Received: 28 March 2023 | Revised: 4 May 2023 | Accepted: 12 May 2023

DOI: 10.1111/joss.12854

ORIGINAL ARTICLE

Journal of Sensory Studies | WILEY

Exploring cider website descriptions using a novel text mining approach

Martha D. Calvert[1] | Elizabeth Cole[1] | Clinton L. Neill[2] | Amanda C. Stewart[1] | Susan R. Whitehead[3] | Jacob Lahne[1]

[1]Department of Food Science and Technology, Virginia Tech, Blacksburg, Virginia, USA

[2]Department of Population Medicine and Diagnostic Sciences, Cornell University, Ithaca, New York, USA

[3]Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia, USA

Correspondence
Martha D. Calvert, Department of Food Science and Technology, Virginia Tech, Blacksburg, VA, USA.
Email: mdcalver@vt.edu

Funding information
USDA-NIFA AFRI, Grant/Award Number: 2020-68006-31682

Abstract

Rapid methods of text analysis are increasingly important tools for efficiently extracting and understanding communication within the food and beverage space. This study aimed to use frequency-based text mining and biterm topic modeling (BTM) as tools for analyzing how cider products are communicated and marketed on cider-producer websites for products made in Virginia, Vermont, and New York. BTM has been previously used to explore topics in small corpora of text data, and frequency-based text mining is efficient for exploring patterns of text across different documents or filters. The present dataset comprised 1115 cider products and their website descriptions extracted from 124 total cider-producer websites during 2020 and 2021. Results of the text mining analyses suggest that cider website descriptions emphasize food-pairing, production, and sensory quality information. Altogether, this research presents the text mining approaches for exploring food and beverage communication.

Practical applications

This research will be valuable to stakeholders in the United States' cider industry by providing relevant insight as to how cider marketing and sensory communication varies based on extrinsic product factors, such as geography and packaging. This research also demonstrates the efficiency and potential of text mining tools for exploring language and communication related to foods, beverages, and sensory quality. Further, this research provides a framework for extracting sensory-specific language from a large corpus of data, which may be adopted by other researchers wishing to apply rapid descriptive methods in the sensory, quality, and consumer research fields.

1 | INTRODUCTION

Various scientific disciplines, including but not limited to linguistics, anthropology, and data science (Jurafsky, 2014; Riley & Paugh, 2018), explore how food is communicated and valued among producers, consumers, and other actors in the food system. In the sensory evaluation field specifically, understanding language and communication is important for understanding food product experiences and perceptions—particularly those which drive repeat consumption and purchasing.

Purchasing and consuming food entails a multisensory experience (Lawless & Heymann, 2010b; Riley & Paugh, 2018) that is heavily influenced by environmental context (Betancur et al., 2020; Lahne & Trubek, 2014; Lawless & Heymann, 2010a; Meiselman, 2019; Stelick & Dando, 2018). Context in the sensory experience can include but is not limited to music playing in the background, visual surroundings in the consumption environment, and *information* about a product that is being consumed, whether that information is on a package or online platform. In the case of the latter, text-based communication is a pertinent type of context in sensory experiences that can influence expectations about how a product might be or is perceived. For example, information about the typicality of grape varieties and growing regions facilitates cognitive associations and expectations relevant to the sensory experience (Sáenz-Navajas & Jeffery, 2021; Spence, 2020a; Thomas & Pickering, 2003). Recent research has investigated ways in which *sensory-relevant* terms are used in everyday contexts to describe products in order to get a broader understanding of how sensory quality is communicated both post- and pre-consumption, but this is an area of ongoing research (L. Hamilton, 2022).

Qualitative research methods have been most commonly utilized to explore the language and communication of food. These methods can include content analysis (Krippendorff, 2019; Weber, 1990), document analysis (Bowen, 2009), and reflexive thematic analysis (Braun & Clarke, 2006, 2013) which can identify overarching, consistent themes or topics across textual and visual data. Researchers have previously used these methods and extensions of these methods to explore wine blogs (Beninger et al., 2014; Doyle et al., 2012), cookbooks (Parys, 2013), marketing messages (Vasiljevic et al., 2018), and even written responses of consumers' perceptions of processed foods (Sadler et al., 2022). However, qualitative methods for language analysis can be time intensive and more influenced by researcher bias than comparable quantitative text mining approaches (Braun & Clarke, 2013; Creswell & Poth, 2018; Peschel et al., 2019), thus presenting the need for research methods that can quickly and effectively investigate how food is discussed.

A promising alternative to qualitative research methods for exploring food language and food sensory experiences is the application of automated text mining tools that allow researchers or companies to quickly process and analyze text (Peschel et al., 2019; Tao et al., 2020). Natural language processing (NLP) is the general area of research and application of automated, quantitative language processing that involves developing and training a statistical model to process unstructured text. For the purposes of sensory-specific research, NLP is an increasingly popular tool and has been used in different ways for different purposes (Jurafsky, 2014; Souza Gonzaga et al., 2020; Tian et al., 2021; Yoo et al., 2023). An overview of NLP and its application in the sensory science field can be found in L. M. Hamilton and Lahne (2023).

Within NLP is a branch of text analysis called topic modeling, which explores patterns in word co-occurrences in order to form broad topics or themes. Analyzing word associations can be done based on how often terms occur together, based on semantic similarity (i.e., trees, orchard, apples, apple seeds), or based on the two approaches combined. This text mining process can allow researchers to quickly investigate themes within text-based data in a way similar to that of reflexive thematic analysis (Braun & Clarke, 2013). Latent Dirichlet allocation (LDA) is a standard topic-modeling methodology that analyzes word co-occurrences within different documents, which comprise a very large overall dataset (An, 2022; Blei et al., 2003; Blei & Lafferty, 2009; Silge & Robinson, 2017; Vidal et al., 2022).

Recently, Cheng et al. (2014) proposed an alternative to LDA that is better suited for investigating topics in data sets with shorter units of text, such as websites and tweets. This approach to explore topics within short texts is called biterm topic modeling (BTM) and overcomes the problem of data sparsity by focusing on co-occurrences of word pairs across the whole data corpus (Cheng et al., 2014; Vidal et al., 2022). BTM has been used to explore topics related to consumer perceptions of vertical farming (Vidal et al., 2022) and turmeric (Feldmeyer & Johnson, 2022), and thus it may be invaluable for exploring sensory language and other information that can potentially influence the sensory experience.

In seeking to understand the language around food and the context around sensory experiences, we note that much research has focused on consumer-based data sets, such as blogs (Beninger et al., 2014), product reviews (L. M. Hamilton & Lahne, 2020), and open-ended responses (Danner & Menapace, 2020; Spinelli et al., 2017; ten Kleij & Musters, 2003). However, few text assessments have been conducted with producer-generated language. Food and beverage producers, and other stakeholders or retailers, add value to products by conveying information about how their products are made and where they are from (Lahne & Trubek, 2014; Paxson, 2013), whether that information is conveyed verbally or in written form. Text-based product information that producers use on their packaging or websites contributes to the context around food and beverages that consumers engage with prior to, during, and after their sensory experience (Bernard & Liu, 2017; Betancur et al., 2020; Kessinger et al., 2020; L. Lee et al., 2006; W. J. Lee et al., 2013).

1.1 | Cider

Cider is an alcoholic beverage made from fermented apple juice that is growing in popularity worldwide. In the United States (US), revenue from the cider industry has grown significantly over the past 10 years (Jacobsen, 2022; Peck & Knickerbocker, 2018; Wood, 2022). Smith et al. (2021) and Yenerall et al. (2022) have all documented various consumer preferences related to cider and the cider sensory experience. These researchers emphasize that information about sensory quality explicitly, as well as elements of sustainability, apple varieties, product origins, and other ethical values can make relevant contributions to consumers' affective and cognitive sensory experiences, further supporting evidence of the sensory experience being contextual (Betancur et al., 2020; Charters & Pettigrew, 2007; Kessinger et al., 2020; Lahne & Trubek, 2014).

The American Cider Association and other researchers have cited inconsistent and unclear marketing and sensory language as barriers to growth in the current American cider industry (Demmon, 2019; Fabien-Ouellet & Conner, 2018; Ostrom et al., 2022). Ostrom et al. (2022) describe cider producers' need for marketing support that increases consumer awareness of place-based qualities and cider-specific apples, echoing the research of Fabien-Ouellet and Conner (2018). Currently, style guidelines are a common tool for communicating cider quality in the US cider industry, with the two dominating styles being "modern" and "traditional." Ciders are also beginning to be described in terms of the explicit apple varieties that they contain, such as a "single varietal cider" made with only Newtown Pippin apples. Yet, there is little social or cultural consensus on how cider quality can or should be described, particularly related to clear and meaningful sensory communication. Therefore, the present research intends to explore how cider producers currently communicate and market cider products in order to gain a better understanding of how descriptive communication can be improved to best support the growing US cider industry.

1.2 | Objectives

The present study aimed to explore the language used by cider producers and marketers to explain their product(s) to consumers. We apply two complementary and easily accessible text-analysis workflows to this problem: frequency-based analyses (Silge & Robinson, 2017) and BTM (Cheng et al., 2014; Yan et al., 2013). We selected these methods because, while they are simple and accessible, they can nevertheless provide broad insight into the topics used to communicate cider products on cider websites, from Virginia, Vermont, and New York. These three states were selected because they are leading cider-producing states in the Northeast and Mid-Atlantic US (West, 2018). A secondary aim of this study will be to report on the effectiveness of the "tidy-text" approach and BTM in conjunction for small-scale, quick speech and language processing projects.

2 | MATERIALS AND METHODS

2.1 | Data collection

Data collection took place from August to December 2020. To determine which cideries to include in the dataset, researchers searched for all cider producers described by Cidercraft Magazine's cider locator (Locator, 2022) and West (2018) to operate in Virginia, Vermont, and New York. After compiling this initial list of cider producers, researchers searched each producer identity or cidery for a website. Cider producers or cideries that did not have an official website with information about their products were excluded from the dataset. From each cidery website, all individual cider product names were extracted along with all product descriptions on the cidery website. All text was directly copied from the cidery website into the dataset. If no information was provided across the website for an individual cider product, then the website description factor was set to "NA."

