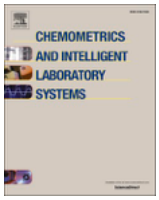
Chemometrics and Intelligent Laboratory Systems 267 (2025) 105521

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Beer's linguistics and chemistry: an investigation opening new research perspectives

Nicola Cavallini [a,*], Francesco Savorani [a], Rasmus Bro [b], Marina Cocchi [c]

[a] Department of Applied Science and Technology, Politecnico di Torino, Corso Duca Degli Abruzzi 24, 10129, Torino, TO, Italy
[b] Department of Food Science, University of Copenhagen, Rolighedsvej 26, 1958, Frederiksberg C, Denmark
[c] Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi 103, 41125, Modena, MO, Italy

ARTICLE INFO

Keywords:
Food
Beer
Spectroscopy
Text analysis
NLP
Chemometrics
GCA

ABSTRACT

In the last two decades, interest in food production and consumption has progressively grown, alongside the booming popularity of craft beer, fueled by micro-breweries and home brewing. Beer is a complex mixture of compounds — from carbohydrates to proteins and ethanol — shaped by the recipe, ingredients, and production process. Less obvious is that the human tongue, in synergy with the oral cavity and nose, acts as a powerful sensor array. Tasting experiences can be viewed as "analytical sessions", where sensory signals processed by the brain determine not only if the beer is appreciated but also which tastes and flavours are perceived.

In our study, we investigated the connection between the "objective" chemical profile of beer and the "subjective" sensory descriptions from user reviews. We analysed 88 beers using near-infrared (NIR), visible, and nuclear magnetic resonance (NMR) spectroscopy, pairing them with text reviews processed through natural language processing (NLP) tools and converted into numerical data via a bag-of-words approach. Principal Component Analysis-Generalized Canonical Analysis (PCA-GCA) revealed correlations between chemical signals and topics like "hops," "brown colour," and "booze". NMR data showed the strongest correlations, especially for hops-related terms, while visible spectra linked to colour descriptors. Automated topic extraction often performed comparably to manual term selection, suggesting potential for scalable studies. Despite limitations like dataset size and beer variety, this approach shows promise for aligning chemical composition with sensory perception, with applications for product development and broader food analysis.

A novel approach integrates text corpora with analytical data through chemometrics, linking language complexity to instrumental responses. Results showed strong correlations, like NMR signals with hops-related terms and visible spectra with beer colour. This previously unexplored connection opens the door to designing food products tailored to consumer preferences. The approach is broadly applicable, from food science to medical diagnosis or aligning expert opinions with factual data.

1. Introduction

In recent years no one can escape cooking shows on television. This reflects a growing trend that has been on the rise for at least a decade. The public is increasingly conscious of various aspects related to food production, preparation, and consumption. Consequently, new restaurants and food production techniques are continuously emerging, accompanied by experimentation with recipes and food pairings. This phenomenon is driven by high quality standards and embodies a language often referred to as the "craft rhetoric", where craft/handmade is opposed to industrial [1], and mass production is opposed to artisanal [2]. While this rhetoric extends beyond the food industry, it particularly thrives in the case of food and beverage [3]. Over the past two decades, the beer industry has also experienced significant transformations, propelled by the proliferation of craft and micro-breweries [4,5] and by the spread of home brewing [6], mostly in the Western countries.

This decade has also witnessed a remarkable surge in data generation, collection, storage, and processing. The line between the physical and digital realms is increasingly blurred, therefore it comes as no surprise that the food sector has become deeply intertwined with information technology [7], which plays a pivotal role not only in aspects related to production, safety, and health but also in market dynamics, where it aims to detect emerging trends and consumer opinions. The concept of "computational gastronomy", introduced by Goel et al., encapsulates the practice of data-driven research encompassing flavour,

* Corresponding author.
E-mail address: nicola.cavallini@polito.it (N. Cavallini).

https://doi.org/10.1016/j.chemolab.2025.105521
Received 3 June 2025; Received in revised form 28 July 2025; Accepted 30 August 2025
Available online 30 August 2025

Abbreviations

GCA	Generalized Canonical Analysis
MCR	Multivariate Curve Resolution
NIR	Near-Infra Red
NLP	Natural Language Processing
NMR	Nuclear Magnetic Resonance
OPTICS	Ordering Points to Identify the Clustering Structure
PCA	Principal Component Analysis
PC	Principal Component
PMD	Penalized Matrix Decomposition
Vis	Visible

food pairing, recipes, health, and nutrition [8]. Data-driven computational approaches are the foundation for numerous transformative changes and decision-making across various fields. Virtually any market can reap benefits from efficiently collecting, processing, and interpreting customer-generated data, for example in textual form sourced from social media, blogs, and specialized websites [8,9]. From the analysis of such text corpora, new directions emerge for customizing products to meet precise consumer expectations and taste preferences. This not only holds the potential for safer [7], healthier, and tastier food products catering to distinct customer segments but also enables manufacturers to expand their market reach and customer base [10].

In the field of computational gastronomy, chemical-physical analyses play a central role. These analyses have proven successful in characterizing a diverse range of food matrices along various production chains, serving multiple purposes such as addressing health and safety concerns, ensuring food authentication [11,12], detecting fraud [12], and ensuring the quality of both raw and processed materials. Analytical chemistry [13] acts as the indispensable toolkit for achieving objective characterizations of these food matrices. Typically, non-destructive and efficient methods are employed, often relying on spectroscopic techniques [14–18]. The study of molecular and chemical aspects of food and its constituents falls under the umbrella of "foodomics" [19] and, when beer is under examination, one can refer to "beeromics" [20,21]. Analytical chemistry in synergy with advanced data analysis, or better, "chemometrics" [22,23], can be profitably used to build new tools to aid consumers when choosing and pairing [24] foodstuff, and producers to meet the consumers' expectations.

The aim of our study is to investigate the links between the "objective" world of analytical chemical profiling – e.g., using spectroscopy – and the "subjective" world of consumers tasting and describing food – e.g., using text descriptions of beer products. The "objective" perspective about beer has been investigated both from the points of view of its chemistry and composition [25–28] and also of the consumers' preferences [1,29], which are traditionally assessed by directly interviewing small groups of people. However, with the growth of the Internet and of online communities, mining online-posted text reviews has increasingly become a useful method for gathering data and assessing product appreciation and reception [30,31]. This leads to the "subjective" perspective about beer in which many psychological, socio-cultural, biological, and situational factors influence the choice and consumption of this beverage [32]. People discuss their preferences in different social settings, including websites like ratebeer.com (website closed in early 2025), beeradvocate.com, and untappd.com, which currently host years of user-generated reviews about beer, with both textual data and numeric ratings. The analysis of online reviews of beer has been mainly investigated by marketing scholars to perform customer segmentation [33], to predict consumers' satisfaction and product success [34] as well as to explain the users' ratings [35]. These approaches and studies considered large groups of beer, with the aim of extracting valuable information about the consumers, for marketing decision-making, with no direct connection with the actual chemical and sensory properties of the product. This aspect could however help with marketing and promotion of certain types of beer, targeting for specific market segments. By knowing the compositional fingerprints of beer and their associated sensory traits the producer can try to change the recipe to obtain by design specific types of sensory responses that the consumers would most likely appreciate.

In our study 88 beers were examined using both analytical chemistry techniques and user-generated text data. These beers were analysed with near-infrared (NIR, [36–39]), visible [40–42] and nuclear magnetic resonance (NMR, [43–47]) spectroscopies. Text reviews about the same 88 beers were obtained from a popular beer rating website and were then processed and analysed by using simple natural language processing (NLP, [48,49]) tools. Text analysis methods [50,51] were applied to process the text reviews and to convert them into numeric format, using the *bag-of-words* approach [50]. The obtained complete data matrix consisted of wordcounts of relevant English terms used to describe the 88 beers. To fully investigate the possible links with the collected analytical chemical signals a chemometric pipeline was developed. First, meaningful subsets of terms were extracted from the complete text dataset, using two approaches: topics extraction [52] by penalized matrix decomposition (PMD, [53]) and manually-defined sets of terms related to specific aspects of beer making and tasting. Then, the links between individual subsets of terms and each spectral datasets were investigated by using principal component analysis-generalized canonical analysis (PCA-GCA, [54]). The results showed interesting correlations of the spectral datasets with general topics such as flavours of "hops", the appearance stated as "brown colour", the alcohol content seen as a way to get drunk ("booze"), and the idea of "refreshment".

To the best of the authors' knowledge, very few studies attempted to link experimental chemical-physical measurements to user-generated text data. A study was published in 2021 by Fox et al., who tried to link reviews obtained from ratebeer.com to chemical and tribology data [55]. They also modelled the numerical ratings from the website, which in our case were not included in the study due to the abundance of information and results to be discussed, as three spectral datasets were collected and processed. The major difference between the two studies is that, although the analysis of the textual data and the chemical-physical data was multivariate, it was not conducted jointly. Moreover, the sample size and composition are extremely different: while Fox et al. focused on four beer products and their non-alcoholic versions, in our study a much larger and more diverse collection of beers was considered, and at the same time three different spectroscopic techniques were employed (namely NIR, NMR and visible). The "Brewfinder" by Price [56] is a project that resulted in an online visualization tool [57] based on a principal component analysis of a matrix of frequencies of flavour terms obtained from ratebeer.com. However, in this case no chemical data are included or related to the text data.

A novel approach is here proposed to integrate and compare the information content of a text with analytical instrumental responses where chemometrics is the key to link this very diverse data. Beer products are an ideal benchmark to test such an approach, and the results of our study, even if possibly affected by limitations due to the dataset size, suggest an important and promising direction for further matching the chemical composition of food with the consumers' appreciation and taste, with interesting practical directions for the producers as well.

