more analytical style with few social words relative to unadopted profiles, which had a narrative style and many social words per profile (Markowitz, 2020). No meaning extraction was performed on these data prior to the writing of this paper.

These texts were chosen to demonstrate the Meaning Extraction Method for several reasons. First, the data are public, and the entire collection is substantially larger than what would be practical to code with humans. Second, the data are highly relevant to communication researchers. Markowitz (2020) evaluated language patterns of social influence from other settings to test how well they reflected persuasion dynamics in pet adoption (Larrimore et al., 2011). Since persuasion is a top focal phenomenon in communication research (Rains et al., 2018), offering additional ways to evaluate persuasion (e.g., in this case, a pet being adopted or not as indicated by language patterns), would likely be important and welcomed by a diversity of scholars. With the Meaning Extraction Method, a previously unexplored research question can be addressed: What themes associate with persuasion success in pet adoption profiles?

In sections that follow, a basic demonstration of the meaning extraction process is performed on the pet adoption and persuasion data, though there are other approaches briefly reviewed as well. All outputs for the current study are available on the Open Science Framework (OSF: https://osf.io/trf5z/). For more information about the corpora of adopted and unadopted pet profiles, please reference the published paper (Markowitz, 2020).

Below, the method and results sections are annotated with commentary and footnotes for the purposes of clarifying an approach and statistical tests, which might be helpful for first-time Meaning Extraction Method users or qualitatively oriented researchers. By providing the method and results in this manner, a tacit aim is to offer the reader a model of how to generally report meaning extraction results as well.

METHOD
The goal of the Meaning Extraction Method, like other dimension reduction approaches, is to form a simple and interpretable number of themes from text data using content words (e.g., nouns, verbs, adjectives). The extraction process therefore removes function words that form the connective tissue of a sentence (e.g., articles, prepositions, pronouns) and low base-rate words to retain content. This procedure can be automated by different methods in Python (e.g., Natural Language Toolkit; Bird et al., 2009), but a graphical user interface called the Meaning Extraction Helper (MEH; version 2.1.07) simplifies the task (Boyd, 2017; Boyd, 2018a). MEH can process texts in spreadsheets and text files, ultimately providing a series of outputs for analysis.

In the current example, the unit of analysis—or the specific segment of text being investigated—is the individual pet adoption profile. The unit of analysis is often determined by the researcher and some considerations are made during this process to decide how texts are segmented (or delimited).
If a researcher is interested in how a construct might change over time or throughout the course of a single text (e.g., Boyd et al., 2020), they might segment a text into various “chunks.” For example, Boyd et al., 2020 evaluated the narrative structure of stories by separating novels, newspaper articles, and Supreme Court opinions into five equal-length segments, or the typical narrative structure according to prior research. Therefore, one way to segment or delimit texts is to divide them into equal-length units and track how a particular social or psychological construct operates over time linguistically. Delimiting the unit of analysis is crucial for automated text analysis because excessively small units hinder the reliable estimation of co-occurrences and prevent the identification of non-trivial co-occurrences, while very large units can make it difficult to identify an individualized theme.

### Preprocessing

MEH begins by converting words or phrases to single units (e.g., contractions such don’t are converted to do not) and shorthand to longform (e.g., gf is converted to girlfriend). There is a preset conversion list of 296 words in MEH and researchers rarely need to amend the list. MEH also automates a process called lemmatization, which transforms words to their “basic form.”

For example, Boyd (2017) suggests that the words drive, driving, and drove are converted to drive, which helps to identify word patterns at the conceptual level. Put another way, with the Meaning Extraction Method, “we tend to care less about specific variations of words than the concepts reflected by each of the words” (Boyd, 2017, p. 168).

There are several decisions the researcher must make in the meaning extraction process with MEH. First, the researcher must decide if they want to exclude texts that contain low word counts (e.g., less than 5 words). This is often performed in order to prevent texts with low word frequencies from positively skewing the results. If this is an interest for the researcher, the process can be automated with MEH. Second, the researcher must decide if they want to exclude words that have generally low frequencies across the collection of texts, a process that can also be automated with MEH. Low-frequency words will be difficult to form themes from and therefore, they are often excluded. Researchers can exclude words from the analysis if they appear below a particular threshold (e.g., ≤5% of texts). It is important to note that the number of words to retain in the analysis is often dataset-dependent and a reflection of the total number of texts and words in the corpus.

