frontiers in Communication

METHODS
published: 23 February 2021
doi: 10.3389/fcomm.2021.588823

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

OPEN ACCESS

Edited by:
Sidarta Ribeiro, Federal University of Rio Grande do Norte, Brazil

Reviewed by:
Luís Faísca, University of Algarve, Portugal
Pekka Isotalus, Tampere University, Finland

***Correspondence:**
David M. Markowitz
dmark@uoregon.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *29 July 2020*
Accepted: *11 January 2021*
Published: *23 February 2021*

Citation:
Markowitz DM (2021) The Meaning Extraction Method: An Approach to Evaluate Content Patterns From Large-Scale Language Data. Front. Commun. 6:588823. doi: 10.3389/fcomm.2021.588823

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[1] 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

[1]In this paper, a *corpus* refers to a collection of texts.

data based on the findings (Berelson, 1952; Riffe et al., 1998; Neuendorf, 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.[2] 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).[3] 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 Benzécri and colleagues, who popularized the application of correspondence analysis to textual data (Benzécri, 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

[2]The 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.

[3]A 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 (Petty 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).

[4]Components are the themes formed in the meaning extraction process with language data. For the purposes of this paper, components and themes are interchangeable terms.

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.[5] 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[6]. 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. With 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.

TABLE 1 | The 50 most frequent unigrams across corpora.

Adopted corpus				Unadopted corpus			
Word	***n***	**Word**	***n***	**Word**	***n***	**Word**	***n***
Adoption	865,480	Mix	164,561	Love	141,438	Train	30,944
Love	612,586	Look	162,313	Adoption	139,974	Email	30,868
Application	422,273	Email	156,655	Cat	115,445	Fill	29,362
cat	411,582	Great	154,859	Application	72,865	Great	28,991
Adopt	347,980	Call	153,392	Adopt	63,973	Best	28,964
Animal	327,526	Urllink	152,112	Foster	62,229	Information	28,792
Fee	311,033	month	149,957	Pet	57,809	Contact	28,533
Foster	304,740	Information	148,931	Animal	55,740	Girl	28,151
Pet	304,324	Fill	144,619	Rescue	47,690	Microchip	27,676
Meet	272,348	Age	136,085	Family	45,973	pm	27,526
Spay	269,519	Give	135,822	year	45,640	Call	27,143
Neuter	268,650	day	133,242	Fee	44,573	Urllink	27,056
Family	261,025	Test	132,747	Meet	44,132	Care	26,053
Shelter	252,524	People	131,110	Time	43,748	Live	24,851
Rescue	235,096	Best	129,176	Sweet	43,394	Test	24,155
Sweet	219,714	Vaccination	127,918	Spay	42,354	Forever	24,141
Puppy	216,982	Girl	125,070	Neuter	40,743	Find	23,696
year	209,845	Care	123,484	Play	40,639	day	23,164
Play	196,797	Website	120,570	Shelter	38,477	Age	22,979
Time	193,232	Forever	116,727	Look	35,908	See	22,540
Microchip	177,483	Find	116,274	Interest	34,251	Life	22,274
Train	174,488	week	114,485	Visit	33,542	8/	22,117
Interest	173,922	Vet	114,230	People	33,100	Website	21,942
Kitten	170,168	Available	109,147	Kitten	32,104	Boy	21,237
Visit	167,530	Contact	109,053	Give	31,244	Kitty	20,435

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[7] ("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*[8] (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 =

[5]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).
[6]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/ for more information.

[7]Conventions for MEH place the processing date at the beginning of each filename.
[8]A careful reader might suggest that *adoption* should be lemmatized to *adopt*. However, both words are retained, as *adopt* as a verb has a different lemma than *adoption* as a noun.

141,438), *adoption* (n = 139,974), *cat* (n = 115,445), *application* (n = 72,865), *adopt* (n = 63,973). The frequency list also describes the percent of texts that contained each word, which when sorted, might slightly change the order of the most frequent words. Such descriptive information is crucial for an overview of each dataset and for a cursory glance at what people communicated. Next, the Binary Matrix is used in the dimension reduction process, which is outlined below.