2. Materials and methods

The first part of the materials and methods section is devoted to the text data: first, the text data collection will be described (Section 2.1.1); then, the methods applied to process the text corpora for obtaining the numerical wordcounts matrix are discussed (Section 2.1.2); then the analysis methods used for processing the wordcounts matrix are reported (Sections 2.1.3 and 2.1.4). Part two covers the spectral datasets

used in this study (Section 2.2). Finally, part three deals with the method used for linking the text to the spectral data (Section 2.3). The whole experimental data analysis workflow is depicted in Fig. 1.

2.1. Text analysis

The aim of text analysis [50,58] methods is to extract information from text documents by converting the text data to a format suitable for analysis. With these methods, it is possible to operate on different levels of detail [50,58], from the simplest approach of counting the occurrences of words or groups of words (i.e., their count or frequency of appearance in the processed documents) to the analysis of whole sentences with methods like sentiment analysis [59,60] or latent semantic analysis [61] (i.e., extraction of the semantic structure of text by considering relationships and relative positions among words).

In the present study, the text data were converted to wordcounts. This approach envisages the creation of a so-called *bag-of-words* model [50], which consists of a list of terms (a *vocabulary*), and their counts in each document used for building the model. The wordcounts are arranged in an array whose rows represent the documents (or samples), and each column is associated with one term. This matrix can be analysed with common multivariate methods, and the results can be interpreted according to the most influent terms.

2.1.1. Text data collection

The text data describing the beer samples under examination consisted of user comments/reviews obtained from ratebeer.com, a website currently shut down (not operative since February 2025) on which the users could review any beer they have tasted by writing a comment and giving scores according to five parameters (aroma, appearance, taste, palate, overall). Many users of such websites, especially the most prolific in tasting and reviewing, often use in their written comments descriptors taken from the "standard" sensory terminology (more details about this in Section 3.1.1), probably being inspired by the different aspects of the five rating parameters. All comments were collected during August 2017 and were organized as an array of 88 text corpora, i.e., one corpus for each beer sample, each corpus containing the concatenated comments that were gathered from the website.

2.1.2. Text data processing: bags-of-words models and wordclouds

Text data processing consists of a series of steps aimed at converting the input text corpora into a format suitable for further analysis. To this aim, Natural Language Processing [48,49] is used. The approach of counting how many times each word has occurred in the input text corpora is called *bag-of-words* modelling [30,50,62,63]. When only the occurrence of individual words (e.g., "warpigs", "galaxy", "awesome") is considered, then the distance relationships among the words in a sentence are neglected (e.g., the sentences: 1. "Warpigs galaxy is awesome" and 2. "Many years ago I visited the galaxy. The trip lasted two months and I could experience different landscapes and people, such as the Warpigs. Overall, the trip was awesome" would translate in a data row of occurrence 1 1 1 in correspondence of the columns referring to the terms "warpigs", "galaxy" and "awesome", i.e. the individual words are considered separately and independently, and their distance relationship in the sentence is ignored). This type of text model is called bag-of-words of *unigrams*. If instead the occurrences of groups of *n* words are considered (e.g., "warpigs galaxy awesome" all together), the model would be a bag-of-words of on *n*-grams (i.e. in this case sentence 1. would translate in a data row of occurrence 1 in correspondence of the column "warpigs galaxy awesome" and sentence 2. would translate in a data row of occurrence 0, since the words are not appearing close to each other). The bag-of-words is composed by its own "vocabulary", which is the list of all the individual words/terms (each one being either an uni- or an *n*-gram) contained in the modelled text corpus, and the corresponding wordcounts. By processing more text corpora together, the resulting wordcount matrix would have as many rows as the number of processed corpora, and as many columns as the terms in the vocabulary. In the present work the unigrams bag-of-words model was used, so the

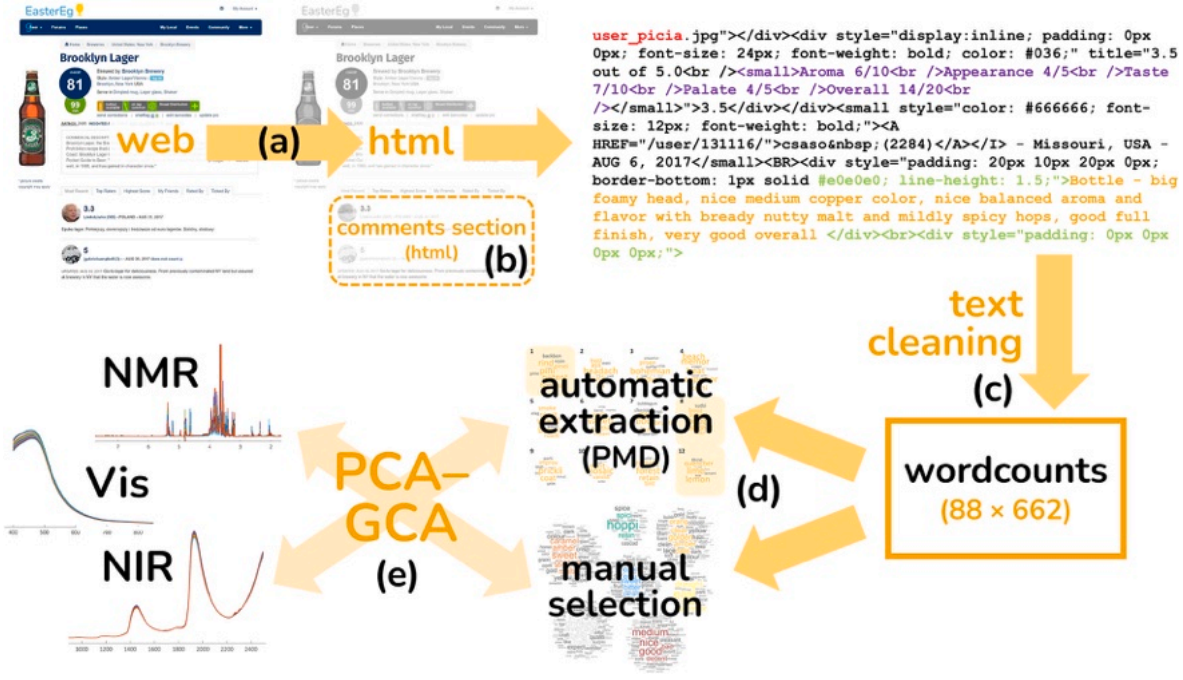


Fig. 1. Data analysis workflow: web-scraping (a) is followed by the extraction of the comment's section (b), whose text undergoes the text cleaning producing the wordcounts matrix (c). Automatic topics extraction (with PMD) and manual selection of meaningful groups of terms are performed on the wordcounts matrix (d). To each topic and group of terms PCA-GCA is applied to inspect the link with the spectral data (e). The example comment shown in the figure is real (only the username was changed).

columns of the wordcounts matrix correspond to individual words. The rows of this matrix correspond to one beer, whose text corpus consists of the user-generated comments concatenated in an individual string of text.

The information stored in a bag-of-words can be represented using a *wordcloud*, a visualization tool [64] that allows to highlight the most important terms of a collection of words. The importance of a term is usually determined by its frequency of appearance within the text corpus (information contained in the wordcount matrix), and the clearest way of representing it is done by using both size and colour coding: the bigger the word in the cloud, the more frequent/important it is. The top five words generally have a colour different from black.

2.1.3. Text data processing: cleaning and English selection

To convert the text data to a numeric format suitable for analysis, a series of cleaning steps is required. The cleaning procedure is fundamental: since punctuation and very common but uninformative particles (e.g., "and", "the", "is", "with") are always prevalent in any text corpora, they should be removed. A detailed description of the cleaning steps taken in this work is provided in the Supplementary Materials (Section S1).

We also decided to focus on the English language only. This allowed working with a "more homogeneous" dataset, avoiding splitting the same information across terms belonging to different languages. All the terms belonging to other languages were therefore excluded from the final Counts dataset. The procedure for precisely selecting the English terms is rather interesting but for conciseness it was moved to the Supplementary Materials (Section S2).

The cleaning processing resulted in a bag-of-words consisting of 662 English terms (represented as a wordcloud in Fig. 2), corresponding to a reduction of 99.09 % from the starting text corpus. It is worth remembering that the main aim of these processing steps was to obtain a set of terms useful for describing beers, i.e., sensory-like terms. An extremely high reduction rate must therefore be expected. From this moment on, this list of 662 terms will be referred to as the "experimental beer vocabulary".

Fig. 2. Wordcloud of the top 100 terms of the experimental beer vocabulary, by frequency. The top 5 terms are also highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

2.1.4. Wordcount data preprocessing

The traditional preprocessing of wordcounts matrices goes under the acronym of *tf-idf* (term frequency-inverse document frequency, [65]). This approach is usually employed for machine learning purposes and in the classification context [66,67]. However, in the present work a "chemometric perspective" was preferred with the aim of being able to interpret the data and understand their structure, instead of "only using" them. Therefore, a preprocessing consisting of row-normalization and column-autoscaling (unit variance scaling) was chosen. The wordcounts matrix was row-normalized using the sum along the rows, to account for the different numbers of comments associated with each sample (i.e., each beer). As an alternative, the number of comments was also tried for row-scaling when studying the most suitable preprocessing method, and it was found less efficient than the row sum, probably because not only the number of comments is important, but also their length and complexity, whose information may be better captured by the word counts. Autoscaling was finally employed for giving all terms the same importance prior to any modelling step.