### Table 1

| Word    | Adopted corpus | Unadopted corpus |
|---------|----------------|------------------|
| Adoption | 865,480        | 141,438          |
| Love    | 612,586        | 164,581          |
| Application | 422,273   | 154,859          |
| cat     | 347,980        | 152,392          |
| Animal  | 327,526        | 152,112          |
| Fee     | 311,033        | 149,967          |
| Foster  | 304,740        | 148,931          |
| Pet     | 304,324        | 144,619          |
| Meet    | 272,348        | 136,085          |
| Spay    | 269,519        | 135,822          |
| Nauser  | 288,650        | 133,245          |
| Family  | 261,025        | 132,747          |
| Shiber  | 252,524        | 131,110          |
| Svea    | 219,714        | 116,727          |
| Pupcy   | 216,982        | 123,484          |
| year    | 209,845        | 123,484          |
| Play    | 196,797        | 120,670          |
| Time    | 193,232        | 116,727          |
| Microch | 177,483        | 116,274          |
| Train   | 174,498        | 114,485          |
| Fist    | 173,922        | 114,230          |
| Kitten  | 170,168        | 109,147          |
| Visit   | 167,530        | 109,035          |

| Word    | Adopted corpus | Unadopted corpus |
|---------|----------------|------------------|
| Adoption | 865,480        | 141,438          |
| Love    | 612,586        | 164,581          |
| Application | 422,273   | 154,859          |
| cat     | 347,980        | 152,392          |
| Animal  | 327,526        | 152,112          |
| Fee     | 311,033        | 149,967          |
| Foster  | 304,740        | 148,931          |
| Pet     | 304,324        | 144,619          |
| Meet    | 272,348        | 136,085          |
| Spay    | 269,519        | 135,822          |
| Nauser  | 288,650        | 133,245          |
| Family  | 261,025        | 132,747          |
| Shiber  | 252,524        | 131,110          |
| Svea    | 219,714        | 116,727          |
| Pupcy   | 216,982        | 123,484          |
| year    | 209,845        | 123,484          |
| Play    | 196,797        | 120,670          |
| Time    | 193,232        | 116,727          |
| Microch | 177,483        | 116,274          |
| Train   | 174,498        | 114,485          |
| Fist    | 173,922        | 114,230          |
| Kitten  | 170,168        | 109,147          |
| Visit   | 167,530        | 109,035          |

A small number of texts and words, more conservative thresholds might be appropriate relative to a large number of texts and words. The simplest and most conventional approach typically excludes words that appear in a low percentage of texts that still yield a meaningful and interpretable number of themes.

For the present analysis, content words were retained if they appeared in ≥10% of the adopted and unadopted datasets. In practice, then, MEH excluded content words that appeared in 9% or fewer of the texts from each individual corpus (adopted and unadopted profiles). This process retained a total of 89 unigrams (single words) in the adopted corpus and 80 unigrams in the unadopted corpus. While the present analysis used single words as the linguistic unit, MEH can process different n-grams as well (e.g., bigrams or two-word phrases, trigrams or three-word phrases). This criterion can be specified by the researcher before meaning extraction.

The meaning extraction process is run on each corpus separately to ensure that the themes are unique to each dataset (adopted vs. unadopted). Several outputs are provided by MEH and the most immediately relevant files are the Frequency List and Binary Matrix. The Frequency List (“2020-04-14_MEH Freq List_adopted.csv” on the OSF) provides a raw count of the number of times a word appears in the adopted profile dataset. The top 50 words from each corpus are located in Table 1. The five most frequent words in the adopted corpus were adoption (n = 865,480), love (n = 612,586), application (n = 422,273), cat (n = 411,582), and adopt (n = 347,980). The same five words were most frequently represented in the unadopted dataset, but in a different order: love (n =