Using all information provided on the cidery websites, including individual product descriptions, other product attributes were extracted including: "Alcohol-By-Volume (ABV)," "Packaging Format," "Flavored," "Single v. Blend varietal," and "Apple Varieties." When listed on the cider website, "ABV" was extracted for every individual cider product. If a product did not have an explicitly stated ABV, the product factor was set to "NA." "Packaging Format" was made as a factor column to indicate the packaging format for every individual cider product. If the packaging format was not available from any information provided on the website, the factor value was set to "NA." If an individual cider product came in multiple packaging formats, then the factor value was set to "multiple." The factor column "Flavored" was set as "Yes" if a product description explicitly indicated that the cider was flavored with non-apple adjuncts. If the product description was clear that no adjuncts were used for flavoring or if flavoring could not be clearly inferred from the product description, then the factor was set as "No." "Single v. Blend varietal" was made as a factor column to indicate whether a cider is made with only one type of apple versus a blend of different apples, which are common production decisions in cider-making (Lea, 2008; Proulx & Nichols, 1980). If a cider product was explicitly described to be made with one apple variety, the "Single v. Blend" factor column was set as "Single" and the "Apple Varieties" factor column contained the explicitly mentioned apple variety. If a cider product was explicitly described to be made from a blend of different apples, then the appropriate factor column was set as "Blend" and all apple varieties were extracted and recorded in the "Apple Varieties" factor column. If the blend of apple varieties was not explicitly stated, then the "Apple Varieties" factor column was set as "NA." If a product's description had no mention of apple varieties or blending format, the "Single v. Blend" column was set as "NA." In some instances, where apple varieties were of an unknown identity described as "wild-foraged," "heirloom apples," "rare," or of similar language, the "Single v. Blend" factor was recorded as "Blend" and the "Apple Varieties" factor was recorded as "undesignated."

All data were manually retrieved (i.e., copy and paste) by first and second authors and stored in an electronic spreadsheet.

2.2 | Data analysis

All data were analyzed using R software version 4.1.2 (R Core Team, 2023) adapting procedures outlined in Silge and Robinson (2017), Wickham (2014), Cheng et al. (2014), Wijffels et al. (2021), and Yan et al. (2013). For data processing, the apple varieties described for each cider product were tokenized and variety frequencies were explored across state, apple variety use (i.e., single variety or blend), and packaging format. Website descriptions were filtered to remove English stop words, using the stop word list provided in the stopwords package for R. Website descriptions were tokenized and term frequency-inverse document frequency (tf-idf) was run across state, apple variety use, and packaging. Term frequency refers to the number of times individual terms occur in a document. tf-idf is a method of showcasing how important a given term is to a given document by minimizing words

that appear across multiple documents and emphasizing words that appear in select documents (Silge & Robinson, 2017).

BTM was adapted from Yan et al. (2013) and Wijffels et al. (2021) and inspired by the work of Vidal et al. (2022). BTM was selected to accommodate the smaller corpus of data as well as to accommodate the relatively small units of unstructured text (i.e., website descriptions) in the present study. The methods of BTM utilized in the present study involved assessing biterm co-occurrences within individual product descriptions (also known as "skipgrams") in the entire corpus, and within reviews from each of the three states individually, removing stop words, and filtering parts of speech to only include nouns, adjectives, and verbs in the topic models.

In order to train a BTM, specifying three hyper-parameters is necessary: the number of topics (*K*), alpha, and beta. The latter values, alpha and beta, control topic density and were set to recommended values (alpha = 50/*K*, beta = .01) based on Cheng et al. (2014). The BTM is trained to assume an underlying number of topics across a corpus, and the model will find distribution trends of words across these topics. Although there is no standard, pre-established way to determine the optimal number of topics for any given corpus, researchers have recommended training a BTM with different numbers of topics in order to discover which number best suites the data by yielding a higher subjective topic coherence (Cheng et al., 2014; Vidal et al., 2022). The present data set was trained using multiple biterm topic models for *K* (*K* = 1, 2, 3, 5, 7, 9) for the whole corpus of biterms, and two biterm topic models for *K* (*K* = 5, 7) for the corpus of biterms separated across the three states. First and last author read the words associated with each topic and qualitatively decided on the number of topics which best fit the data. Topics were deemed coherent when the researchers unanimously agreed that a clear topic or theme could be identified from the clusters with minimal repetition of non-related terms across topics. All topic models were visualized using cluster graphs following procedures described by Vidal et al. (2022) and Wijffels et al. (2021).

3 | RESULTS

The present data set, which we refer to as the "Cider Catalog" moving forward, comprises 1115 cider products, representing 15 cider producers from Vermont, 32 cider producers from Virginia, and 77 producers from New York (124 cider production companies total). In the following section, we describe results of our text mining, and we visualize trends across the data set according to factors including apple varieties, packaging, and state.

3.1 | Frequency-based results

3.1.1 | Apple varieties

To explore the use of named and unnamed apple varieties in ciders across the Cider Catalog, a word-frequency model was employed. The manually scraped apple varieties were treated as the tokens (units of analysis) in these analyses. In the Cider Catalog, Vermont had 6 single varietal ciders (representing 4.0% of all Vermont ciders), Virginia had 39 single varietal ciders (representing 12.6% of all Virginia ciders), and New York had 32 single varietal ciders (representing 4.9% of all New York ciders). Figure 1 showcases the top 15 most commonly used apple varieties across cider products from Virginia, Vermont and New York, indicating broad differences in the uses of individual apple varieties across the three states. Some apple varieties were more commonly used in certain states and regions, such as the McIntosh and Golden Russet apples appearing much more commonly in New York and Vermont than in Virginia. The Rhode Island Greening apple appeared in 19 cider products in New York versus 3 cider products in Vermont and 0 cider products in Virginia. As well, apple varieties that were planted historically in specific states were very common in that respective state. For example, the Newtown Pippin apple was used commonly in New York ciders, the Albemarle Pippin apple was used commonly in Virginia ciders, and the Dolgo Crab apple was used commonly in Vermont ciders (Calhoun, 2010; Proulx & Nichols, 1980; Pucci & Cavallo, 2021).

When exploring the use of different apple varieties in single varietal and blend ciders, the use of non-apple flavorings impacted the types of apple varieties that were used in cider products. In non-flavored single varietal ciders, the most common apple varieties were Northern Spy, Harrison, Albemarle Pippin, Gold Rush, Golden Russet, Arkansas Black, Winesap, Ashmead's Kernel, Baldwin, and Black Twig. In flavored single varietal ciders, Red Delicious, Granny Smith, and Pink Lady apple varieties were included in the list of commonly used apple varieties. This suggests that dessert or table apple varieties may be more common in flavored ciders overall, and that cider-specific apples may be more common in unflavored ciders. In cider products which were a blend of different apples ($n = 386$), apple variety names were not stated at all in 122 descriptions; but all descriptions for single varietal ciders ($n = 77$) explicitly stated which apple varieties were used in the product.

The most commonly used packaging formats described for cider products on the cider websites were 750 ml glass bottles ($n = 230$), 12 ounce cans ($n = 161$), 16 ounce cans ($n = 85$), and 500 ml bottles ($n = 27$). Many websites did not mention the packaging format of cider products ($n = 486$). The use of different apple varieties across canned and bottled packaging is shown in Figure 2. Bottled ciders often contained more cider-specific apple varieties, and explicitly stated apple varieties more often than canned ciders.

3.1.2 | Website description analyses

In order to explore unique communication patterns across website descriptions, tf-idf analyzes were conducted for all cider products. Figure 3 shows the results of high tf-idf terms across Virginia, Vermont, and New York. This visualization showcases how products made in different places are described in different ways, primarily with emphasis on the geography of origin: for example, "ny," "vermont," "vt," "burlington," "virginia," "shenandoah," "northeast," and