2.1.5. Extraction of meaningful topics via penalized matrix decomposition (PMD)

To work with subsets of the wordcount matrix, different ways for automatically defining groups of related terms can be used. A set of related words can be defined as a "topic" [52]. Many methods for automatically extracting meaningful topics were considered and tested: principal component analysis (PCA [68]), multivariate curve resolution (MCR, [69]), archetypical analysis [70,71], coclustering [72], latent Dirichlet allocation [73] and penalized matrix decomposition (PMD, [53]). Since the two main aims were (i) obtaining meaningful groups of terms and (ii) that these groups should have reduced dimensionality, PMD was chosen for the quality of the obtained topics and the sparseness of the groups (groups with maximum ten terms were obtained). PMD is a bilinear decomposition method for obtaining meaningful clusters based on non-negative matrix factorization [74] which applies two-sided sparsity. This means that both scores and loadings are sparse, and this corresponds to our desired situation: the groups of terms should correspond to groups of beers, capturing the peculiar aspects that makes them a cluster. In other words, a topic can emerge if a specific feature or set of features is shared among groups of beers (i.e., if the scores are sparse). A brief recall to the basics of the method can be found in the Supplementary Materials (Section S3).

2.1.6. Use of the text data

As stated before, the text data could not be fruitfully used as a whole, but an approach based on the selection of meaningful subsets of words had to be applied. The most interesting subsets, namely two manually selected topics and four topics selected by PMD, will be further discussed and visually represented in Section 3.1.2. Please notice that even if *groups of terms* and *topics* are basically synonyms, it was decided to refer to the manually defined topics as "groups of terms" and to use "topics" for referring to the ones automatically obtained by PMD.

2.2. Spectral data

The spectral datasets of our study were obtained using visible, NIR and NMR spectroscopies. The present study was developed in parallel with two other works by the authors based on the same datasets: a technical work proposing a features extraction/data fusion method [75], in which the visible, NIR and NMR datasets were used for testing the methodology, and a metabolomics-oriented article focused on the NMR beer dataset alone [21]. The NMR dataset of this paper is like the one of Cavallini et al. [21], but without the ethanol resolved signals (i.e., the dataset has two fewer columns). An important difference from the two related works is that for the present study all datasets were reduced from 100 to 88 samples, since 12 of them were replicates that underwent thermal treatment: these samples were removed, as it is more logical to

consider the non-treated samples more similar to the products that consumers can buy, taste and review, making them a more reliable match with the text dataset. Additional details about the beer samples are provided in Table S1.

2.3. Linking datasets: Principal component analysis-generalized canonical correlation (PCA-GCA)

The issue of linking two datasets is at the core of the present study. Different approaches can be taken, and one of the most established chemometric methods for inspecting the relations between two data blocks is partial least squares (PLS, [76,77]) regression. However, the PLS regression models obtained using the extracted topics provided very poor results, suggesting that strict prediction of single terms or even topics is probably not attainable with the available data.

For this reason, it was decided to investigate the connections between spectral and text data using a method able to find common subspaces among data blocks, which is a combination of principal component analysis and generalized canonical correlation (PCA-GCA, [54]). The aim of PCA-GCA is to find linear combinations of the blocks under examination, fitting as well as possible a set of orthogonal common components. This set of common components is not necessarily contained in the column space of any data block, but it is instead in the combined space represented by the concatenated matrix obtained by joining the data blocks in the variables' direction [54]. GCA aims at finding the trends in the data blocks that correlate the strongest. Since this method only focuses on correlation, within-block variability may be not captured at best. PCA operated on the individual blocks prior to GCA helps enhancing the stability of the extracted common components [54].

Combining PCA with GCA leads to a stepwise procedure in which the first step is devoted at determining the subspaces' dimensionalities (each block having its own dimensionality). Then, the correlation coefficients and the block-wise explained variances are obtained by GCA and are used for deciding the number of common components. Each common component is in fact obtained by separately estimating one component from each data block, in a way that all the block-related components are as similar as possible (i.e., they are as strongly correlated as possible). The number of common components is chosen according to how strong the correlation is and by evaluating the amount of variance explained from each data block. A brief recall to the basics of the method can be found in the Supplementary Materials (Section S4).

2.4. Software

All analyses were carried out under MATLAB environment (2017b and 2021a, Mathworks, MA, USA). Text data processing and visualization was done using the functions provided by the Text Analytics Toolbox (version 1, Mathworks, MA, USA) and in-house written scripts. The text corpora were processed using word embeddings trained using MATLAB's *word2vec* function, whose code is based on the original algorithm developed by Mikolov et al. [78,79]. Word embedding visualization was performed using the t-SNE function provided in the MATLAB Statistics and Machine Learning Toolbox (version 11.2). The PMD algorithm used for topics extraction was written by Jose Camacho Paez and it is based on the work by Witten et al. [53]. The PCA-GCA toolbox for MATLAB [80] was provided as a curtesy by Ingrid Måge and it is based on [54,81].

3. Results

The results section is organized into two parts. First, the beer vocabulary will be described and inspected, starting from the established standard terminology (Section 3.1.1), which will be compared with the experimental vocabulary (i.e., the results of the terms extraction from the text data will be reported, in Section 3.1.2). Then, the results of linking the text data with spectral data by means of PCA-GCA will be discussed in Section 3.2.

3.1. Investigation of the beer vocabulary: established terminology and experimental topics

To better understand the results of linking the text data to the spectral data, a comparison between the content of the text data and the established terminology about beer sensory tasting is fundamental. This section will therefore start with a description of the standard established beer sensory vocabulary (Section 3.1.1), which will be followed by some considerations about the experimental beer vocabulary as obtained from the users' comments (Section 3.1.2).

3.1.1. Established terminology of beer flavour and aroma

Sensory description of foodstuff is generally done by means of a *lexicon* [82,83], which is a set of terms used for documenting and describing the sensory perceptions of a selected food [82]. By defining a lexicon, it is possible to perform sensory studies with a panel of evaluators who are "speaking the same language", so that the results of their assessments can be compared and quantitatively elaborated. During the years, the beer flavour terminology [84] has been constantly enriched, updated and used for characterizing [85] different commercial beers, but it also made easier the communication between stakeholders, scientists and also consumers. An example of the beer flavour wheel, first introduced by Meilgaard et al. [84], is shown in Fig. 3.

The purpose of this visual device is to allow locating quickly and easily any sensory term. The sensory terms are grouped into classes (middle ring with stronger background colour in Fig. 3) and organized in tiers within each class. The second-tier terms are located on the external ring, which represents the highest level of detail in this classification system: as defined by Meilgaard et al., each "separately identifiable flavour characteristic has its own name" [84]. If the second-tier terms represent the most detailed description of a flavour characteristic, the first-tier terms represent the family to which these terms belong to: for instance, within the *aromatic, fragrant, fruity, floral* class, the first-tier term *fruity* represents the flavour characteristics directly related to typical fruit flavours, like *citrus*, *apple*, *banana*, *blackcurrant*, *melon*, *pear*, *raspberry,* and *strawberry*.

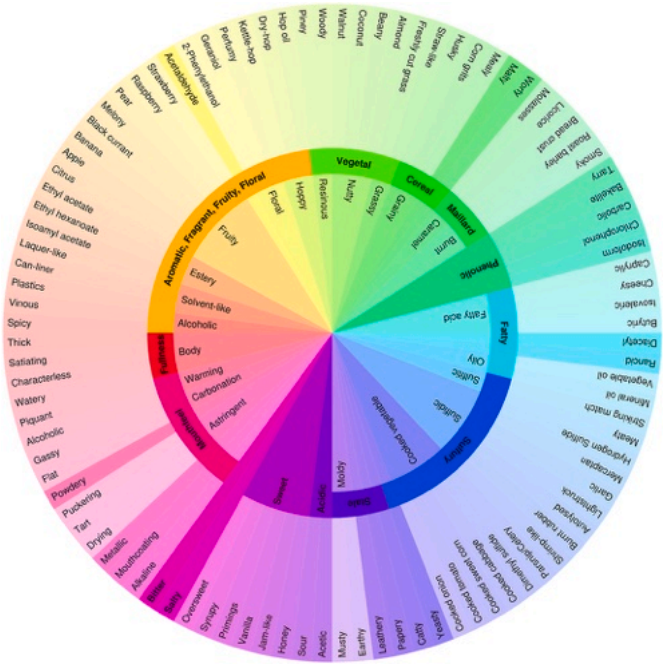


Fig. 3. The beer flavour wheel generated by Abraham Kabakoff [86].

3.1.2. Experimental beer vocabulary

Many of the 662 terms composing the experimental beer vocabulary are clearly derived from the standard beer flavour terminology. However, it can be noted that the experimental vocabulary also includes many other "sensory-like" terms which often are just synonyms of the standard terms. Also, additional descriptors covering the different aspects of beer tasting, e.g., its appearance, the real-word circumstance of consuming it and the consumer's opinions about the products appear in the experimental vocabulary. As motivated in Section 2.1.5, the whole dataset was reduced to six sub-sets characterized by a "topic". To facilitate the inspection and comparison among the six topics, a visual representation is given in Fig. 4.

Six groups of terms were manually defined, according to macro aspects of beer tasting, such as general sensory-like terms, recognition of malt- or hops-related flavours, appearance/colour, experience-memory linked to the consumption of beer, and personal judgement. Only the manually defined hops- and appearance/colour groups (Fig. 4a and b) were used in the modelling steps, mainly as a sort of benchmark for assessing the similar topics automatically extracted by PMD (i.e., PMD-hops and PMD-brown colour).