---

Notes:
1. If researchers are interested in segmenting texts to evaluate how they evolve linguistically, they should consider an appropriate word count delimiter, which likely differs across settings and samples. Researchers should look at the average word count per text to determine if an a priori number of segments is appropriate. For example, if a researcher decides on five segments, but the average word count is 10 words per text (2 words per segment), this might produce less interpretable results than if the average word count is 100 words per text (20 words per segment).
2. There are many customizable options in MEH not reviewed here. The majority of analyses will use MEH presets, but customization (e.g., creating dictionary lists, or words to retain, even if they are low-base rates or function words) is possible. Please see [https://www.ryanboyd.io/software/meh/options/](https://www.ryanboyd.io/software/meh/options/) for more information.
3. Conventions for MEH place the processing date at the beginning of each filename.
4. A careful reader might suggest that adoption should belemmatized to adopt. However, both words are retained, as adopt as a verb has a different lemma than adoption as a noun.
Dimension Reduction

The dimension reduction technique discussed here is Principal Component Analysis (PCA). There are many others, including Latent Dirichlet Allocation, which are types of supervised machine learning common for text analysis (van der Meer, 2016). They require the researcher to make analytic decisions a priori (e.g., the number of themes to retain). PCA, on the other hand, is a form of unsupervised machine learning that does not, by default, require the researcher to specify a number of components (themes) to retain. PCA is similar—but not identical to—Exploratory Factor Analysis, a common data reduction technique (Bryant and Yarnold, 1995). Since PCA is one of the most common approaches for dimension reduction with language in the social sciences, it is applied here.9

We first rely on an output from MEH that describes if a content word is present or absent from a text, not the magnitude of its presence in a text. The Binary Matrix ("2020-04-14_MEH_DTM_Binary_adopted.csv" on the OSF) contains columns for each extracted content word and rows for each text (profile). Each content word has an associated binary score: 1 = the word is present in each respective text, 0 = the word is absent from each respective text. One can think of each word as a traditional scale item that will cluster with other words to form themes.

Best practices for PCA with language data often recommend that each extracted component (theme) is statistically independent from each other. To achieve this, components need to be rotated in a multidimensional space, a process that is automated by conventional statistics software (e.g., R, SPSS, Stata). Each component is comprised of items (words) that are correlated within component as indicated by item loadings (see below),11 but we often want them to be independent between components (e.g., no two components are correlated with each other). This is mathematically achieved by rotating the axes of each component with varimax rotation,12 further confirmed by simple bivariate correlations between components (rs = 0.000, ps = 1.00). Each component is perfectly uncorrelated with each other to ensure separate themes emerged.

PCAs also provide the loadings for each component. Loadings are essentially the correlations between content words that cluster together and are often lower than those on traditional questionnaires. There is a practical reason for this: people often say a phrase and do not need to repeat it, thus abiding by ideals of conversation and speaking (Grice, 1975). Therefore, component loadings are often lower than scale measures, but still reliable.

In the present analysis, items (words) are retained if they load ≥ |0.20| onto each component. This threshold is dataset-dependent, however, based on the number of cases in each corpus. Large samples with short texts (e.g., Tweets) might require more conservative thresholds (e.g., retained item loadings ≥ |0.30|), whereas large datasets with longer texts allow for less conservative thresholds (e.g., retained item loadings ≥ |0.10|). Cutoffs for item loadings should be chosen based on the interpretability of the retained themes (Chung and Pennebaker, 2008; Boyd, 2017).

It is also important to note that cross-loadings (e.g., words that appear on multiple themes or components) are quite common in PCA with language data. There is a practical reason for this as well: the same content word (e.g., store) might be used across settings but with different meanings (e.g., “I need to go to the store” vs. “I need to store items for winter”). Therefore, compared to scale items that might be removed from a PCA if they cross-load with another component at ≥ |0.40|, if most words do not cross-load onto another component(s), researchers tend to ignore them and move forward with the analysis (Chung and Pennebaker, 2008; Millar and Hunston, 2015; Boyd, 2017; Blackburn et al., 2018; Markowitz and Griffin, 2020; Markowitz and Slovic, 2020).