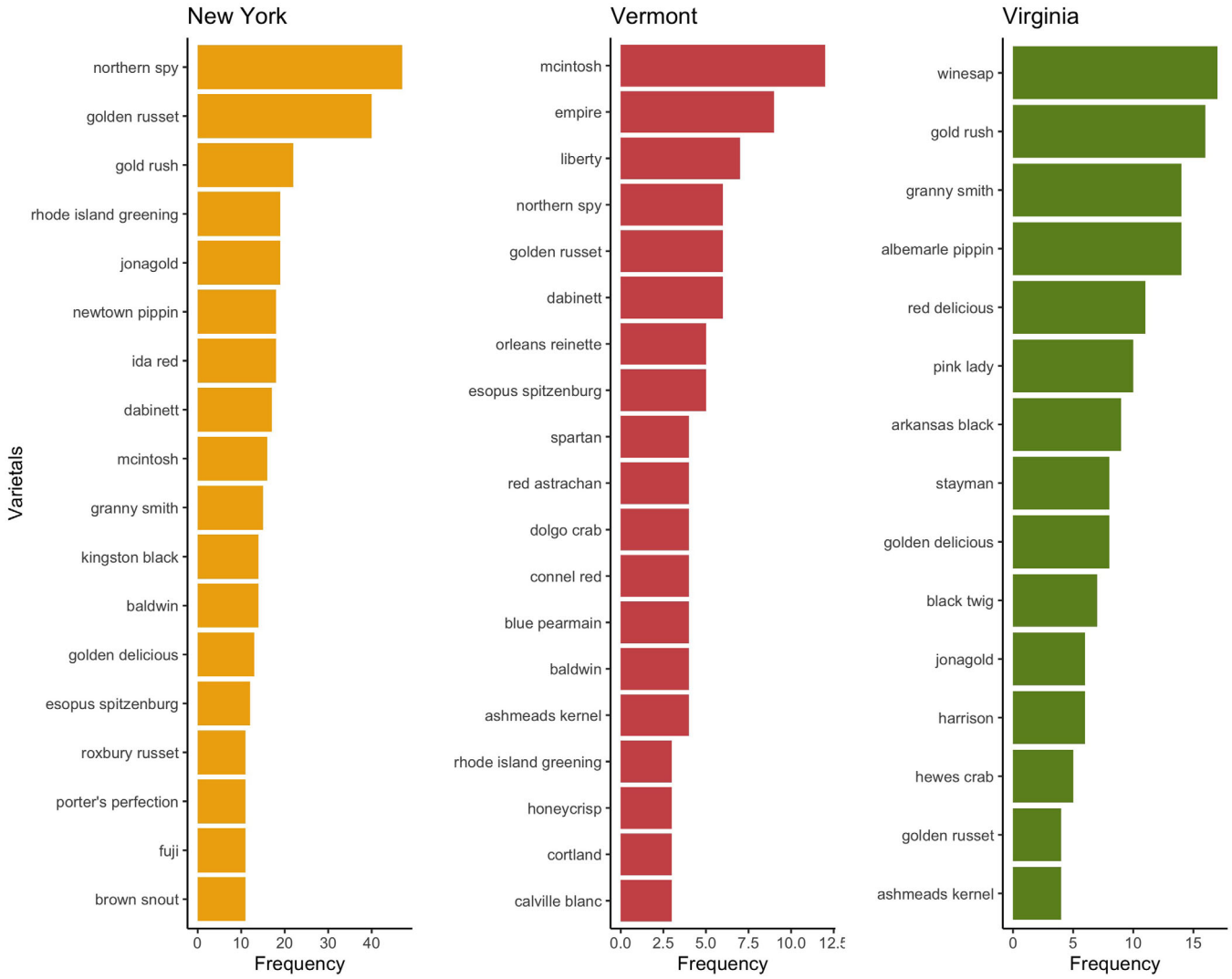

FIGURE 1 Specific apple varieties mentioned in website descriptions were counted across cider products from Virginia, Vermont, and New York. Apple varieties accounted for in the visualization must have been explicitly stated on cidery websites. Null values (NA) were removed from the visualizations to improve readability of the varietal frequencies. The proportion of website descriptions that did not explicitly mention apple varieties were 35.66% (435/1220) in New York, 47.98% (202/421) in Virginia, and 35.23% (99/281) in Vermont.

"champlain" are all high tf-idf terms for ciders from the respective state or region. All three states have high tf-idf terms indicative of the sensory experience including "palate," "aromas," and "sight." Virginia also has more high tf-idf terms that are potentially related to sensory quality such as "earthiness," "brightness," "mouthfeel," and "caramel." Virginia has "foods" has a high tf-idf term, suggesting that food pairing information is notably more common among Virginia cider descriptions and products. Virginia's and New York's set of high tf-idf terms includes multiple terms related to apple varieties; suggesting that this information is often conveyed in the descriptions of products from these two states. With New York product descriptions, high tf-idf terms related to apple varieties include "russet," "island," "newtown," "spy," "rhode," "greening," and "baldwin." With Virginia product descriptions, high tf-idf terms related to apple varieties include "albemarle," "harrison," "twig," and "winesap." Interestingly, one of Vermont's highest tf-idf terms is "manufactured," though this term is not common in New York or Virginia cider descriptions. Unique to Vermont, various high tf-idf terms are related to the name of a cider product itself (e.g., "arlo" and "americran" are two cider product names) or the producer (e.g., "windfall," "citizen," and "eden" are all Vermont cider brands). Finally, New York also has "ph," "ta," and "0.0" (presumably referring to pH, titratable acidity, and 0.0 g residual sugar, respectively) indicated as high tf-idf terms, suggesting that cider chemistry is a unique feature of website descriptions for products from this state.

An analysis of tf-idf terms across blended and single varietal ciders revealed that blended ciders contain many high tf-idf terms related to the cider-making process (i.e., "yeast," "time," "lees," "vintage," and "mix") and cider sensory quality (i.e. "aromas," "balance," and "structure"). Blended ciders also include "summer" as a high tf-idf term, indicating that these products may be marketed for seasonal or occasion-based consumption. However, no obvious

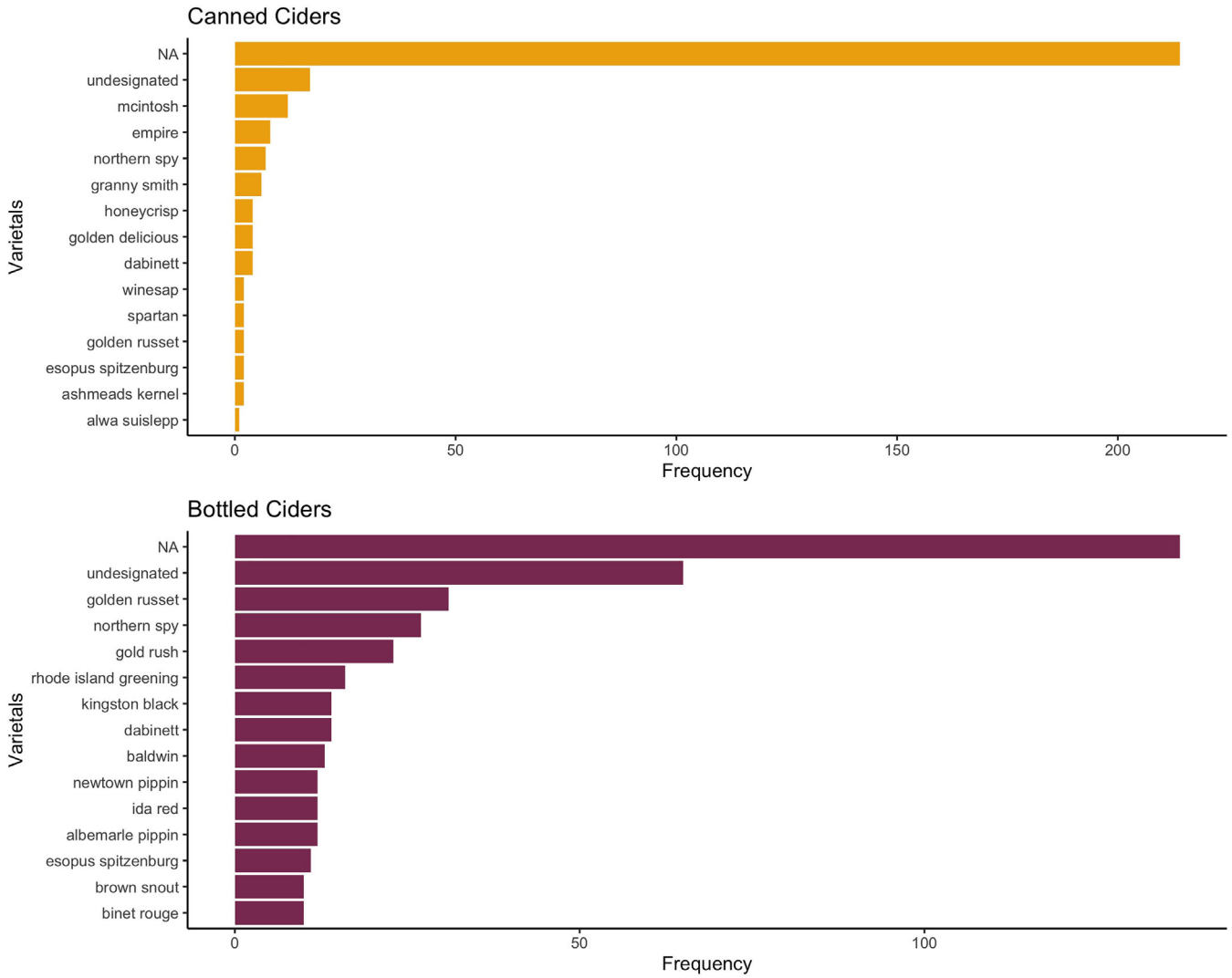

FIGURE 2 Specific apple varieties mentioned in website descriptions were counted across cider packaging formats. Apple varieties accounted for in the visualization must have been explicitly stated for cider products on the cider website, or the null value (NA) indicates that no apple varieties were indicated. Packaging formats accounted for in the visualization must have been explicitly stated on cidery websites. These results indicate that canned ciders very frequently do not indicate the apple varieties (NA) used, whereas bottled ciders are more likely to include mention of apple varieties in website descriptions.

themes were able to be identified from the high tf-idf terms consistent with single varietal ciders and visualizations of these analyses were deemed not critically meaningful by the researchers.

Figure 4 visualizes high tf-idf terms from the website descriptions across the two primary packaging formats (i.e., glass bottles and aluminum cans). Figure 4 indicates subtle yet important differences in the language used to describe the products across packaging formats. Ciders in cans are distinguished by website descriptions without a clear overarching topic, though multiple high tf-idf terms appear to refer to nutrition-related information such as "carb" and "cal." As well, the terms "3g" and "6g" potentially refer to residual sugar content and the terms "140" and "155" potentially refer to calorie contents (see Figure 4). In addition, tf-idf analyses were calculated across 12 ounce cans, 16 ounce cans, 500 ml bottles, and 750 ml bottles (results not shown). Twelve ounce cans in particular had high tf-idf terms such as "fermented," "pressed," and "concentrate," and the term "tart," as an unique sensory descriptor. Ciders in bottles were distinguished by website descriptions with terms related to apple varieties, chemistry, and cider production methods. For example, the terms "perfection," "russet," "newtown," "esopus," and "goldrush" are suggestive of apple varieties and the terms "disgorged" and "champenoise" are related production methods (see Figure 4).