Twenty topics were also extracted, using PMD. Twelve of them showed clear links to the sensory world: many terms came directly from the standard beer flavour terminology. Two topics seemed to be related to negative experiences, with characteristic terms, sometimes vulgar, such as *waste*, *ass*, *piss*, *urine*, *suck*, *headache*, *blah* or *cost* (from beers that did not quite match quality/price expectations?). On the contrary, three topics seemed to be related to positive experiences (*goto*, *gateway*, *reliable*, *staple*, *favorite*, *flagship*), bringing back memories (*memory*, *reliable*) and specific situations in which thirst needs to be quenched (*summer*, *beach*). Finally, two topics included very mixed terms that hindered the identification of a common theme. Of these automatically PMD-extracted topics only four of them provided interesting and interpretable results, which will be discussed in Section 3.2.

3.2. Linking the topics to the spectral data with PCA-GCA

The PCA-GCA association with the spectral data was computed for all PMD topics, but only those showing correlations above 0.70 were further inspected: the results for four of them (corresponding wordclouds in Fig. 4c–f) are reported. A summary of the nine most meaningful and interesting comparisons is provided in Table 1.

Table 1
List of the nine discussed comparisons. For each general topic the pairs of compared data blocks are reported, together with the corresponding correlation value and the figure in which the results are displayed. The correlation values that changed significantly when highly influential samples were removed are marked with *.

General topic	Data blocks: Spectra	Text (wordcloud for comparisons)	Correlation	Fig.
Hops	NMR	PMD-hops (Fig. 4d)	0.91	5
	NMR	hop-terms (Fig. 4a)	0.91	6
Brown colour	NMR	PMD-brown colour (Fig. 4e)	0.85 (0.51*)	7
	Vis	PMD-brown colour (Fig. 4e)	0.75	8
	NMR	appearance/colour terms (Fig. 4b)	0.86	9
	Vis	appearance/colour terms (Fig. 4b)	0.88	10
Booze	NMR	PMD-booze (Fig. 4c)	0.74	11
	NIR	PMD-booze (Fig. 4c)	0.74 (0.56*)	12
Refreshment	NMR	PMD-refreshment (Fig. 4f)	0.76	13

3.2.1. "Hops"

3.2.1.1. Hops: NMR and PMD-hops. PMD-hops provided the best common component correlation result with the NMR data (0.91) and it appears to be related to the presence of hops [87,88]. Terms like *resin* and *pine* (in the Counts loadings, Fig. 5c) refer to the resins extracted from the hops' cones while boiling the beer wort [89], during the beer brewing process. Another term from the beer flavour terminology is *piney*, and it can be found under the "vegetal" class and the first-tier term *resinous*, in the beer flavour wheel (Fig. 3). This also confirms the topic's association with the hops.

The top-scoring chemical compound in the NMR loadings is trigonelline, a plant metabolite generally found and studied in relation to coffee [90,91] for its pharmacological [92,93] and health benefits [91]. It has recently been found in beer [43,45,94] and described as a plant-associate metabolite whose concentration increases with boiling [95]. Hops are generally added right before boiling the beer wort, so that

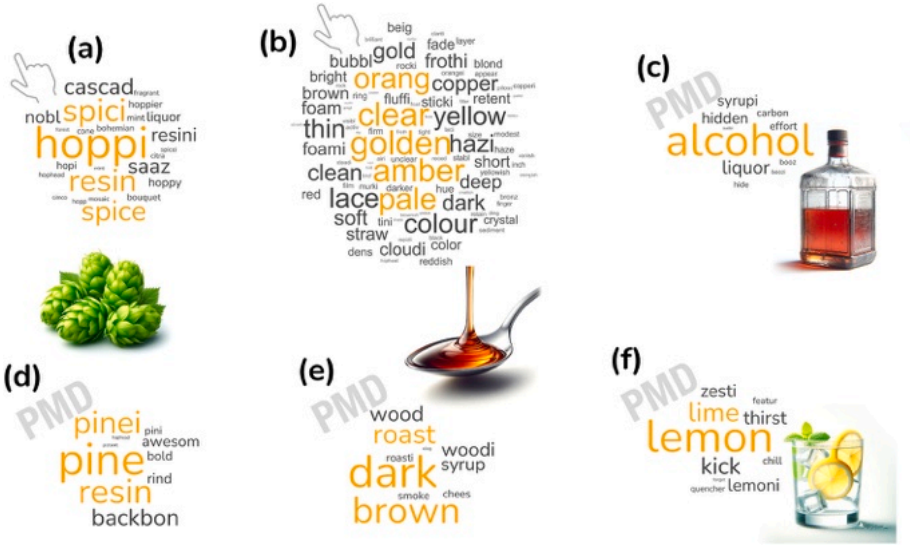


Fig. 4. The topics and groups of terms modelled with PCA-GCA. In (a) the hops-related and in (b) the appearance/colour-related manually defined groups of terms. In (c) PMD-booze, in (d) PMD-hops, in (e) PMD-brown colour, and in (f) PMD-refreshment. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

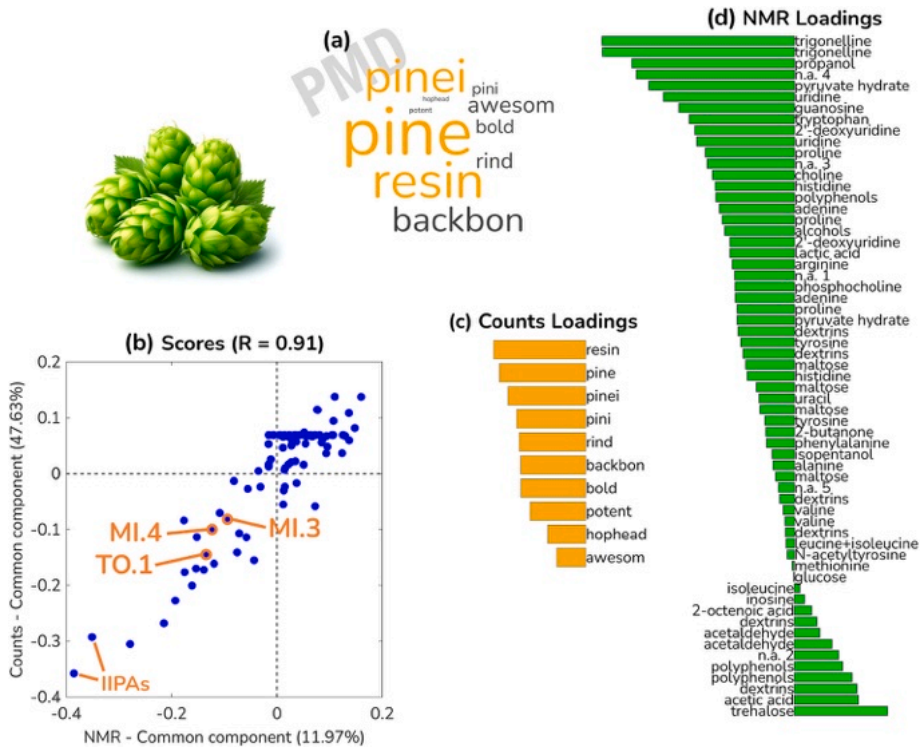


Fig. 5. Hops: NMR and PMD-hops (10 terms). PCA-GCA results obtained by comparing the NMR dataset and PMD-hops (a). The common scores are in (b), the Counts loadings are in (c), the NMR loadings are in (d).

heat allows converting the hop acids to become soluble and extracting them. For these reasons, trigonelline can also be associated with the hops. Since the loadings associated with *resin*, *pine* on the Counts and trigonelline on the NMR data share the same negative direction, a connection between them can be deduced, confirming that PMD-hops can be related to the hops.

By inspecting the samples' distribution in the scores plot of Fig. 5b, at negative scores values India pale ales (IPAs), imperial IPAs (IIPAs) and generic ales are mostly found. The presence of three lagers can be explained with their recipes, which are rich in hops: sample MI.3 (*Helping Hand*, Mikkeller) is described as a "hoppy pilsner" [96]; sample TO.1 (*Hop Love Pils*, To Øl) is described as "brewed with lots of hops"; sample MI.4 (*American Dream*, Mikkeller) is described as "packed with American hops". Terms like *bold* and *potent* also contribute to the samples' separation along the scores: the most extreme samples at negative scores in Fig. 5b belong to IIPA style, whose first "I" stands for "imperial": this attribute is used for very strong beers, both from the point of view of alcoholic strength and flavour richness. A clear link with the terms *bold* and *potent* can therefore be recognized and justified.

On the opposite end of the plot, at positive scores, only lagers from producers such as Hite, Heineken, Budwiser, Pilsner Urquell and San Miguel are present. These are very widespread brands, and their style does not involve much addition of hops or spices. They seem to be mainly characterized by variables related to sugars (dextrins and trehalose) and malt (polyphenols), as if in absence of a rich/peculiar bouquet of flavours the most basic taste of beer emerges. This is also confirmed by the opposite direction of the topic's terms *bold* and *potent*, which are logically distant from beers with more common and "flat" flavours.

An interesting contrast can be identified between two metabolites from the NMR loadings: trehalose and pyruvate (hydrate). Trehalose [97] is a disaccharide that is involved in the anaerobic carbohydrate metabolism in yeast cells, as an intermediate on the path for the formation of glycogen [98], an "energy storage" compound for yeast cells. Pyruvate, on the contrary, is an intermediate on the path that leads to the production of ethanol, alcohols, aldehydes, and esters. The opposite directions that trehalose and pyruvate have also correspond to the two main beer style families, namely ales and lagers. Ale beers tend to be richer in flavour and to have higher alcoholic strength: these products can be related to fermentation processes in which the yeasts produce a larger variety of metabolites. A link with the pyruvate path can therefore be traced, as opposed to the production of lagers, where the yeasts may also express to a larger extent the metabolic path related to trehalose and glycogen.