Analytic Approach

How many themes are extracted by the PCA process in each corpus? The number of components to retain is generally informed by the “greater than one” approach or the “fixed factor” approach. The “greater than one” approach refers to components that are retained based on eigenvalues (e.g., a number that describes the magnitude of importance for a component, reported by statistical software). The largest eigenvalue indicates the most statistically important component (e.g., it accounts for the most explained variance), though eigenvalues greater than one indicate generally reliable components and eigenvalues less than one indicate generally unreliable components (Kaiser, 1960). A scree plot (Cattell, 1966) and the amount of variance explained by each component can also assist with this approach to find reliable, interpretable components that should remain. The “fixed factor” approach, on the other hand, specifies an exact number of components that should be extracted from a dataset, typically prescribed by theory, empirical evidence, or to aid in the interpretability of components. Both approaches are presented for comparison.

Note, since language data are different than questionnaire items, dimension reduction guidelines are recommendations, not absolutes. Adjusting thresholds (e.g., the words to retain in the analysis, number of components to retain) might be effective to achieve interpretable themes.
The goal of the Meaning Extraction Method is to extract the simplest number of themes that are meaningful and interpretable, while capturing as many broad themes as possible as well. For example, a theme about sports might include words like bat, ball, pad, equipment, and field. If an excessive number of themes are extracted, one component with the words bat and ball might emerge, in addition to another component with pad and equipment. These themes are interpretable but often too narrow. Therefore, the general aim of this approach is not to isolate specific units within a theme; we care more about the general theme instead (sports).

**RESULTS**

The adopted corpus, (Kaiser-Meyer-Olkin Measure of Sampling Adequacy = 0.867\(^1\), Bartlett’s Test of Sphericity = \(\chi^2(3,916) = 7,751,318, p < 0.001\)), and unadopted corpus (Kaiser-Meyer-Olkin Measure of Sampling Adequacy = 0.874, Bartlett’s Test of Sphericity = \(\chi^2(3,160) = 1,302,846, p < 0.001\)), were well-suited for PCA using language data.

**Adopted Corpus**

In the “greater than one” analysis (Table 2), a scree plot and variance explained evidence recommended that five components be retained: pet health (Component 1), the pet adoption process (Component 2), pet upkeep (Component 3), descriptions of giving pets a happy life (Component 4), and descriptions of cats (Component 5). The “fixed factor” approach (\(n = 5\); Table 3) had several common themes (pet health, the pet adoption processes, pet upkeep), but two themes were more distinct using this extraction method. The unadopted corpus contained shelter information (Component 1) and descriptions about interacting with people (Component 3). Therefore, across PCA extraction approaches, pets that are unadopted tend to have about five reliable themes in the written portion of the profile.

There are several notable outcomes of the PCAs. First, the number of extracted themes indicates the thematic variability of each corpus. Adopted profiles are more thematically consistent than unadopted profiles, since the number of reliable components was smaller in the adopted corpus (\(n = 3\)) vs. the unadopted corpus (\(n = 5\)). The unadopted profiles were more thematically variable with an additional two components extracted. Since both corpora contained three themes in common (e.g., descriptions of the pet adoption process, pet upkeep, and pet health), the additional two components are crucial and indicate one point of difference that separates adopted vs. unadopted pet profiles. It is therefore reasonable to suggest a distinguishing feature of adopted vs. unadopted profiles is the number of themes in a corpus.

A second distinguishing feature is the themes themselves, which are rich and describe social and psychological characteristics of both pet ad types. Profiles in the adopted corpus tend to focus on core characteristics of the pet and the adoption process, presumably allowing the potential owner to

---

\(^1\)Kaiser-Meyer-Olkin (KMO) statistics often describe how suitable the data are for PCA. Statistics textbooks often suggest that values <0.70 indicate that a sample is unsuitable for PCA. This rule of thumb is often relaxed for language data. These statistics are worth reporting, but they are not as meaningful as they would be for a traditional PCA on a traditional (non-language) dataset.
understand the responsibility of ownership. These findings support evidence offered by Markowitz (2020), who suggested adopted profiles tend to have an analytic style (e.g., contain more articles and prepositions relative to pronouns and other storytelling words; Pennebaker et al., 2014) compared to unadopted profiles. Put another way, adopted profiles list the facts of adoption in terms of both style words and content words: they provide owners with a formulaic account of expectations for the pet’s health, how to apply, and upkeep at home.