Network bigram visualization

To explore words which commonly occurred together throughout the entire dataset, tokenization of website descriptors was performed using bigrams and a cluster network graph was visualized in Figure 5. This visualization was not segmented based on extrinsic product factors in order to maintain simplicity, given that bigrams are sparser than unigrams (Jurafsky & Martin, 2021), and to gain an overall glimpse of terms which are commonly used together. The biterm cluster related to residual sugar indicates a trend of website descriptions emphasizing sugar

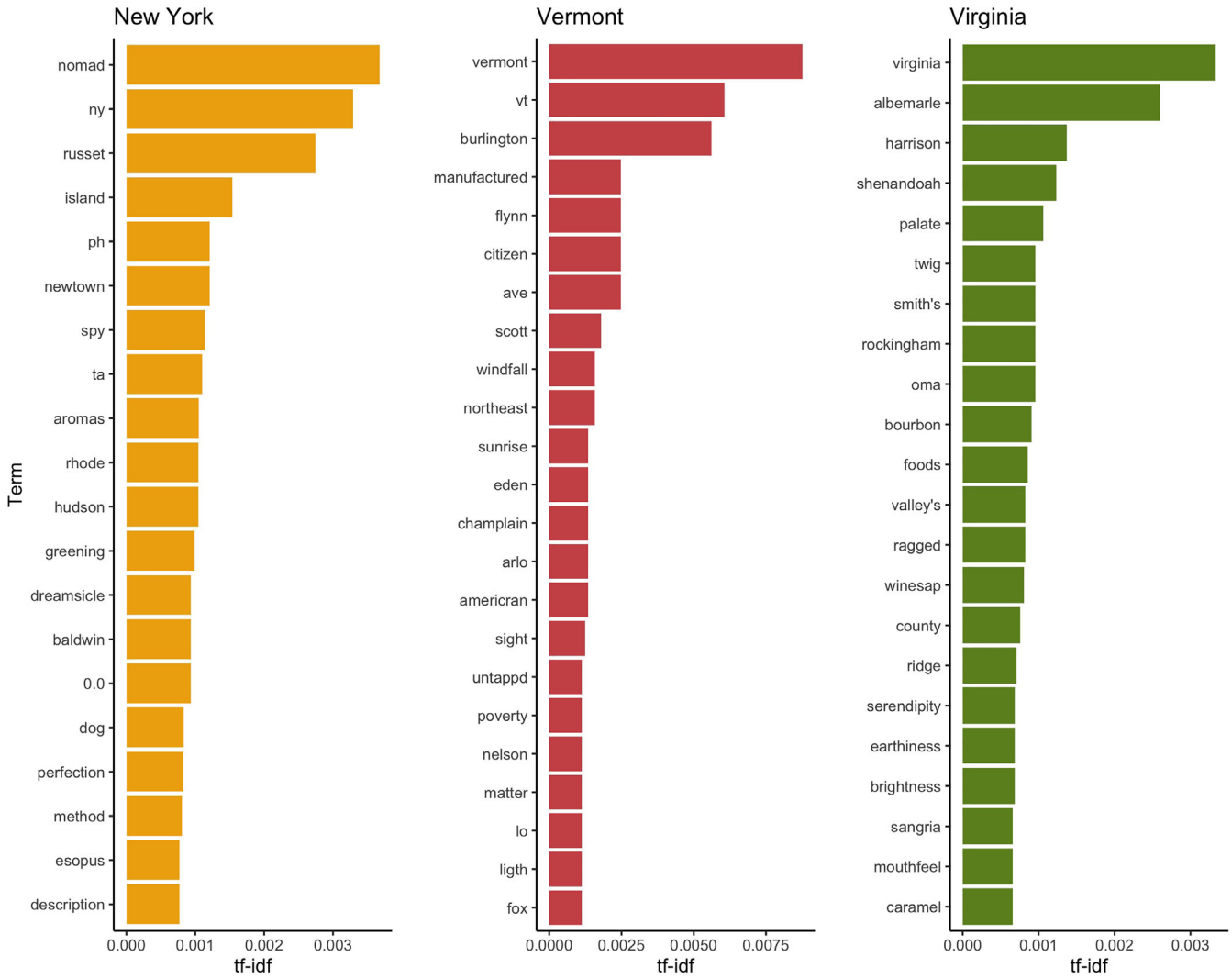

FIGURE 3 Term frequency-inverse document frequency (tf-idf) of terms from website descriptions, compared across Virginia, Vermont, and New York, shows that all three states have important terms related to proper nouns and geography. More sensory and food-related terms are unique to cider descriptions for products from Virginia, whereas the mention of chemical parameters (e.g., "ph" and "ta") are unique to New York cider product descriptions. Both New York and Virginia cider products have more unique terms related to apple varieties.

content, among other attributes, in website descriptions. Many biterms refer to apple varieties, such as "granny smith" and "kingston black," and places, such as "blue ridge" and "finger lakes." Some biterms are suggestive of production-related language, such as "bottle conditioned" and "6 months," and "secondary fermentation," and others are suggestive of sensory terminology, such as "soft tannins" and "fruit forward." Interestingly, there is a bigram for "taste process," suggesting that many producers use website descriptions to share information about the connections between the cider-making process and sensory attributes.

3.2 | BTM results

BTM is an unsupervised learning method that finds "topics" in texts by identifying terms that occur in the same small context window more often than would be expected by chance. In order to explore topics across the text used in website descriptions, different "document corpora" were selected for fitting the BTMs: first, BTMs with various numbers of topics were fit to the entire cider catalog; second, separate BTMs were fit for each state for various numbers of topics. For each BTM, topics were manually labeled following agreement from the researchers. A topic was deemed coherent when the researchers agreed on topic identity based on the relevant biterms and deemed incoherent when the researchers could not identify a consistent theme according to the biterms. For all models, only coherent topics are labeled.

Figure 6 shows the seven-topic model of biterms across the whole corpus. In this figure, the five coherent topics were labeled ***Food Pairing***, ***Apple Varieties***, ***Production***, ***Sensory***, and ***Flavorings*** as relevant themes emerging from website descriptions of cider products. The topic ***Sensory*** was the most stable topic in the corpus, originally appearing as a subjectively distinguishable cluster in the three-

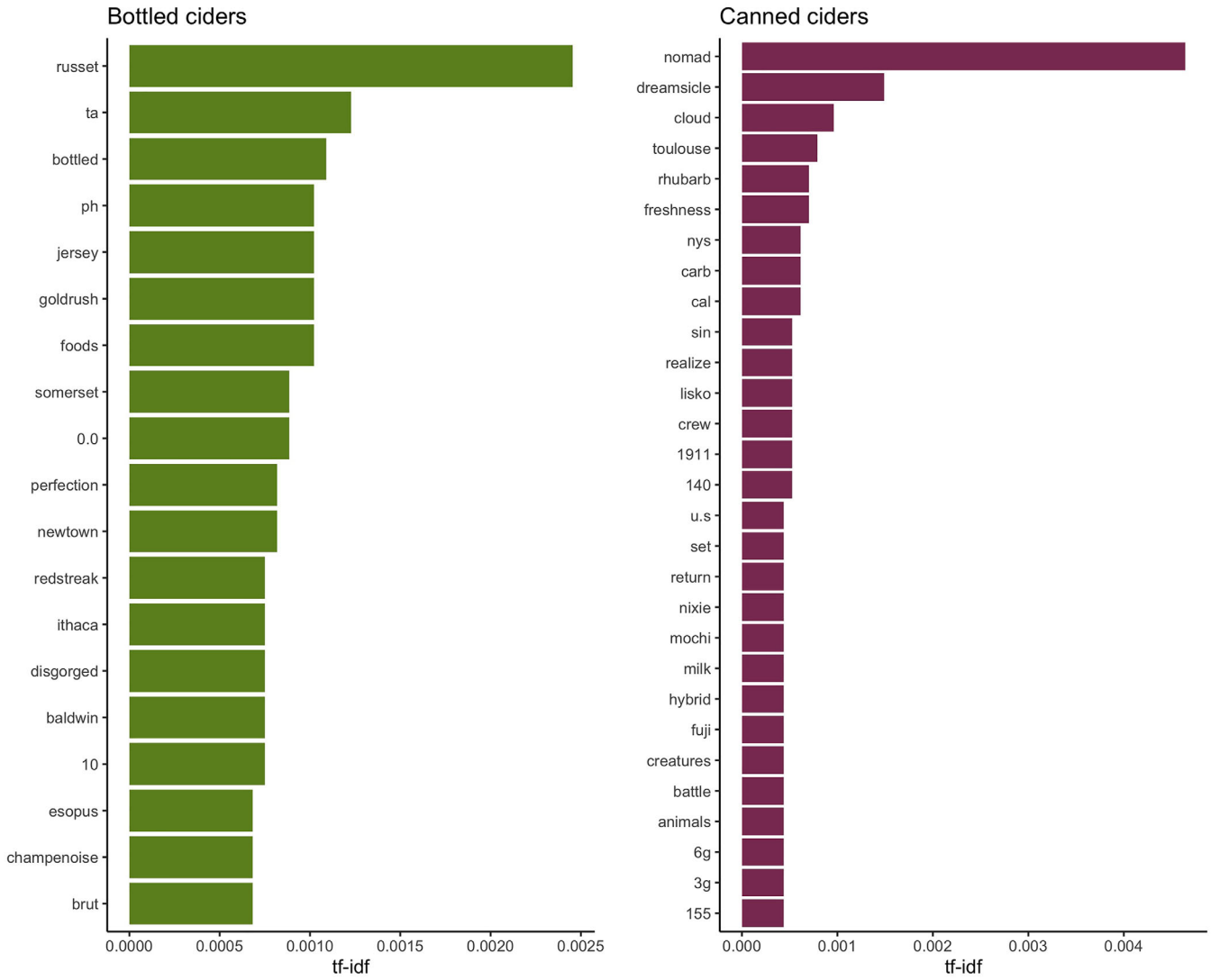

FIGURE 4 Term frequency-inverse document frequency (tf-idf) of terms from website descriptions were compared across the two most common packaging formats, cans and bottles revealing that bottled ciders tend to emphasize terms clearly related to apple varieties, chemistry, and fermentation processes. Packaging formats accounted for in the visualization must have been explicitly indicated on the cidery websites.

topic model. When the topic model was expanded to the nine-topic model, the ***Flavorings*** cluster separated into three clusters each containing different flavoring-related biterms in addition to inconsistent terms. For example, a cluster containing the term "cranberry" also contained the terms "sweet" and "refresh," whereas another cluster containing the term "hop" also contained the words "valley" and "local." This cluster was named to refer to non-apple adjunct flavorings which are common in the US cider industry (Alexander & Ewing Valliere, 2020).