It is interesting to notice that in the Counts scores there is a set of samples that share the same score values: the words belonging to this topic may have been used in an extremely similar way for these samples, which may have ended up practically identical from the terms' point of view.

3.2.1.2. Hops: NMR and hop-terms (manual selection). The results obtained with PMD-hops closely resemble those obtained from the model built using the manually selected terms related to the hops: as shown in Fig. 6, similar common component correlation values were obtained. The inclusion of more terms probably made it possible to break the group of samples with very similar values along the Counts scores (i.e., the group of samples "horizontally aligned" in the scores plot in Fig. 5b).

3.2.2. "Brown colour"

3.2.2.1. Brown colour: NMR and PMD-brown colour. PMD-brown colour is characterized by interesting terms such as *wood*, *brown*, *roast*, *syrup* and *dark*. This combination suggests that features related to beers with darker colours and brownish hues were captured by the topic: hence, the name "brown colour". In the model relating the NMR data with PMD-brown colour, the common component correlation value looks good at a first glance, but by inspecting the distribution in the score plot of Fig. 7b it becomes very clear that one sample is driving the correlation. As a matter of fact, if sample SL.1 (at very negative scores in both components) is removed, the correlation value drops to 0.51, meaning

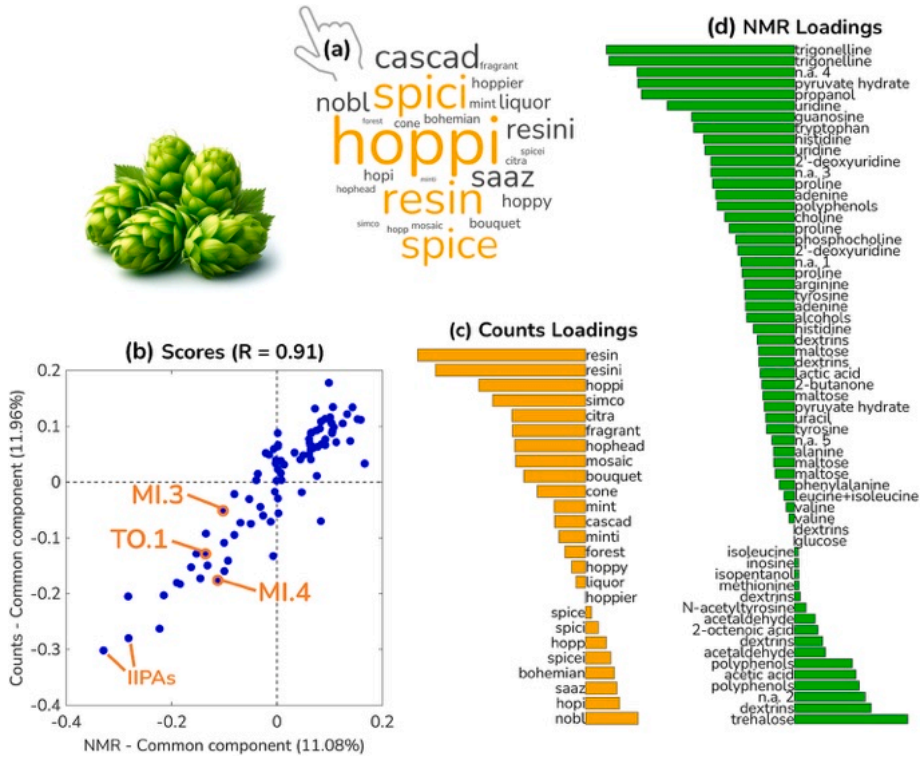


Fig. 6. Hops: NMR and hop-terms (manually selected, 25 terms). PCA-GCA results obtained by comparing the NMR dataset and the hop-terms (a). The common scores are in (b), the Counts loadings are in (c), the NMR loadings are in (d).

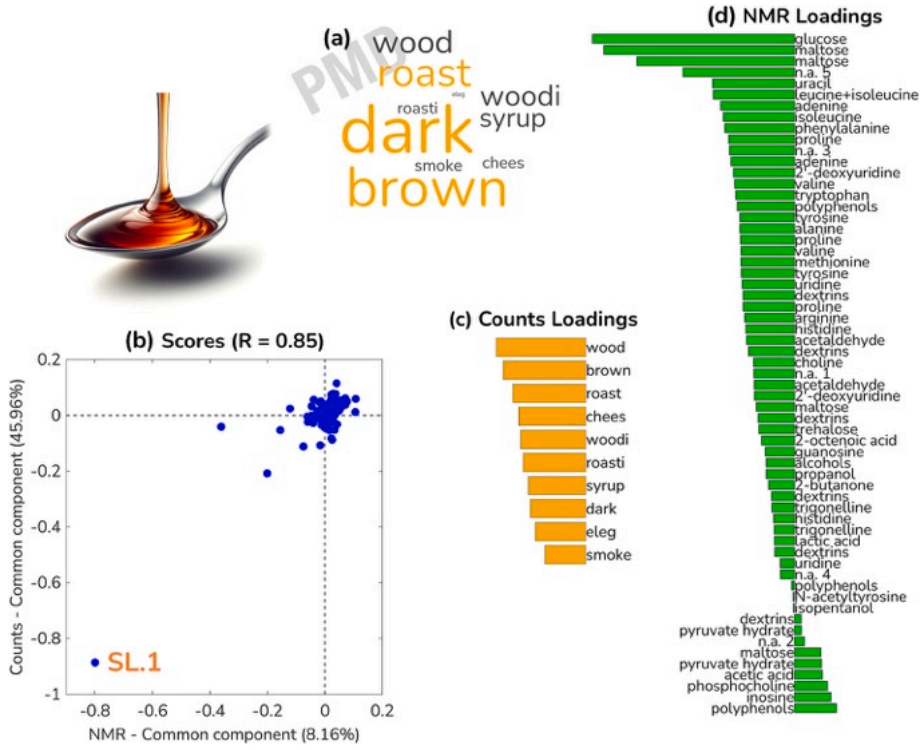


Fig. 7. Brown colour: NMR and PMD-brown colour (10 terms). PCA-GCA results obtained by comparing the NMR dataset and PMD-brown colour (a). The common scores are in (b), the Counts loadings are in (c), the NMR loadings are in (d). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

that even if a connection between glucose and maltose and the brown/roasted colour of PMD-brown colour seems plausible, the situation described in the figure might not be real.

3.2.2.2. Brown colour: Vis and PMD-brown colour. However, by inspecting PMD-brown colour in relation to the Vis data, a completely different situation was found: the result is a common component

correlation of 0.75, as shown in Fig. 8. The most correlated terms appear to be *dark*, *roast*, *brown* and *roasty*, suggesting that the brown and dark hues are captured by the Vis data, even if the spectral loadings interpretation can be quite difficult.

In this case, the common component correlation is mainly driven by the samples at positive scores in both directions, as represented in Fig. 8b: OR.3 strong ale, FU.1 amber lager, TY.2 amber lager, SL.1 brown ale, MA.2 lager strong, SL.3 Oktoberfest. These samples are mainly strong and darker beers, which is in line with the terms they are associated with.

3.2.2.3. Brown colour: NMR-Vis and appearance/colour terms (manual selection). PMD-brown colour seems therefore to be related, at least to some extent, to the appearance of the beer, mainly to its colour. If all the appearance-related terms of the experimental beer vocabulary are considered (defined by manual selection), both the NMR and Vis datasets perform well: NMR provides a correlation of 0.86 (Fig. 9), while Vis provides a slightly better correlation of 0.88 (Fig. 10).

In the case of the manually-selected appearance/colour-related terms (group b in Fig. 4), the most frequent ones are directly linked to beer colour: *orange*, *amber*, *golden*, *clear*, *pale* and *colour* followed by more specific terms such as *hazy*, *foamy*, *dark*, *lace* (the web-like pattern produced by the beer's foam when it dries on the glass' walls), "*copper*", "*clean*", "*yellow*" and so on. For the NMR dataset, terms referring to darker (*orange*, *black*, *amber*, *copper*) and hazy (*cloudy*, *hazy*, *murky*, *opaque*) beers results related to metabolites typical of ales and hopped beers (trigonelline, pyruvate and propanol). On the contrary, NMR signals such as trehalose, dextrins and polyphenols result more linked to terms like *clear*, *thin*, *yellow*, *pale*, *golden*, *straw* and *gold*, which are characteristic of lighter and clearer beers. The direction of the loadings of Fig. 9d corresponds to the trend in which at negative scores mainly ales are found, and on the other end of the distribution almost only lagers are found.

A situation similar to Fig. 9 is reported in Fig. 10: the Count loadings are ordered in almost the same way, with the light/clear beer characteristic terms on one end (negative scores), and the terms related to darker colours on the other end (positive scores). The Vis loadings of Fig. 10d seem to indicate that stronger absorption occurs in the 400–500 nm region, which corresponds to the blue/violet absorption interval, whose observed colour is yellow/orange, the colour of beer [99]. Positive association with the Vis loadings may therefore be linked to stronger absorption of light, which means darker colour; on the contrary, in the case of terms like *yellow*, *pale*, *straw* and *golden*, which are negatively correlated with the Vis loadings, this means that less light is absorbed, and therefore the observed colour should result less intense, as it is generally observed with the lagers and light beers linked to these terms.