Profiles in the unadopted corpus had a broader range of themes. They still covered necessary information for adoption (e.g., descriptions of the pet’s health, the adoption process of applying online, and what to expect at home), but two points of departure (e.g., happy life and/or time with people, depending on the extraction method) are social in nature. For example, the happy life theme (Component 4, Table 4), describes living with a human and the pet having a long and happy life (words in italics are items from each component). The time with people theme (Component 3, Table 5) describes experiences with people and family, enjoying playtime, and experiencing love. These themes, characteristic of the unadopted corpus, also are consistent with the findings from Markowitz (2020), who suggested that social words and humanizing details tend to undermine adoption efforts and are found in high rates for unadopted profiles (and in low rates for adopted profiles).

Taken together, the writing style patterns via function words (Markowitz, 2020) and content words reported here tell a consistent story about how adopted and unadopted pet adoption profiles are communicated. Adopted pet profiles are often written in an analytic and formal manner focusing on the adoption process and telling owners that pet immunizations are consistent story about how adopted and unadopted pet profiles are communicated. Adopted pet profiles are often written in an analytic and formal manner focusing on the adoption process and telling owners that pet immunizations are consistent story about how adopted and unadopted pet profiles are communicated. Adopted pet profiles are often written in an analytic and formal manner focusing on the adoption process and telling owners that pet immunizations are consistent story about how adopted and unadopted pet profiles are communicated. Adopted pet profiles are often written in an analytic and formal manner focusing on the adoption process and telling owners that pet immunizations are consistent story about how adopted and unadopted pet profiles are communicated.
and story-telling manner focusing on the adoption process and social aspects of the pet owning experience. Style and content therefore matter to betray pet adoption persuasion dynamics from a large archive of adoption profiles.

**DISCUSSION**

This article provided a lightweight tutorial on how to automatically prepare and statistically analyze content from text data to form themes using the Meaning Extraction Method. This bottom-up approach extracted themes based on conventions of PCA using published data from pet adoption profiles (Markowitz, 2020). The data suggest there are three common themes among adopted and unadopted pet adoption profiles. Two additional themes were extracted from the unadopted corpus and suggest further word-level differences that separate the two adoption types. Therefore, the Meaning Extraction Method offers valuable insights into how content can be explored at scale and how themes emerge in different corpora.

The goal of this article was to demonstrate an underused method that can analyze large-scale communication data to form themes from language. The relative ease and scope of this approach—statistical tests aside—deserve to be emphasized. The PCA process evaluated the writing style of nearly 700,000 pet adoption profiles, removed function words and low base-rate words, retained content (e.g., nouns, verbs) that appeared in a least 10% of each corpus, and indicated whether such words appeared in a text or not. To process the complete dataset with human coders would require a massive undertaking and it is unlikely that coders would be accurate or reliable in their assessments of data at this scale. On a mid-range laptop computer with 16 GB of RAM, the entire extraction and analytic process took approximately 40 minutes; nearly 25 minutes consisted of processing approximately 700,000 texts with MEH. The Meaning Extraction Method offers a replicable, reproducible, and scalable approach that can allow researchers to observe the themes people communicate across a diversity of settings.

**Appraisal of the Meaning Extraction Method**

Having read about the meaning extraction process, it is also important to review the benefits and challenges of this approach. The benefits of the Meaning Extraction Method include the ability to analyze massive and complete datasets for themes, the replicability and reproducibility of the approach, the use of statistical methods that are mainstream for social scientists and require little coding expertise and the ability to facilitate thematic exploration that complements theory. This does not suggest that exploratory work of this kind is atheoretical, but exploration can offer insights into theory that other models or frameworks might miss. For example, mobile dating deception research suggests that most lies are communicated as a result of self-presentation (e.g., to appear likable, interesting, or attractive) or availability management goals (e.g., to avoid meeting another person, but still maintain the relationship) (Markowitz and Hancock, 2018). Applying the Meaning Extraction Method to mobile dating conversations might clarify the specific aspects of self-presentation that people focus on when lying (e.g., appearance, entertainment interests) and additional availability management strategies people use to avoid an activity (e.g., discussing work, their phone died).