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 unsupervized 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"[10] 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 $\geq |0.20|$ 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 $\geq |0.30|$), whereas large datasets with longer texts allow for less conservative thresholds (e.g., retained item loadings $\geq |0.10|$). 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 $\geq |0.40|$, 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.

[9]It is also worth noting that performing a PCA versus an Exploratory Factor Analysis produces substantively equivalent results.

[10]DTM = document term matrix. Documents or texts are represented in rows and terms (words) are represented in columns.

[11]Loadings range from −1 (a perfect inverse relationship to the theme) to +1 (a perfect positive relationship to the theme).

[12]There are other types of rotations (IBM, 2020), but for the components to be entirely uncorrelated, varimax rotation is used.

TABLE 2 | PCA results for adopted Petfinder corpus using the "greater than one" method.

C1		C2		C3	
Pet health		**Pet adoption process**		**Pet upkeep**	
λ	**%**	**λ**	**%**	**λ**	**%**
4.26	**4.79**	**2.47**	**2.77**	**2.24**	**2.51**
Word	**Loading**	**Word**	**Loading**	**Word**	**Loading**
Neuter	0.702	Fill	0.799	Train	0.731
Spay	0.700	Application	0.735	House	0.691
Microchip	0.679	Online	0.693	Crate	0.681
Test	0.637	Urllink	0.352	Work	0.417
Fee	0.560	Interest	0.315		
Heartworm	0.556	Adopt	0.272		
Vaccination	0.554				
Rabies	0.535				
Flea	0.525				
Adoption	0.443				

Note. C1–C3 = component numbers. λ = eigenvalues. % = percent variance explained by each component. At most, the top 10 words per component are displayed.

TABLE 3 | PCA results for adopted Petfinder corpus using the "fixed factor" method.

C1		C2		C3	
Pet adoption process		**Pet upkeep**		**Pet health**	
λ	**%**	**λ**	**%**	**λ**	**%**
5.22	**5.86**	**4.64**	**5.21**	**4.59**	**5.16**
Word	**Loading**	**Word**	**Loading**	**Word**	**Loading**
Application	0.576	Train	0.583	Neuter	0.697
Adoption	0.487	Crate	0.560	Spay	0.694
Fill	0.478	Walk	0.512	Microchip	0.629
Process	0.478	Leash	0.504	Test	0.569
Adopt	0.476	Love	0.441	Fee	0.559
Website	0.464	House	0.439	Heartworm	0.549
Visit	0.457	Great	0.388	Rabies	0.533
Interest	0.434	Family	0.370	Flea	0.516
Urllink	0.415	Play	0.366	Vaccination	0.498
Online	0.408	Work	0.360	Vaccine	0.379

Note. C1–C3 = component numbers. λ = eigenvalues. % = percent variance explained by each component. At most, the top 10 words per component are displayed.

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[13], 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, a scree plot and variance explained evidence recommended that three components should be retained (see **Table 2**): information about pet health (Component 1), the pet adoption process (Component 2), and pet upkeep (Component 3). The "fixed factor" approach ($n = 3$) produced substantively identical results (**Table 3**), though the importance of each component was slightly rearranged.

Unadopted Corpus

In the "greater than one" analysis (**Table 4**), 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 5**) 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

[13]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.

TABLE 4 | PCA results for unadopted Petfinder corpus using the "greater than one" method.

C1		C2		C3		C4		C5	
Pet health		**Pet adoption process**		**Pet upkeep**		**Happy life**		**Cats**	
λ	**%**	**λ**	**%**	**λ**	**%**	**λ**	**%**	**λ**	**%**
3.78	**4.72**	**3.24**	**4.04**	**2.12**	**2.65**	**2.06**	**2.58**	**1.90**	**2.37**
Word	**Loading**	**Word**	**Loading**	**Word**	**Loading**	**Word**	**Loading**	**Word**	**Loading**
Spay	0.734	Application	0.767	Train	0.644	Life	0.582	cat	0.616
Neuter	0.729	Fill	0.724	House	0.613	Live	0.469	Kitten	0.523
Microchip	0.719	Online	0.598	Kid	0.536	Long	0.444	Kitty	0.521
Vaccination	0.662	Website	0.574	Great	0.355	Happy	0.380	Fiv	0.494
Test	0.635	Urllink	0.513	Walk	0.346	Time	0.303	Mix	–0.434
Fee	0.629	Interest	0.484	Work	0.319	Learn	0.300		
		Adoption	0.479	People	0.260	Human	0.286		
		Adopt	0.441						