3.2.3. "Booze"

3.2.3.1. Booze: NMR and PMD-booze. PMD-booze is very interesting because of its most important terms: *boozy*, *alcohol* and *syrupy*. According to Urban Dictionary [100], the top definition for the term *booze* in everyday English slang is: "*An alcoholic beverage, specifically any type of beer. It doesn't matter which* […]". This suggests that the information captured by PMD-booze may be related both to the sensory-like detection of alcohol and to drinking aimed to drunkenness. No meaningful interpretation for terms like *hidden* or *hide* could be found.

The NMR dataset performs quite well with PMD-booze, with a common component correlation of 0.74. However, the strongest beers in the dataset do not end up at positive scores, as the loading sign of the terms *boozy* and *alcohol* may suggest (Fig. 11c). A quite strong association with polyphenols was however found, and since the source of most polyphenols in beer is barley [101], it is possible that the *syrupy* term is related to the sweet/malty taste of beer. However, no terms such as *sweet*, *malty* or *barley* are associated with this topic, therefore this link is just hypothetical.

3.2.3.2. Booze: NIR and PMD-booze. The same situation is found with

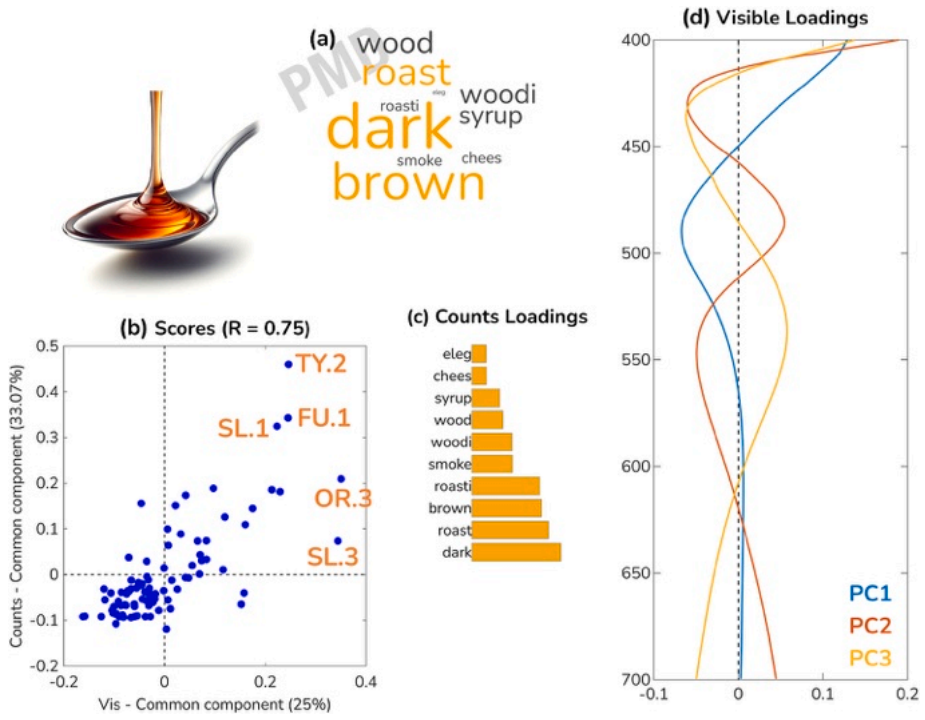


Fig. 8. Brown colour: Vis and PMD-brown colour (10 terms). PCA-GCA results obtained by comparing the Vis dataset and PMD-brown colour (a). The common scores are in (b), the Counts loadings are in (c), the Vis loadings are in (d). The Vis data were compressed by PCA, and 3 PCs were used for computing the PCA-GCA model; for this reason, there are three loading vectors in this figure. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

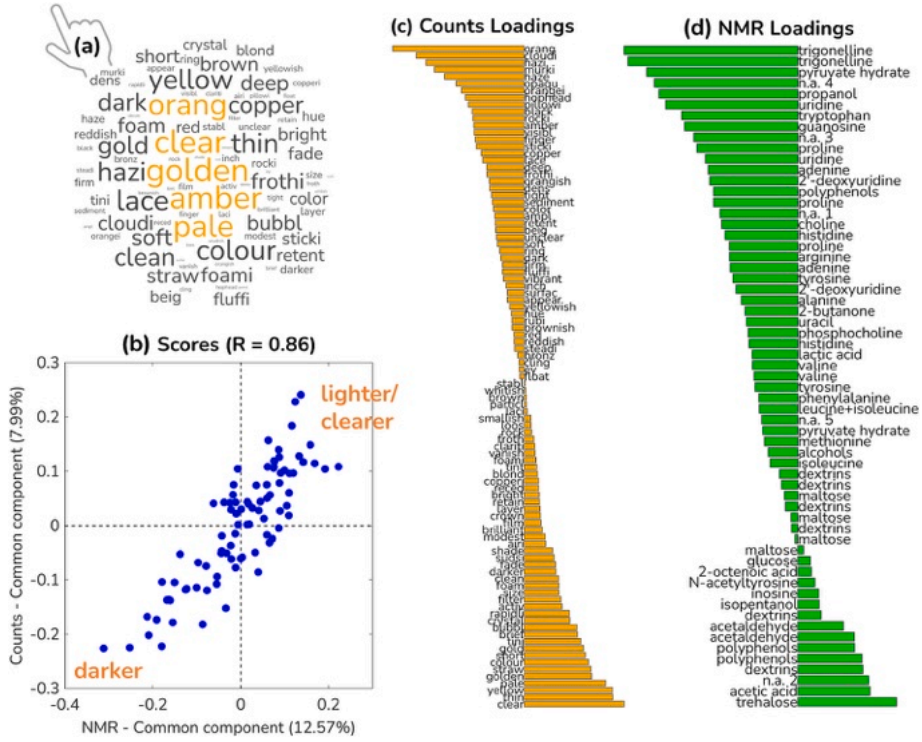


Fig. 9. Brown colour (broad meaning): NMR and appearance/colour terms (manually selected, 95 terms). PCA-GCA results obtained by comparing the NMR dataset and the appearance/colour terms (a). The common scores are in (b), the Counts loadings are in (c), the NMR loadings are in (d). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

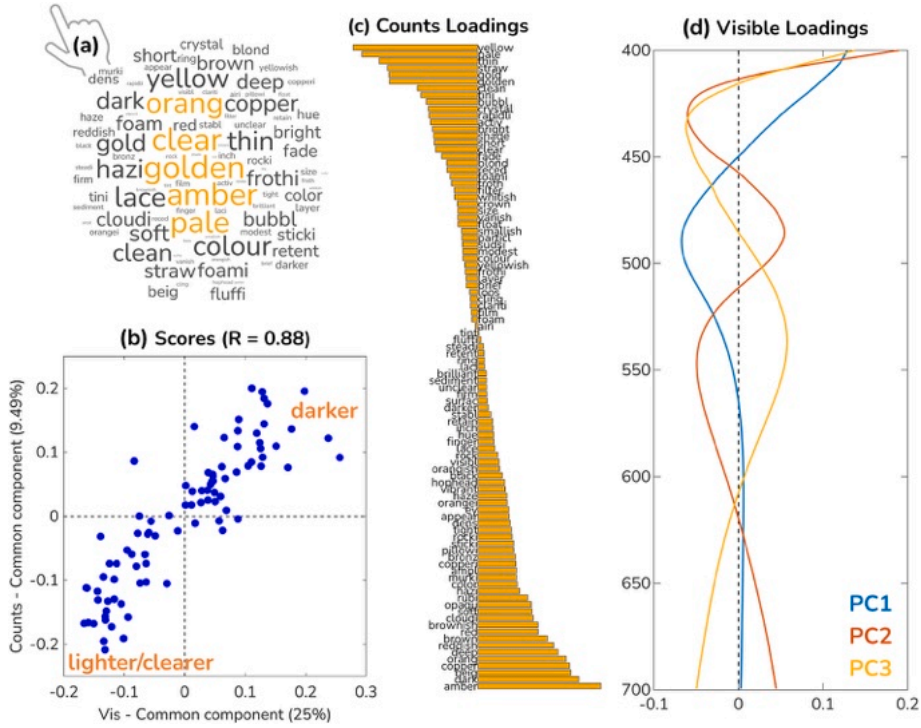


Fig. 10. Brown colour (broad meaning): Vis and appearance/colour terms (manually selected, 95 terms). PCA-GCA results obtained by comparing the Vis dataset and the appearance/colour terms (a). The common scores are in (b), the Counts loadings are in (c), the Vis loadings are in (d). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

the NIR dataset, with a common component correlation of 0.74. However, the samples' distribution suggests that this correlation is highly influenced by few samples, which are located at negative scores values (Fig. 12). These samples are FB.3 and TO.2 and are the strongest beers in the dataset (ABV respectively 10 % and 9.3 %). If these two samples are excluded, the common component correlation drops to 0.56. The rather