The Meaning Extraction Method is also useful because it lowers the cost of entry into analyzing content when resources might be scarce. One could imagine a scenario where the cost of starting a qualitative content analysis is burdensome. Qualitative content analyses require time and human resources that are finite (e.g., attention). Therefore, an option might be to use the Meaning Extraction Method as a first glance into the data to investigate if it is worth performing human coding. If themes fail to emerge or they seem unintuitive, there are at least two possible explanations: 1) the data were not fit for PCA and require a human to review the texts, or 2) there is little consistency in the data, and it is unclear if they would be worth reviewing qualitatively. The Meaning Extraction Method deserves attention from scholars who want to use an automated approach to extract themes or those who want to evaluate their data and assess if qualitative review would be valuable.

The challenges of this approach are also nontrivial and deserve attention. First, the Meaning Extraction Method does not make any assumptions about data quality. Therefore, bad data (e.g., many misspellings, noisy text; Subramaniam et al., 2009) might produce uninterpretable PCA results (e.g., the “garbage in, garbage out” principle). Researchers should ensure that their texts are preprocessed and cleaned according to best practices for automated text analysis before meaning extraction is performed (Boyd and Pennebaker, 2015; Boyd, 2017). Second, at some level, results from the PCA are largely descriptive and cannot explain why people wrote with particular thematic patterns. Removing function words and low base-rate words also remove some interpretability of the text and limits the researcher’s ability to infer what people were trying to accomplish with their language patterns. Often, a qualitative review of these texts after meaning extraction is beneficial and necessary to disambiguate unclear themes or “sanity check” the data. Third, while the meaning extraction process can be performed on a number of non-English texts, more work is needed to validate the procedure for other languages (for examples, see Ramirez-esparza et al., 2008; Rodriguez-Arauz et al., 2017; Ikizer et al., 2019). Fourth, humans are still involved in the development of theme names and their interpretation. Objectivity might be an issue when forming themes using this approach.

Finally, the Meaning Extraction Method is still a “bag of words” approach that ignores the order of words and the contextual meaning of words within a sentence. It therefore relies on a probabilistic model of how people use words to make inferences about word patterns and themes (Harris, 1954; Boyd, 2018b). For example, people who use the words happy, amazing, and awesome have a high probability of feeling and expressing positive affect, and people who use the words hate, disgust, and awful have a high probability of feeling and expressing negative affect. How words are counted with this approach is important to be aware of, but despite its
limitations, the Meaning Extraction Method does not introduce systematic misinterpretation concerns for most large-scale analyses. More language per text and texts per sample are better for “bag of words” analyses, since the “true signal” of a corpus will be revealed when enough words reflect and identify a specific construct (Boyd, 2017).

A Note on Validity

How do communication researchers know if their extracted themes are valid? Principles of measurement validity from quantitative content analysis are also relevant here, specifically face validity and concurrent validity (Riffe et al., 1998). Face validity, or the degree to which the measurement of a concept seems reasonable, is paramount with the Meaning Extraction Method because the researcher makes many decisions in the analytic process that will affect thematic interpretation (e.g., the number of themes to extract, the factor loading threshold, the number of words to retain, the types of words to retain, among many others). A researcher should ask: Do the extracted themes make broad clusters of topics, or are they too specific and narrow? Are the extracted themes reasonable given what we know about the corpus and its construction? What are the costs and benefits of extracting more (or less) themes? Face validity checks help the researcher identify if the Meaning Extraction Method is appropriate for their dataset, or if the aforementioned decisions during the analytic process might need reconsideration.

Concurrent validity is the idea that a measure in one study is associated with the measure from a second study. For the Meaning Extraction Method, this would be similar to the findings from a top-down dictionary tool (e.g., LIWC) being consistent with bottom-up thematic extraction. Recall, Markowitz (2020) observed that social words and humanizing details tend to hurt adoption efforts and are more strongly associated with unadopted vs. adopted pet profiles. Social themes, via the Meaning Extraction Method, were more apparent in the unadopted corpus than the adopted corpus according to the thematic extraction (e.g., time with people). Therefore, the cross-check between two analyses strengthens the theory that adopted and unadopted pet adoption profiles tend to have different patterns of social references. Concurrent validity, if established, is an important way to identify that extracted themes support a robust finding across measurement techniques.