Note. C1–C5 = component numbers. λ = eigenvalues. % = percent variance explained by each component. Negative loadings suggest that a particular word correlates negatively with the theme, overall. Therefore, in C5, unadopted pet profiles that tend to discuss cats use words like cat, kitten, kitty, and fiv (e.g., feline immunodeficiency virus), but tend to not use the word mix. At most, the top 10 words per component are displayed.

TABLE 5 | PCA results for unadopted Petfinder corpus using the "fixed factor" method.

C1		C2		C3		C4		C5	
Shelter information		**Pet health**		**Time with people**		**Pet adoption process**		**Pet upkeep**	
λ	**%**	**λ**	**%**	**λ**	**%**	**λ**	**%**	**λ**	**%**
3.83	**4.79**	**3.81**	**4.76**	**3.80**	**4.74**	**3.30**	**4.13**	**2.35**	**2.94**
Word	**Loading**	**Word**	**Loading**	**Word**	**Loading**	**Word**	**Loading**	**Word**	**Loading**
Visit	0.487	Spay	0.713	Love	0.555	Fill	0.702	Train	0.522
Adopt	0.435	Neuter	0.711	Play	0.484	Online	0.694	Mix	0.488
Animal	0.426	Microchip	0.699	Toy	0.431	Application	0.691	Walk	0.412
Information	0.422	Test	0.654	People	0.374	@8	0.659	House	0.337
Care	0.413	Vaccination	0.619	Time	0.360	pm	0.643	year	0.290
Rescue	0.404	Fee	0.618	Enjoy	0.340	Adoption	0.469	cat	–0.386
Email	0.383	Fiv	0.505	Live	0.325	Website	0.445	Kitty	–0.459
Interest	0.370	Age	0.308	Give	0.324	Urllink	0.336	Kitten	–0.488
Available	0.359	month	0.251	Human	0.324				
Shelter	0.353			Family	0.320				

Note. C1–C5 = component numbers. λ = eigenvalues. % = percent variance explained by each component. Negative loadings suggest that a particular word correlates negatively with the theme, overall. Therefore, in C5, unadopted pet profiles that tend to discuss pet upkeep use words like train, mix, walk, *and* house *(e.g., words that might be associated with pet upkeep for dogs), but tend to not use the word* cat, kitty, *or* kitten*. At most, the top 10 words per component are displayed.*

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 play time*, 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 in order. Unadopted pet profiles are often written in a narrative

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; Rodríguez-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 statistics, whereas content analyses are largely top-down processes that allow 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 and 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

Boyd, R. L., Blackburn, K. G., and Pennebaker, J. W. (2020). The narrative arc: revealing core narrative structures through text analysis. *Sci Adv.* 6 (32), eaba2196. doi:10.1126/sciadv.aba2196

Boyd, R. L. (2018b). Mental profile mapping: a psychological single-candidate authorship attribution method. *PloS One* 13 (7), e0200588. doi:10.1371/journal.pone.0200588

Boyd, R. L. (2018a). Meaning extraction helper (2.1.07). Available at: https://meh.ryanb.cc (Accessed April 10, 2020).

Boyd, R. L., and Pennebaker, J. W. (2015). "A way with words: using language for psychological science in the modern era," in *Consumer psychology in a social media world*. Editors C. Dimofte, C. Haugtvedt, and R. Yalch (Abingdon, United Kingdom: Routledge), 222–236.

Boyd, R. L. (2017). "Psychological text analysis in the digital humanities," in *Data analytics in digital humanities*. Editor S. Hai-Jew (Cham, Switzerland: Springer International Publishing), 161–189. doi:10.1007/978-3-319-54499-1_7

Bryant, F. B., and Yarnold, P. R. (1995). "Principal-components analysis and exploratory and confirmatory factor analysis," in *Reading and understanding multivariate statistics*. Editors L. G. Grimm and P. R. Yarnold (Washington, DC: American Psychological Association), 99–136.

Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in context space: words, sentences, discourse. *Discourse Process.* 25 (2–3), 211–257. doi:10.1080/01638539809545027

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behav. Res.* 1 (2), 245–276. doi:10.1207/s15327906mbr0102_10

Chung, C. K., and Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: an automated meaning extraction method for natural language. *J. Res. Pers.* 42 (1), 96–132. doi:10.1016/J.JRP.2007.04.006

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., et al. (2020). An agenda for open science in communication. *J. Commun.* 1–26 doi:10.1093/JOC/JQZ052

Dixon, T. L., Schell, T. L., Giles, H., and Drogos, K. L. (2008). The influence of race in police-civilian interactions: a content analysis of videotaped interactions taken during cincinnati police traffic stops. *J. Commun.* 58 (3), 530–549. doi:10.1111/j.1460-2466.2008.00398.x

Feldman, L., Hart, P. S., and Milosevic, T. (2017). Polarizing news? Representations of threat and efficacy in leading US newspapers' coverage of climate change. *Publ. Understand. Sci.* 26 (4), 481–497. doi:10.1177/0963662515595348

Giles, H., and Smith, P. (1979). "Accommodation theory: optimal levels of convergence," in *Language and social psychology*. Editors H. Giles and R. N. St. Clair (Oxford, United Kingdom: Blackwell), 45–65.

Gillespie, T. (2018). *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media*. London, United Kingdom: Yale University Press.

Glaser, B. G., and Strauss, A. L. (1967). *The discovery of grounded theory: strategies for qualitative research*. New Delhi, India: Aldine.

Grice, H. P. (1975). "Logic and conversation," in *Syntax and semantics 3: speech acts*. Editors P. Cole and J. Morgan (Cambridge, MA: Academic Press), 3, 41–58.

Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162. doi:10.1080/00437956.1954.11659520

Hsieh, H. F., and Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qual. Health Res.* 15 (9), 1277–1288. doi:10.1177/1049732305276687

IBM (2020). Factor analysis rotation. Available at: https://www.ibm.com/support/knowledgecenter/SSLVMB_23.0.0/spss/base/idh_fact_rot.html (Accessed April 10, 2020).

Ikizer, E. G., Ramírez-Esparza, N., and Boyd, R. L. (2019). #sendeanlat (#tellyourstory): text analyses of tweets about sexual assault experiences. *Sex. Res. Soc. Pol.* 16 (4), 463–475. doi:10.1007/s13178-018-0358-5

John, O. P., and Srivastava, S. (1999). "The big five tait taxonomy: history, measurement, and theoretical perspectives," in *Handbook of personality: theory and research*. Editors L. A. Pervin and O. P. John (New York, NY: Guilford Press), 102–138.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* 20 (1), 141–151. doi:10.1177/001316446002000116

Kazerooni, F., Taylor, S. H., Bazarova, N. N., and Whitlock, J. (2018). Cyberbullying bystander intervention: the number of offenders and retweeting predict likelihood of helping a cyberbullying victim. *J. Computer-Mediated Commun.* 23 (3), 146–162. doi:10.1093/JCMC/ZMY005

Keating, D. M., and Totzkay, D. (2019). We do publish (conceptual) replications (sometimes): publication trends in communication science, 2007–2016. *Annals of the International Communication Association.* 43 (3), 225–239. doi:10.1080/23808985.2019.1632218

Kim, E., Hou, J., Han, J. Y., and Himelboim, I. (2016). Predicting retweeting behavior on breast cancer social networks: network and content characteristics. *J. Health Commun.* 21 (4), 479–486. doi:10.1080/10810730.2015.1103326

Krippendorff, K. (2018). *Content analysis: an introduction to its methodology*. Thousand Oaks, CA: Sage Publications.