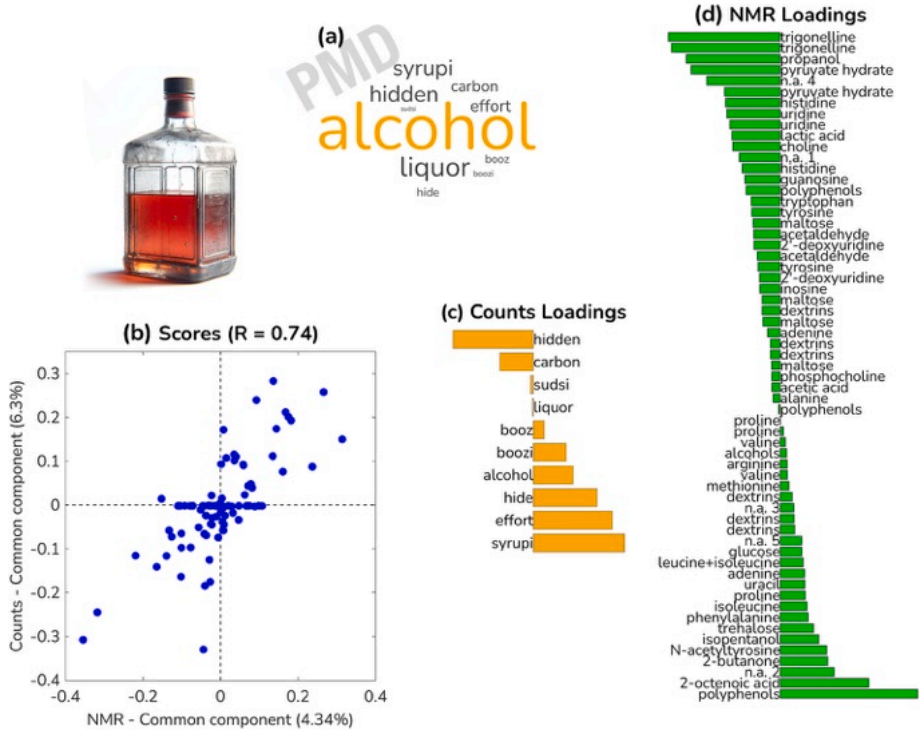


Fig. 11. Booze: NMR and PMD-booze (10 terms). PCA-GCA results obtained by comparing the NMR dataset and PMD-booze (a). The common scores are in (b), the Counts loadings are in (c), the NMR loadings are in (d).

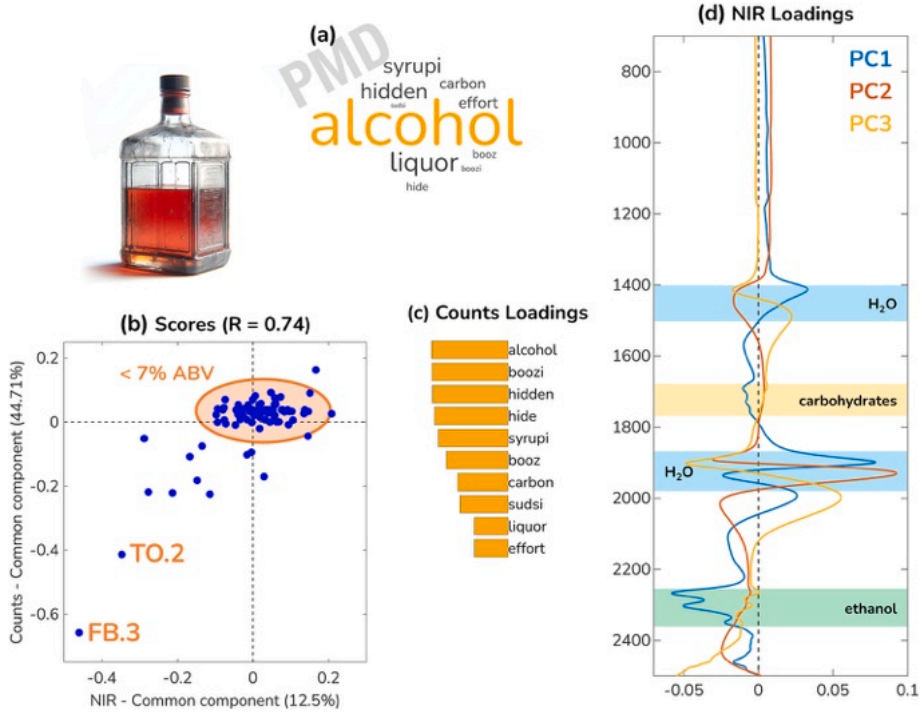


Fig. 12. Booze: NIR and PMD-booze (10 terms). PCA-GCA results obtained by comparing the NIR dataset and PMD-booze (a). The common scores are in (b), the Count loadings are in (c), the NIR loadings are in (d). The NIR data were compressed by PCA, and 7 PCs were used for computing the PCA-GCA model. For the sake of clarity, only the first three loading vectors are depicted together with general assignments of water, carbohydrates, and ethanol.

grouped set of samples close to the origin of the score plot represent the bulk beers that have ethanol content lower than 7 %. This unbalanced situation is probably the direct result of having very few extreme samples and a substantial group of "average" samples. However, the fact that terms like *alcohol* and *boozy* are the most related to these samples and the NIR signals of ethanol (they both have negative loadings) suggests that this correlation may exist. The inclusion of samples with higher ABV% values could demonstrate this, as it is evident that the original study sampling is quite unbalanced from this perspective.

3.2.4. "Refreshment"

PMD-refreshment is mainly characterized by terms like *lemony*, *chill*, *thirst*, *quencher* and *lime*. Hence, the name "Refreshment". Fresh and light beers can be found at positive scores, corresponding to the direction of these terms (Fig. 13b). Metabolites such as acetaldehyde, dextrins and trehalose also share this direction, as opposed to the strongly hops-related metabolites trigonelline, propanol and pyruvate. Acetaldehyde is a key component of lemon [102], and may be associated to freshness. Samples like LE.1 - "Sommerøl" (= Danish for "summer beer"), MA.4 - "San Miguel Fresca" (= Spanish for "fresh"), TO.4 - "Sun Dancer" and NO.2 - "Lemon Ale" are found at positive scores, in line with this "freshness" trend. In slight opposition to these groups of freshness-related terms is *zesty*, a term generally related to the citrus flavour, but also very used in the beer flavour description in association with the flavour of hops.

At negative scores in Fig. 13b are most ales and IPAs, in the same direction as trigonelline, but also propanol, which is linked to *alcohol*, *ripe*, *fruit* aromas [103]. If PMD-refreshment is about getting refreshment by looking for fresh, lemony flavours, IPAs and ales do not fit for this purpose, being more spiced and stronger in general (higher ABV, but also richer in flavour).

4. Conclusions and further developments

In the present study a novel approach to integrate and compare the information content of text corpora with analytical instrumental responses by means of chemometric methods was presented, applied and discussed. The general aim of our work was to assess the links between the analytical information and the user-generated descriptions of a set of beer samples. From the point of view of text data, subsets of meaningful terms were selected employing both automatic (PMD, for the definition of a number of "topics") and manual approaches, with different outcomes. The obtained results showed interesting correlations between the chemistry of beer and the words people use to describe their beer taste experiences. A summary of the correlation findings is given in Table 1.

From the point of view of the spectral data, the NIR dataset proved to be the least informative in relation to the text data. Unexpectedly, the NMR dataset showed correlations with many analysed topics and subsets. In some cases, such as with PMD-brown colour, the visible dataset showed a stronger correlation with the topic's information compared to the NMR dataset. However, even if the correlation improved, interpreting the Vis loadings is more complex, since the visible bands cannot be directly related to specific chemical compounds.

Regarding the automatic topics extraction, it is interesting to notice that the PMD selection procedure often produced subsets of the (generally larger) manually selected sets. Since in many cases comparable results were obtained, this could mean that an automated procedure applied to different/larger datasets of chemical characterizations and textual reviews could be fruitfully processed in an automated way using the PMD method for topic extraction. Considering the other topics (not present in Table 1), even though many of the twenty PMD-extracted topics made sense, no significant correlation with any of the spectral datasets was found. This part of distinct information surely deserves to be more deeply investigated.

Based on the present results, different directions may be taken. For instance, other automatic topic extraction methods may be evaluated, potentially using PMD topic extraction as a benchmark reference. Moreover, many topics may be fused based on the computation of a correlation index. Combinations of the spectral dataset through low- or mid-level data fusion approaches may be worth investigating.

Considering the dataset's limitations, future developments should be based on a more homogeneous beer dataset. The aim could be expanding the view, either by collecting a much larger pool of beer samples (with particular attention to balancing the beer styles, production sites, producers, and ABV content) and replicate the whole study on a larger scale, and/or by gathering a larger text dataset to study how the current dataset is related to the "rest of the world" of beer, from the point of view of the consumer. Another interesting addition to this study would surely be the generation and integration of sensory evaluations produced by a trained panel of experts.

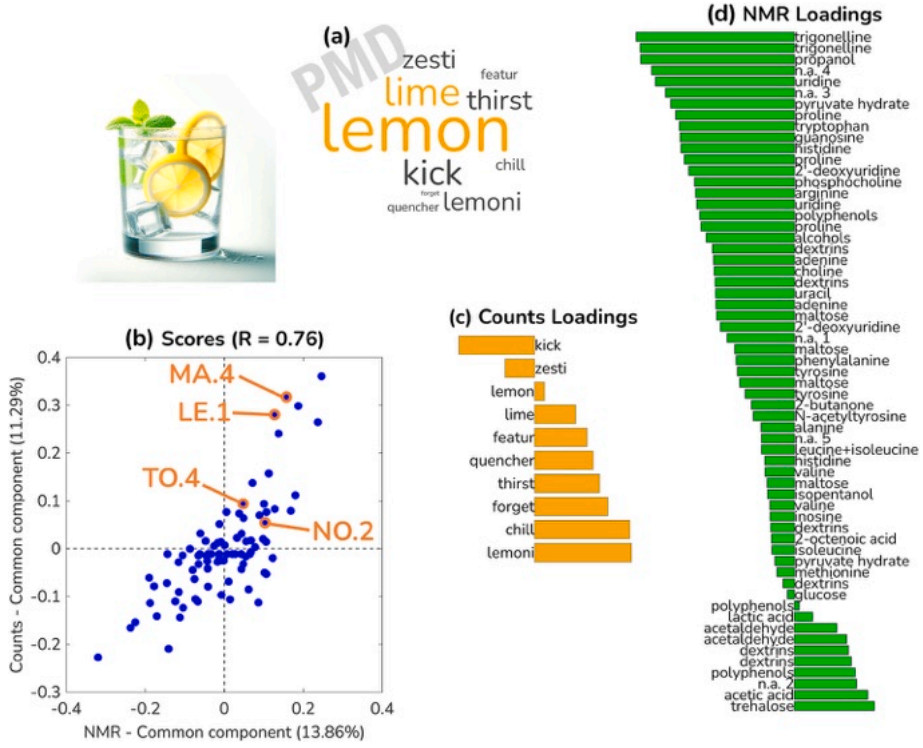


Fig. 13. Refreshment: NMR and PMD-refreshment (10 terms). PCA-GCA results obtained by comparing the NMR dataset and PMD-refreshment (a). The common scores are in (b), the Counts loadings are in (c), the NMR loadings are in (d).