In Figure 7, the five-topic model for cider website descriptions from each of the three states showed the most coherent and non-repetitive topic clusters. Only coherent topics were labeled, including three ***Sensory*** clusters, three ***Production*** clusters, two ***Food-Pairing*** clusters, and one ***Apple Varieties*** cluster. Cider descriptions from New York were unique in their biterm clustering of words related to apple varieties, including the terms "northern," "spy," "golden," and "russet" (see Figure 7a), which refer to two apple varieties commonly used for cider-making: Northern Spy and Golden Russet (Proulx & Nichols, 1980). Vermont ciders had three clusters (red, yellow, and blue clusters) inconsistently including sensory biterms, although the yellow cluster included the majority of sensory-related terms (see Figure 7c). For example, Figure 7c shows the blue cluster including the terms "sweet" and "fresh" and the red cluster including the term "flavor." This suggests that Vermont cider products may use sensory-related language in a way that is unclear on their website descriptions. We elaborate on the inconsistent occurrence of these and other terms in the following section.

4 | DISCUSSION AND CONCLUSIONS

The present study aimed to explore the language used by cider producers to explain and market their products via website descriptions for cider products in three primary cider-producing states in the

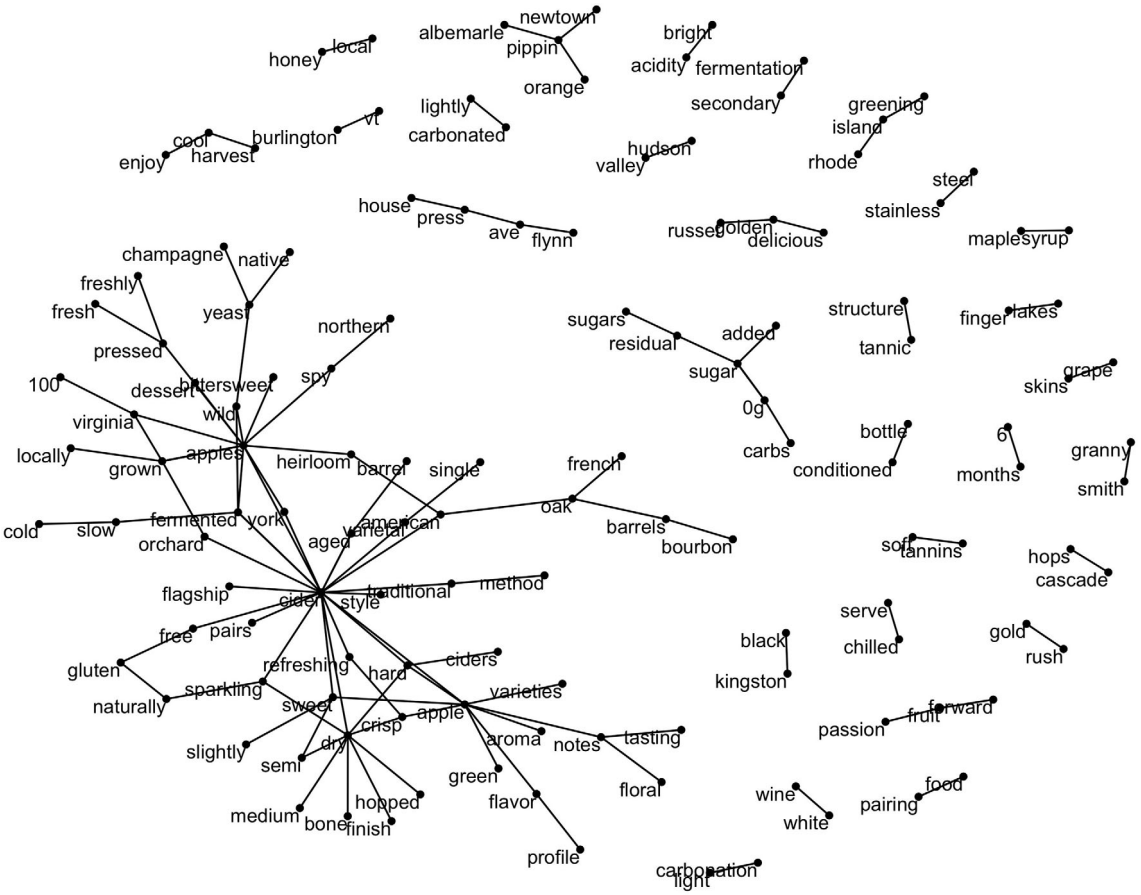

FIGURE 5 A visualized network of the most frequently occurring bigrams from the cider website descriptions using connected nodes. This visualization shows common bigrams in cider website descriptions primarily related to apple varieties, sensory quality, production processes, and places. This visualization showcases some of the overarching topics across website descriptions.

Northeast and Mid-Atlantic US. This study objective was accomplished using two text-mining approaches: frequency-based and BTM workflows which were selected for their accessibility and applicability to small-scale language projects. In general, text frequency, tf-idf, and BTM produced meaningful insights regarding the patterns of communication and marketing on cider website descriptions. The BTM approach was particularly effective at reporting coherent and consistent themes across website descriptions with some limitations, and both text mining approaches were valuable for uncovering patterns related to descriptive sensory language. As we demonstrate in detail below, these relatively simple text-mining approaches can be implemented with low cost and overhead by sensory scientists in a number of circumstances. For example, as was the case in this study, a text-mining approach can be extremely useful in getting a "first look" at a category's possible sensory attributes when a category has relatively little information published about it. Text mining is also extremely valuable as an alternative or additional method for exploring how sensory language is used in vivo: what terms and terminologies are important for product users and consumers outside of the laboratory.

The first portion of the text mining procedures involved simple frequency analyses of apple varieties used and discussed in products from Virginia, Vermont, and New York, as well as across different packaging formats. Frequency analyses revealed that different apple varieties are used with different frequencies across the three states (see Figure 1). The observations of certain apple varieties being more common in certain states aligns with horticulture research regarding how certain apple varieties are more prevalent and successful in certain regions (Miles et al., 2020; Ostrom et al., 2022; Peck & Knickerbocker, 2018). Because apples can be unique to a place, and because cider products can have significant sensory differences as an effect of apple varieties and place (Calvert, Neill, et al., 2022; Littleson et al., 2022), it may be relevant for cider producers and marketers to clearly communicate the apple varieties used in their products. Apple varieties were also a coherent cluster predicted using BTM across all website descriptions (see Figure 7). However, apple varieties were only a coherent cluster for New York cider descriptions, not for Virginia and Vermont ciders (see Figure 7). Recent research has also described consumers' increasing interest in local and regionally sourced apples and historically significant apple varieties (Smith et al., 2021; Yenerall et al., 2022) and so it may be valuable for cider producers, particularly those in Virginia and Vermont, to emphasize their use of regionally distinct apple varieties when applicable.

The present research also reveals broad findings related to the language used to describe cider products across Virginia, Vermont,

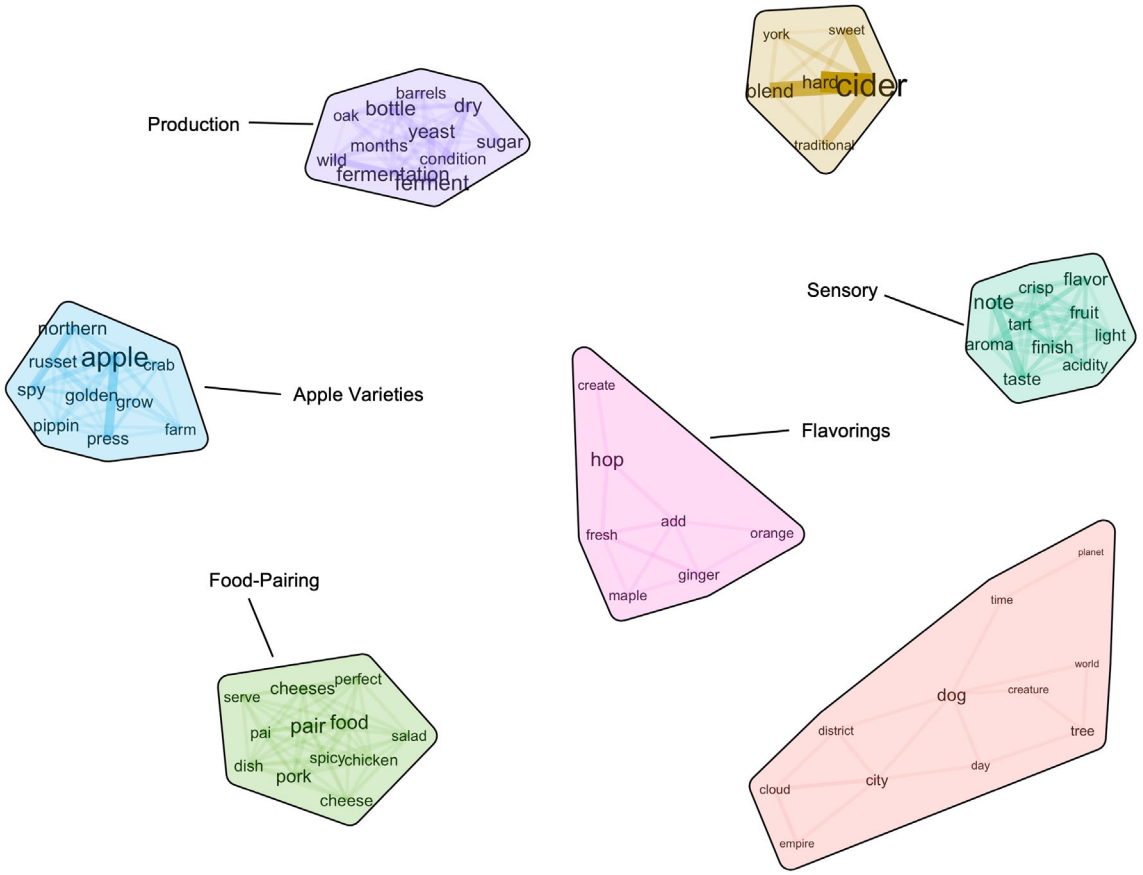

FIGURE 6 A seven-topic ($k = 7$) biterm topic model of topics appearing across the whole corpus of data. Only coherent topics were labeled manually by the researchers. Inside of each topic cluster, the line width indicates how commonly two terms were associated with each other as a biterm. For example, "wild-fermentation" and "blend-cider" were terms commonly associated with each other. The orange, unlabeled cluster represents one incoherent topic cluster which is normal for biterm topic modeling (BTM) (Wijffels et al., 2021).