Lacy, S., Watson, B. R., Riffe, D., and Lovejoy, J. (2015). Issues and best practices in content analysis. *Journal. Mass Commun. Q.* 92 (4), 791–811. doi:10.1177/1077699015607338

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104 (2), 211–240. doi:10.1037/0033-295X.104.2.211

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D. M., and Gorski, S. (2011). Peer to peer lending: the relationship between language features, trustworthiness, and persuasion success. *J. Appl. Commun. Res.* 39 (1), 19–37. doi:10.1080/00909882.2010.536844

LeFebvre, L., LeFebvre, L., Blackburn, K., and Boyd, R. (2015). Student estimates of public speaking competency: the meaning extraction helper and video self-evaluation. *Commun. Educ.* 64 (3), 261–279. doi:10.1080/03634523.2015.1014384

Lovejoy, J., Watson, B. R., Lacy, S., and Riffe, D. (2014). Assessing the reporting of reliability in published content analyses: 1985–2010. *Commun. Methods Meas.* 8 (3), 207–221. doi:10.1080/19312458.2014.937528

Markowitz, D. M., and Griffin, D. J. (2020). When context matters: how false, truthful, and genre-related communication styles are revealed in language. *Psychol. Crime Law.* 26 (3), 287–310. doi:10.1080/1068316X.2019.1652751

Markowitz, D. M., and Hancock, J. T. (2018). Deception in mobile dating conversations. *J. Commun.* 68 (3), 547–569. doi:10.1093/joc/jqy019

Markowitz, D. M. (2020). Putting your best pet forward: language patterns of persuasion in online pet advertisements. *J. Appl. Soc. Psychol.* 50 (3), 160–173. doi:10.1111/jasp.12647

Markowitz, D. M., and Slovic, P. (2020). Social, psychological, and demographic characteristics of dehumanization toward immigrants. *Proc. Natl. Acad. Sci. Unit. States Am.* 117 (17), 9260–9269. doi:10.1073/pnas.1921790117

Millar, N., and Hunston, S. (2015). Adjectives, communities, and taxonomies of evaluative meaning. *Funct. Lang.* 22 (3), 297–331. doi:10.1075/fol.22.3.01mil

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., and Beaver, D. I. (2014). When small words foretell academic success: the case of college admissions essays. *PloS One.* 9 (12), e115844. doi:10.1371/journal.pone.0115844

Pennebaker, J. W., Booth, R. J., Boyd, R. L., and Francis, M. E. (2015). *Linguistic Inquiry and word count: LIWC2015*. Austin, TX, Pennebaker Conglomerates.

Petty, R. E., and Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Adv. Exp. Soc. Psychol.* 19, 123–205. doi:10.1016/S0065-2601(08)60214-2

Rains, S. A., Levine, T. R., and Weber, R. (2018). Sixty years of quantitative communication research summarized: lessons from 149 meta-analyses. *Annals of the International Communication Association.* 42 (2), 105–124. doi:10.1080/23808985.2018.1446350

Ramirez-esparza, N., Ramirez-esparza, N., Chung, C. K., Kacewicz, E., and Pennebaker, J. W. (2008). "The psychology of word use in depression forums in English and in Spanish: testing two text analytic approaches," in Proceedings of the ninth international AAAI conference on Web and social media, Oxford, United Kingdom, May, 2008, 102–108. doi:10.1.1.371.2720

Reinert, M. (1990). Alceste une méthodologie d'analyse des données textuelles et une application: aurelia De Gerard De Nerval. *Bull. Sociol. Methodol.* 26 (1), 24–54. doi:10.1177/075910639002600103

Riffe, D., Lacy, S., and Fico, F. (1998). *Analyzing media messages: using quantitative content analysis in research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rodríguez-Arauz, G., Ramírez-Esparza, N., Pérez-Brena, N., and Boyd, R. L. (2017). Hablo inglés y español: cultural self-schemas as a function of language. *Front. Psychol.* 8, 885. doi:10.3389/fpsyg.2017.00885

Sbalchiero, S. (2018). "Topic detection: a statistical model and a quali-quantitative method," in *Quantitative methods in the humanities and social sciences*. Cham, Switzerland: Springer, 189–210. doi:10.1007/978-3-319-97064-6_10

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.