CONCLUSION

The Meaning Extraction Method relies on a bottom-up framework that allows themes to emerge from human interpretation and deep reading. Communication researchers might view these approaches as apples and oranges. Ultimately, however, apples and oranges are both fruit. The Meaning Extraction Method and content analysis are trying to achieve a similar objective: to observe and make sense of themes from communication content via language patterns.

There is room for the two approaches (and others) to inform each other and the most impactful science likely draws inspiration from both. Mixed-method approaches, those that combine a deep reading of texts and the exploration of content patterns that emerge from automatic theme extraction, might be advantageous for researchers. For instance, combining these approaches might be particularly impactful for social media researchers who acquire sensitive data (e.g., episodes of cyberbullying) and want to parse conversations for themes to understand vulnerable communities (Kazerooni et al., 2018), scholars interested in privacy and content moderation might want to observe how corporate policies differ by platform to understand the companies better (Gillespie, 2018), and science communication researchers might want to evaluate themes in polarized media reporting of climate change to evaluate trends over time (Feldman et al., 2017). The Meaning Extraction Method can offer an exploratory way to conceptualize collections of texts and when paired with deep qualitative analyses, can form a holistic evaluation of what people mean with words.

Taken together, this paper describes and demonstrates the Meaning Extraction Method as an automated approach to understand content patterns and themes. This replicable and reproducible method can also be used as an exploratory technique to measure the word patterns that cluster to form themes. The approach is a useful complement to qualitative coding, a resource-friendly “first pass” to analyze content, and a way to evaluate the boundaries of a theory. Scholarship benefits from multidimensional and interdisciplinary approaches to science and the current paper offers a guide for analyzing content through automated means.

AUTHOR CONTRIBUTIONS

DM wrote the entire article and performed all analyses.

ACKNOWLEDGMENTS

Much gratitude is owed to Ryan Boyd, Amanda Cote, and Maxwell Foxman for their input. I also thank the editor and two reviewers for their thoughtful comments on this paper.

REFERENCES

Beaudouin, V. (2016). Statistical analysis of textual data: benzécri and the French school of data analysis. Glottometrics. 33, 56–72.

Benzécri, J. P. (1980). Pratique de l’analyse des données. Analyse des correspondances et classification. Exposé élémentaire. Paris, France: Dunod.

Berelson, B. (1952). Content analysis in communication research. Mumbai, India: Free Press.

Bird, S., Loper, E., and Klein, E. (2009). Natural language processing with Python. Newton, MA: O’Reilly Media Inc.

Blackburn, K. G., Yilmaz, G., and Boyd, R. L. (2018). Food for thought: exploring how people think and talk about food online. Appetite. 123, 390–401. doi:10.1016/j.appet.2018.01.022
Song, H., Eberl, J.-M., and Eisele, O. (2020). Less fragmented than we thought? Toward clarification of a subdisciplinary linkage in communication science. J. Commun. 70 (3), 310–334. doi:10.1093/joc/jqaa009

Stanton, A. M., Meston, C. M., and Boyd, R. L. (2017). Sexual self-schemas in the real world: investigating the ecological validity of language-based markers of childhood sexual abuse. Cyberpsychol. Behav. Soc. Netw. 20 (6), 382–388. doi:10.1089/cyber.2016.0657

Subramaniam, L. V., Roy, S., Faruquie, T. A., and Negi, S. (2009). “A survey of types of text noise and techniques to handle noisy text,” in Proceedings of the third workshop on analytics for noisy unstructured text data, AND 2009, Barcelona, Spain, July, 2009, 115–122. doi:10.1145/1568296.1568315

Tong, S. T., Corriero, E. F., Wibowo, K. A., Makki, T. W., and Slatcher, R. B. (2020). Self-presentation and impressions of personality through text-based online dating profiles: a lens model analysis. New Media Soc. 22, 875–895. doi:10.1177/1461444819872678

van der Meer, T. G. L. A. (2016). Automated content analysis and crisis communication research. Publ. Relat. Rev. 42 (5), 952–961. doi:10.1016/j.pubrev.2016.09.001

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Markowitz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.