and New York overall. Cider product descriptions from all three states include emphasis on sensory attributes, production information, and geography based on findings from both the frequency-based text mining and BTM. Results of the single-word tf-idf approach suggest that Virginia cider products place more emphasis on food, New York cider products place more emphasis on chemistry and apple varieties, and Vermont cider products place more emphasis on proper nouns (i.e., locations, product names, and brand names). BTM revealed that food pairing was in fact a consistent topic among both Virginia and New York products. Food-pairing information has been historically and culturally common for wine, both on websites and on product labels (Kelley et al., 2015; Mueller et al., 2010; Pettigrew & Charters, 2006; Spence, 2020b), but is less common within the product experience for beer (Martinez et al., 2017; Pettigrew & Charters, 2006). As well, in the case of wine, food-pairing information has been linked with higher consumer valuation (Kelley et al., 2015; Pickering et al., 2022). Although food-pairing is becoming more popular in the craft beer industry (Arellano-Covarrubias et al., 2022; Pettigrew & Charters, 2006; Spence, 2020b; Stoller, 2019), our results of food-pairing information being a dominant topic in cider descriptions may seem to suggest that cider producers are using food-pairing information as a way to analog cider closer to wine in terms of product expectations and valuation. The same predictions may also apply to the use of apple varieties as a common topic among bottled (see Figure 4) and New York products, which may be an artifact of regional beverage and packaging trends given that New York is home to 11 American Viticultural Areas and was ranked third as the most wine producing state in the United States in 2020 (Conway, 2022).

It is worth discussing that Figures 5 and 6 also reveal interesting trends related to cider production as evidenced by the biterms "secondary fermentation," "bottle conditioned," "native yeast," "wild yeast," and "wild fermentation." These biterms suggest that the mention of wild fermentation methods is becoming common in product descriptions, presumably as a way to differentiate products. Wild fermentation methods are also trending in wine-making and brewing (Dillon et al., 2013; Molinet & Cubillos, 2020), though the details of these processes are not well-understood nor is the use of process-oriented terms regulated by industry stakeholders or governmental organizations. To reemphasize, the "taste-process" biterm in Figure 5 suggests that cider producers use cider website descriptions to draw connections between production methods and intrinsic sensory attributes. These findings are consistent with consumer research on cider

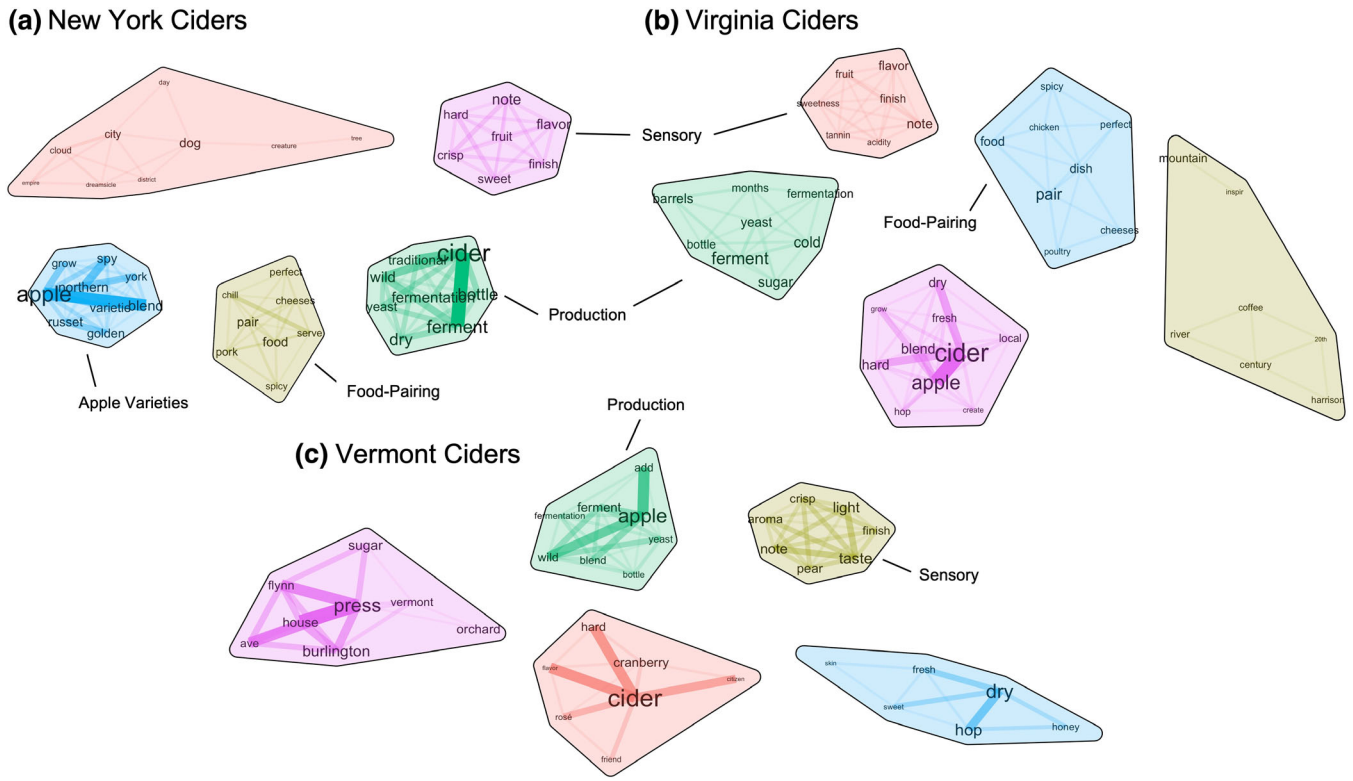

FIGURE 7 Three, five-topic biterm topic models of topics appearing per state, across the whole corpus of data. Topics related to ***Production*** and ***Sensory*** were consistent and coherent across all three states. New York (a) uniquely had a topic cluster related to ***Apple Varieties***, confirming the results of term frequency-inverse document frequency (tf-idf) seen in previous visualizations. Vermont products (c) did not have a coherent ***Food-Pairing*** cluster but did have a cluster including specific places (i.e., "vermont" and "burlington") and two clusters including flavoring-related terms (i.e., "cranberry" and "dry-hop").

and other craft products demonstrating how consumers are interested in cognitive, reflective sensory experiences that are supplemented with cider-making information (Calvert, Neill, et al., 2022; Gómez-Corona et al., 2016, 2017; Lahne & Trubek, 2014; Paxson, 2013).

Sensory terms were a meaningful feature of the present research. Singular sensory terms were minimally showcased by tf-idf analyses across each state in Figure 3 but were the most stable biterm topic predicted with the BTM across the whole data corpus (see Figure 7). This indicates that while few specific sensory terms are unique to cider products from specific places, cider website product descriptions consistently use sensory language. However, the sensory terms shown in Figure 6 and Figure 7 (from the BTM) include fairly simple descriptive terms (e.g., "fruit") though sensory research on cider has demonstrated an abundance of more detailed sensory terms, like those that were unique to Virginia product descriptions (see Figure 3; Calvert, Neill, et al., 2022; Cole et al., 2022; Littleson et al., 2022). These findings suggest that the sensory language of cider marketing is not emphatically descriptive, which may align with discussion of cider descriptive language being unclear (Demmon, 2019; Fabien-Ouellet & Conner, 2018).

In addition, the sensory terms "sweet" and "dry" both appear in clusters separate from the distinctly labeled ***Sensory*** topic in Figure 6. This may seem to suggest that ciders described as "dry" are also commonly described according to how they are made, whereas ciders described as "sweet" do not have any other clearly consistent descriptive attributes. Interestingly, traditional style cider is often described as less "sweet" than modern style cider (Alexander & Ewing Valliere, 2020; Calvert, Neill, et al., 2022)—so, the co-occurrence of "sweet" and "traditional" in the yellow cluster of Figure 6 reaffirms a lack of clarity regarding the sweetness and dryness of cider products in sensory communication (Calvert, Cole, et al., 2022; Kessinger et al., 2020; Phetxumphou et al., 2020). Sensory quality is a driving factor of consumer preference and willingness to pay for American cider products (Cole et al., 2022; Kessinger et al., 2020; Tozer et al., 2015), therefore it is crucial that sensory information be clear and meaningful so that consumers can know what to expect from a cider product before purchasing and consuming.

Although the present study was able to effectively use text mining to explore common terms and topics used to market cider products in an online space, this research has broad limitations. Most importantly, websites are not the primary means of conveying product information between producers and consumers. In addition, this research only uses website descriptions from Virginia, Vermont, and New York as a snapshot of cider product descriptions in the United States. Another limitation of the present research is that all analyses conducted using frequency-based text mining and the BTM

require data processing decisions that are unique to the research team and heavily influential on the present findings (Jaeger & Rasmussen, 2021; Vidal et al., 2022). For example, the pre-processing steps taken in the present research included lemmatization, removal of common stop words, and selection of specific parts of speech (i.e., nouns, adjectives, and verbs), which were done to reduce data sparsity and allow for more coherent topic formation from the BTM. With the frequency-based text mining methods, procedures were closely adapted from Silge and Robinson (2017) though wrangling techniques unique to the data-set and chosen by the researchers were also necessary (i.e., filtering out "NA" when not relevant to the scope of the analysis in question). Other data manipulation and wrangling steps could have been conducted to deepen or change the information shown in the data visualizations, including running tf-idf across all packaging formats or running more general frequency analyses. Nonetheless, both the tidy-text mining and BTM approaches were valuable for understanding broad themes and patterns across cider website descriptions for products from Virginia, Vermont, and New York.

More broadly, the present research offers a framework for the use of frequency-based and BTM approaches for exploring sensory language in the context of marketing and product descriptions. With the hard cider industry as a case study, sensory scientists in research and specialized industries can use this research as a framework for quickly and efficiently exploring how sensory descriptive language is used across food and beverage products—results that can then be used to support other sensory testing protocols and quality assurance programs. For example, these text mining tools could be used to rapidly generate sensory lexicons that could then be used as a starting point for panel training or rapid descriptive profiling. Future work should be conducted to show how the described text mining tools can be adapted to more thoroughly extract sensory terms and can be used in conjunction with or to support panel training, lexicon development, and rapid descriptive profiling. With BTM and frequency-based approaches, text mining is a multifaceted tool that can both support sensory research and provide meaningful product-specific insights.

FUNDING INFORMATION

This research was funded through USDA-NIFA AFRI award #2020-68006-31682.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

Alexander, T. R., & Ewing Valliere, B. L. (2020). *The professional handbook of cider tasting* (1st ed.). Oxfordshire.

An, J. (2022). Examining market-driving innovation in the wine industry: A topic modeling method on consumer reviews on wine documentaries. *Wine Business Journal*, *5*(1), 76–87. https://doi.org/10.26813/001c.31306

Arellano-Covarrubias, A., Escalona-Buendía, H. B., Gómez-Corona, C., & Varela, P. (2022). Pairing beer and food in social media: Is it an image worth more than a thousand words? *International Journal of Gastronomy and Food Science*, *27*, 100483. https://doi.org/10.1016/j.ijgfs.2022.100483

Beninger, S., Parent, M., Pitt, L., & Chan, A. (2014). A content analysis of influential wine blogs. *International Journal of Wine Business Research*, *26*(3), 168–187. https://doi.org/10.1108/IJWBR-09-2013-0036

Bernard, J. C., & Liu, Y. (2017). Are beliefs stronger than taste? A field experiment on organic and local apples. *Food Quality and Preference*, *61*, 55–62. https://doi.org/10.1016/j.foodqual.2017.05.005

Betancur, M. I., Motoki, K., Spence, C., & Velasco, C. (2020). Factors influencing the choice of beer: A review. *Food Research International*, *137*, 109367. https://doi.org/10.1016/j.foodres.2020.109367

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In *Text mining*. Chapman and Hall.

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

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27–41.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. SAGE.

Calhoun, C. L., Jr. (2010). *Old southern apples: A comprehensive history and description of varieties for collectors, growers, and fruit enthusiasts* (2nd ed.). Chelsea Green Publishing.

Calvert, M. D., Cole, E., Stewart, A. C., Neill, C. L., & Lahne, J. (2022). Can cider chemistry predict sensory dryness? Benchmarking the Merlyn dryness scale. *Journal of the American Society of Brewing Chemists*, 1–6. https://doi.org/10.1080/03610470.2022.2121562

Calvert, M. D., Neill, C. L., Stewart, A. C., & Lahne, J. (2022). Sensory descriptive analysis of hard ciders from the Northeast and Mid-Atlantic United States. *Journal of Food Science*, *88*, 1700–1717. https://doi.org/10.1111/1750-3841.16507

Charters, S., & Pettigrew, S. (2007). The dimensions of wine quality. *Food Quality and Preference*, *18*(7), 997–1007. https://doi.org/10.1016/j.foodqual.2007.04.003

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 2928–2941. https://doi.org/10.1109/TKDE.2014.2313872

Cole, E., Stewart, A. C., Chang, E. A. B., & Lahne, J. (2022). Exploring the sensory characteristics of Virginia ciders through descriptive analysis and external preference mapping. *Journal of the American Society of Brewing Chemists*, 1–13. https://doi.org/10.1080/03610470.2022.2119535

Conway, J. (2022). Wine production in the U.S. 2020, by state. Statista, Inc. Industry report.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design* (4th ed.). SAGE Publications.

Danner, H., & Menapace, L. (2020). Using online comments to explore consumer beliefs regarding organic food in German-speaking countries and the United States. *Food Quality and Preference*, *83*, 103912. https://doi.org/10.1016/j.foodqual.2020.103912

Demmon, B. (2019). Crafting the future of the American cider industry. SevenFifty Daily. Retrieved from https://daily.sevenfifty.com/crafting-the-future-of-the-american-cider-industry/

Dillon, S., Jiranek, V., & Grbin, P. (2013). Wild yeast fermentation can allow chemical and sensory differentiation in red and white wines. *Wine & Viticulture Journal*, *28*(6), 23–25.

Doyle, J. D., Heslop, L. A., Ramirez, A., Cray, D., & Armenakyan, A. (2012). Trust building in wine blogs: A content analysis. *International Journal of*

Wine Business Research, *24*(3), 196–218. https://doi.org/10.1108/17511061211259198

Fabien-Ouellet, N., & Conner, D. (2018). The identity crisis of hard cider. *Journal of Food Research*, *7*(2), 54. https://doi.org/10.5539/jfr.v7n2p54

Feldmeyer, A., & Johnson, A. (2022). Using Twitter to model consumer perception and product development opportunities: A use case with turmeric. *Food Quality and Preference*, *98*, 104499. https://doi.org/10.1016/j.foodqual.2021.104499

Gómez-Corona, C., Escalona-Buendía, H. B., Chollet, S., & Valentin, D. (2017). The building blocks of drinking experience across men and women: A case study with craft and industrial beers. *Appetite*, *116*, 345–356. https://doi.org/10.1016/j.appet.2017.05.026

Gómez-Corona, C., Escalona-Buendía, H. B., García, M., Chollet, S., & Valentin, D. (2016). Craft vs. industrial: Habits, attitudes and motivations towards beer consumption in Mexico. *Appetite*, *96*, 358–367. https://doi.org/10.1016/j.appet.2015.10.002

Hamilton, L. (2022). *Translating sensory perceptions: Existing and emerging methods of collecting and analyzing flavor data* [doctoral dissertation]. Virginia Tech.

Hamilton, L. M., & Lahne, J. (2020). Fast and automated sensory analysis: Using natural language processing for descriptive lexicon development. *Food Quality and Preference*, *83*, 103926. https://doi.org/10.1016/j.foodqual.2020.103926

Hamilton, L. M., & Lahne, J. (2023). Natural language processing. In J. Delarue & J. B. Lawlor (Eds.), *Rapid sensory profiling techniques* (pp. 371–410). Woodhead Publishing.

Jacobsen, J. (2022). 2022 beer report: Hard cider market takes on competition. Retrieved from https://www.bevindustry.com/articles/94801-beer-report-hard-cider-market-takes-on-competition

Jaeger, S. R., & Rasmussen, M. A. (2021). Importance of data preparation when analysing written responses to open-ended questions: An empirical assessment and comparison with manual coding. *Food Quality and Preference*, *93*, 104270. https://doi.org/10.1016/j.foodqual.2021.104270

Jurafsky, D. (2014). *The language of food*. W. W. Norton & Company.

Jurafsky, D., & Martin, J. H. (2021). Speech and language processing (3 (draft)). Retrieved from https://web.stanford.edu/~jurafsky/slp3/

Kelley, K., Hyde, J., & Bruwer, J. (2015). U.S. wine consumer preferences for bottle characteristics, back label extrinsic cues and wine composition: A conjoint analysis. *Asia Pacific Journal of Marketing and Logistics*, *27*(4), 516–534. https://doi.org/10.1108/APJML-09-2014-0140

Kessinger, J., Earnhart, G., Hamilton, L., Phetxumphou, K., Neill, C., Stewart, A. C., & Lahne, J. (2020). Exploring perceptions and categorization of Virginia hard ciders through the application of sorting tasks. *Journal of the American Society of Brewing Chemists*, *79*(2), 1–14. https://doi.org/10.1080/03610470.2020.1843927

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology*. SAGE Publications. https://doi.org/10.4135/9781071878781

Lahne, J., & Trubek, A. B. (2014). "A little information excites us." Consumer sensory experience of Vermont artisan cheese as active practice. *Appetite*, *78*, 129–138. https://doi.org/10.1016/j.appet.2014.03.022

Lawless, H. T., & Heymann, H. (2010a). Context effects and biases in sensory judgment. In H. T. Lawless & H. Heymann (Eds.), *Sensory evaluation of food: Principles and practices* (pp. 203–225). Springer. https://doi.org/10.1007/978-1-4419-6488-5_9

Lawless, H. T., & Heymann, H. (2010b). Introduction. In H. T. Lawless & H. Heymann (Eds.), *Sensory evaluation of food: Principles and practices* (pp. 1–18). Springer. https://doi.org/10.1007/978-1-4419-6488-5_1

Lea, A. (2008). *Craft cider making* (3rd ed.). GoodLife Press.

Lee, L., Frederick, S., & Ariely, D. (2006). Try it, you'll like it: The influence of expectation, consumption, and revelation on preferences for beer. *Psychological Science*, *17*(12), 1054–1058. https://doi.org/10.1111/j.1467-9280.2006.01829.x

Lee, W. J., Shimizu, M., Kniffin, K. M., & Wansink, B. (2013). You taste what you see: Do organic labels bias taste perceptions? *Food Quality and Preference*, *29*(1), 33–39. https://doi.org/10.1016/j.foodqual.2013.01.010

Littleson, B., Chang, E., Neill, C., Phetxumphou, K., Sandbrook, A., Stewart, A., & Lahne, J. (2022). Sensory and chemical properties of Virginia hard cider: Effects of apple cultivar selection and fermentation strategy. *Journal of the American Society of Brewing Chemists*, *0*, 1–14. https://doi.org/10.1080/03610470.2022.2057780

Locator. (2022). CIDERCRAFT magazine. Retrieved from https://cidercraftmag.com/locator/

Martinez, D. C., Hammond, R. K., Harrington, R. J., & Wiersma-Mosley, J. D. (2017). Young adults' and industry experts' subjective and objective knowledge of beer and food pairings. *Journal of Culinary Science & Technology*, *15*(4), 285–305. https://doi.org/10.1080/15428052.2016.1256243

Meiselman, H. L. (2019). *Context: The effects of environment on product design and evaluation*. Woodhead Publishing.

Miles, C. A., Alexander, T. R., Peck, G., Galinato, S. P., Gottschalk, C., & van Nocker, S. (2020). Growing apples for hard cider production in the United States—Trends and research opportunities. *HortTechnology*, *30*(2), 148–155. https://doi.org/10.21273/HORTTECH04488-19

Molinet, J., & Cubillos, F. A. (2020). Wild yeast for the future: Exploring the use of wild strains for wine and beer fermentation. *Frontiers in Genetics*, *11*, 1–8. https://doi.org/10.3389/fgene.2020.589350

Mueller, S., Lockshin, L., Saltman, Y., & Blanford, J. (2010). Message on a bottle: The relative influence of wine back label information on wine choice. *Food Quality and Preference*, *21*(1), 22–32. https://doi.org/10.1016/j.foodqual.2009.07.004

Ostrom, M. R., Conner, D. S., Tambet, H., Smith, K. S., Sirrine, J. R., Howard, P. H., & Miller, M. (2022). Apple grower research and extension needs for craft cider. *HortTechnology*, *32*(2), 147–157. https://doi.org/10.21273/HORTTECH04827-21

Parys, N. (2013). Cooking up a culinary identity for Belgium. Gastrolinguistics in two Belgian cookbooks (19th century). *Appetite*, *71*, 218–231. https://doi.org/10.1016/j.appet.2013.08.006

Paxson, H. (2013). *The life of cheese*. University of California Press.

Peck, G., & Knickerbocker, W. (2018). Economic case studies of cider apple orchards in New York state. *Fruit Quarterly*, *26*(3), 6.

Peschel, A. O., Kazemi, S., Liebichová, M., Sarraf, S. C. M., & Aschemann-Witzel, J. (2019). Consumers' associative networks of plant-based food product communications. *Food Quality and Preference*, *75*, 145–156. https://doi.org/10.1016/j.foodqual.2019.02.015

Pettigrew, S., & Charters, S. (2006). Consumers' expectations of food and alcohol pairing. *British Food Journal*, *108*(3), 169–180. https://doi.org/10.1108/00070700610650990

Phetxumphou, K., Cox, A. N., & Lahne, J. (2020). Development and characterization of a check-all-that-apply (CATA) lexicon for Virginia hard (alcoholic) ciders. *Journal of the American Society of Brewing Chemists*, *78*(4), 299–307. https://doi.org/10.1080/03610470.2020.1768784

Pickering, G. J., Duben, M., & Kemp, B. (2022). The importance of informational components of sparkling wine labels varies with key consumer characteristics. *Beverages*, *8*(2), 27. https://doi.org/10.3390/beverages8020027

Proulx, A., & Nichols, L. (1980). *Cider: Making, using, & enjoying sweet & hard cider* (3rd ed.). Garden Way Publishing.

Pucci, D., & Cavallo, C. (2021). *American cider*. Ballantine Books.

R Core Team. (2023). R: A language and environment for statistical computing. (4.2.3). http://www.R-project.org/

Riley, K. C., & Paugh, A. L. (2018). *Food and language: Discourses and foodways across cultures* (1st ed.). Routledge. https://doi.org/10.4324/9781315695235

Sadler, C. R., Grassby, T., Hart, K., Raats, M. M., Sokolović, M., & Timotijevic, L. (2022). "Even We Are Confused": A Thematic Analysis of Professionals' Perceptions of Processed Foods and Challenges for

Communication. *Frontiers in Nutrition*, 9. https://www.frontiersin.org/articles/10.3389/fnut.2022.826162

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach* (1st ed.). O'Reilly Media Inc. Retrieved from https://www.tidytextmining.com/

Smith, M., Lal, P., Oluoch, S., Vedwan, N., & Smith, A. (2021). Valuation of sustainable attributes of hard apple cider: A best-worst choice approach. *Journal of Cleaner Production*, *318*, 128478. https://doi.org/10.1016/j.jclepro.2021.128478

Souza Gonzaga, L., Capone, D. L., Bastian, S. E. P., Danner, L., & Jeffery, D. W. (2020). Sensory typicity of regional Australian Cabernet Sauvignon wines according to expert evaluations and descriptive analysis. *Food Research International*, *138*, 109760. https://doi.org/10.1016/j.foodres.2020.109760

Spence, C. (2020a). Wine psychology: Basic & applied. *Cognitive Research: Principles and Implications*, *5*(1), 22. https://doi.org/10.1186/s41235-020-00225-6

Spence, C. (2020b). Food and beverage flavour pairing: A critical review of the literature. *Food Research International*, *133*, 109124. https://doi.org/10.1016/j.foodres.2020.109124

Spinelli, S., Dinnella, C., Masi, C., Zoboli, G. P., Prescott, J., & Monteleone, E. (2017). Investigating preferred coffee consumption contexts using open-ended questions. *Food Quality and Preference*, *61*, 63–73. https://doi.org/10.1016/j.foodqual.2017.05.003

Stelick, A., & Dando, R. (2018). Thinking outside the booth—The eating environment, context and ecological validity in sensory and consumer research. *Current Opinion in Food Science*, *21*, 26–31. https://doi.org/10.1016/j.cofs.2018.05.005

Stoller, G. (2019). Craft beer and food: Finding the perfect pairing. *Forbes* Retrieved from https://www.forbes.com/sites/garystoller/2019/02/26/craft-beer-food-finding-the-perfect-pairing/

Sáenz-Navajas, M.-P., & Jeffery, D. W. (2021). Perspectives on wines of provenance: Sensory typicality, quality, and authenticity. *ACS Food Science & Technology*, *1*(6), 986–992. https://doi.org/10.1021/acsfoodscitech.1c00128

Tao, D., Yang, P., & Feng, H. (2020). Utilization of text mining as a big data analysis tool for food science and nutrition. *Comprehensive Reviews in Food Science and Food Safety*, *19*(2), 875–894. https://doi.org/10.1111/1541-4337.12540

ten Kleij, F., & Musters, P. A. D. (2003). Text analysis of open-ended survey responses: A complementary method to preference mapping. *Food Quality and Preference*, *14*(1), 43–52. https://doi.org/10.1016/S0950-3293(02)00011-3

Thomas, A., & Pickering, G. (2003). The importance of wine label information. *International Journal of Wine Marketing*, *15*(2), 58–74. https://doi.org/10.1108/eb008757

Tian, G., Lu, L., & McIntosh, C. (2021). What factors affect consumers' dining sentiments and their ratings: Evidence from restaurant online review data. *Food Quality and Preference*, *88*, 104060. https://doi.org/10.1016/j.foodqual.2020.104060

Tozer, P. R., Galinato, S. P., Ross, C. F., Miles, C. A., & McCluskey, J. J. (2015). Sensory analysis and willingness to pay for craft cider. *Journal of Wine Economics*, *10*(3), 314–328. https://doi.org/10.1017/jwe.2015.30

Vasiljevic, M., Coulter, L., Petticrew, M., & Marteau, T. M. (2018). Marketing messages accompanying online selling of low/er and regular strength wine and beer products in the UK: A content analysis. *BMC Public Health*, *18*(1), 147. https://doi.org/10.1186/s12889-018-5040-6

Vidal, L., Ares, G., & Jaeger, S. R. (2022). Biterm topic modelling of responses to open-ended questions: A study with US consumers about vertical farming. *Food Quality and Preference*, *100*, 104611. https://doi.org/10.1016/j.foodqual.2022.104611

Weber, R. P. (1990). *Basic content analysis*. SAGE.

West, E. (2018). United States cider map—Cider guide. Retrieved from https://www.ciderguide.com/cider-maps/united-states/

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*, 1–23. https://doi.org/10.18637/jss.v059.i10

Wijffels, J., BNOSAC, & Yan, X. (2021). BTM: Biterm topic models for short text (0.3.6) [R]. Retrieved from https://cran.rproject.org/web/packages/BTM/index.html

Wood, G. (2022). Good harvest: The industry will likely benefit from an influx of new entrants (No. OD5335). IBISWorld Industry Report OD5335.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). Biterm topic models for short text. Institute of Computing Technology, CAS.

Yenerall, J., Jensen, K., Hughes, D. W., Trejo-Pech, C., & DeLong, K. L. (2022). Demographics, alcoholic beverage purchase patterns, and attitudes driving hard cider expenditures. *Journal of Food Products Marketing*, *28*(5), 228–241. https://doi.org/10.1080/10454446.2022.2096423

Yoo, R., Kim, S.-Y., Kim, D.-H., Kim, J., Jeon, Y. J., Park, J. H. Y., Lee, K. W., & Yang, H. (2023). Exploring the nexus between food and veg*n lifestyle via text mining-based online community analytics. *Food Quality and Preference*, *104*, 104714. https://doi.org/10.1016/j.foodqual.2022.104714

How to cite this article: Calvert, M. D., Cole, E., Neill, C. L., Stewart, A. C., Whitehead, S. R., & Lahne, J. (2023). Exploring cider website descriptions using a novel text mining approach. *Journal of Sensory Studies*, e12854. https://doi.org/10.1111/joss.12854