Finally, improvements to the text analysis procedure should be investigated as well, focusing on further refinement of the text data of the present study. For instance, *n*-grams instead of unigrams may be used for creating the bags-of-words, so that the relations linking more words can also be included in the text data for modelling, leading to more precise characterizations of the beer samples. The methods of sentiment analysis are also an interesting direction to explore.

We think that the proposed approach can be potentially extended to and tested in any context where it is necessary to integrate subjective textual information with objective analytical data, for instance medical diagnosis with clinical data, or expert opinions with factual data.

CRediT authorship contribution statement

Nicola Cavallini: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Francesco Savorani:** Writing – review & editing, Visualization, Validation, Supervision. **Rasmus Bro:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Methodology, Funding acquisition, Conceptualization. **Marina Cocchi:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT by OpenAI to improve the readability of the manuscript, as all authors are not English native speakers. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article. The images in Figures from 5 to 13 depicting hops, syrup, a bottle of whiskey and a glass of fresh water were generated using Microsoft Designer's Image Creator. No graphs or plots related to the data or the results were generated using AI or AI-assisted tools.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Jose Camacho Paez is greatly thanked for his help and support in evaluating the different topics extraction approaches.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2025.105521.

Data availability

Data will be made available on request.

References

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

[2] J. Rice, Craft Rhetoric, Commun. Crit. Stud. 12 (2015) 218–222, https://doi.org/10.1080/14791420.2015.1014186.

[3] H. Everett, Craft Food and Drink: the Movement, Guard. Lib, Voice, 2014. https://guardianlv.com/2014/08/craft-food-and-drink-the-movement/.

[4] D.W. Murray, M.A. O'Neill, Craft beer: penetrating a niche market, Br. Food J. 114 (2012) 899–909, https://doi.org/10.1108/00070701211241518.

[5] K.G. Elzinga, C.H. Tremblay, V.J. Tremblay, Craft beer in the United States: history, numbers, and geography, J. Wine Econ. 10 (2015) 242–274, https://doi.org/10.1017/jwe.2015.22.

[6] M. Wolf, W. Ritz, S. McQuitty, Prosumers who home brew: a study of motivations and outcomes, J. Mark. Theory Pract. 28 (2020) 541–552, https://doi.org/10.1080/10696679.2020.1801321.

[7] C. Qian, S.I. Murphy, R.H. Orsi, M. Wiedmann, How can AI help improve food safety? Annu. Rev. Food Sci. Technol. 14 (2023) https://doi.org/10.1146/annurev-food-060721-013815.

[8] M. Goel, G. Bagler, Computational gastronomy: a data science approach to food, J. Biosci. 47 (2022) 1–10, https://doi.org/10.1007/S12038-021-00248-1.

[9] A. Singh, A. Glińska-Neweś, Modeling the public attitude towards organic foods: a big data and text mining approach, J. Big Data 9 (2022) 1–21, https://doi.org/10.1186/S40537-021-00551-6.

[10] K. Asseo, M.Y. Niv, Harnessing food product reviews for personalizing sweetness levels, Foods 11 (2022) 1872, https://doi.org/10.3390/foods11131872.

[11] G.P. Danezis, A.S. Tsagkaris, F. Camin, V. Brusic, C.A. Georgiou, Food authentication: techniques, trends & emerging approaches, TrAC, Trends Anal. Chem. 85 (2016) 123–132, https://doi.org/10.1016/j.trac.2016.02.026.

[12] S. Saadat, H. Pandya, A. Dey, D. Rawtani, Food forensics: techniques for authenticity determination of food products, Forensic Sci. Int. 333 (2022) 111243, https://doi.org/10.1016/j.forsciint.2022.111243.

[13] M. Gallo, P. Ferranti, The evolution of analytical chemistry methods in foodomics, J. Chromatogr. A. 1428 (2016) 3–15, https://doi.org/10.1016/j.chroma.2015.09.007.

[14] E. Hatzakis, Nuclear Magnetic Resonance (NMR) spectroscopy in food science: a comprehensive review, Compr. Rev. Food Sci. Food Saf. 18 (2019) 189–220, https://doi.org/10.1111/1541-4337.12408.

[15] J.U. Porep, D.R. Kammerer, R. Carle, On-line application of near infrared (NIR) spectroscopy in food production, Trends Food Sci. Technol. 46 (2015) 211–230, https://doi.org/10.1016/j.tifs.2015.10.002.

[16] S. Grassi, C. Alamprese, Advances in NIR spectroscopy applied to process analytical technology in food industries, Curr. Opin. Food Sci. 22 (2018) 17–21, https://doi.org/10.1016/j.cofs.2017.12.008.

[17] S. Lohumi, S. Lee, H. Lee, B.K. Cho, A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration, Trends Food Sci. Technol. 46 (2015) 85–98, https://doi.org/10.1016/j.tifs.2015.08.003.

[18] A.C. Power, J. Chapman, S. Chandra, D. Cozzolino, Ultraviolet-visible spectroscopy for food quality analysis, Eval. Technol. Food Qual. (2019) 91–104, https://doi.org/10.1016/B978-0-12-814217-2.00006-8.

[19] M. Bevilacqua, R. Bro, F. Marini, Å. Rinnan, M.A. Rasmussen, T. Skov, Recent chemometrics advances for foodomics, TrAC, Trends Anal. Chem. (2017), https://doi.org/10.1016/j.trac.2017.08.011.

[20] C.A. Hughey, C.M. McMinn, J. Phung, Beeromics: from quality control to identification of differentially expressed compounds in beer, Metabolomics 12 (2016) 11, https://doi.org/10.1007/s11306-015-0885-5.

[21] N. Cavallini, F. Savorani, R. Bro, M. Cocchi, A metabolomic approach to beer characterization, Mol 26 (2021) 1472, https://doi.org/10.3390/molecules26051472, 26 (2021) 1472.

[22] S. Wold, M. Sjöström, Chemometrics, present and future success, Chemom. Intell. Lab. Syst. 44 (1998) 3–14, https://doi.org/10.1016/S0169-7439(98)00075-6.

[23] J.M. Amigo, Data mining, machine learning, deep learning, chemometrics: definitions, common points and trends (Spoiler alert: VALIDATE your models!), Brazilian J. Anal. Chem. 8 (2021) 22–38, https://doi.org/10.30744/brjac.2179-3425.AR-38-2021.

[24] G. Donadini, M.D. Fumi, M. Lambri, A preliminary study investigating consumer preference for cheese and beer pairings, Food Qual. Prefer. 30 (2013) 217–228, https://doi.org/10.1016/j.foodqual.2013.05.012.

[25] I. Duarte, A. Barros, P.S. Belton, R. Righelato, M. Spraul, E. Humpfer, A.M. Gil, High-resolution nuclear magnetic resonance spectroscopy and multivariate analysis for the characterization of beer, J. Agric. Food Chem. 50 (2002) 2475–2481, https://doi.org/10.1021/jf011345j.

[26] S. Rossi, V. Sileoni, G. Perretti, O. Marconi, Characterization of the volatile profiles of beer using headspace solid-phase microextraction and gas chromatography-mass spectrometry, J. Sci. Food Agric. 94 (2014) 919–928, https://doi.org/10.1002/jsfa.6336.

[27] F.A. Iñón, S. Garrigues, M. de la Guardia, Combination of mid- and near-infrared spectroscopy for the determination of the quality properties of beers, Anal. Chim. Acta 571 (2006) 167–174, https://doi.org/10.1016/j.aca.2006.04.070.

[28] B.V. Humia, K.S. Santos, A.M. Barbosa, M. Sawata, M. da C. Mendonça, F. F. Padilha, Beer molecules and its sensory and biological properties: a review, Mol 24 (2019) 1568, https://doi.org/10.3390/MOLECULES24081568, 24 (2019) 1568.

[29] C. Gómez-Corona, M. Lelievre-Desmas, H.B. Escalona Buendía, S. Chollet, D. Valentin, Craft beer representation amongst men in two different cultures, Food Qual. Prefer. 53 (2016) 19–28, https://doi.org/10.1016/j.foodqual.2016.05.010.

[30] K. Christensen, K.H. Liland, K. Kvaal, E. Risvik, A. Biancolillo, J. Scholderer, S. Nørskov, T. Næs, Mining online community data: the nature of ideas in online communities, Food Qual. Prefer. 62 (2017) 246–256, https://doi.org/10.1016/j.foodqual.2017.06.001.

[31] G.D. Jacobsen, Consumers, experts, and online product evaluations: evidence from the brewing industry, J. Public Econ. 126 (2015) 114–123, https://doi.org/10.1016/j.jpubeco.2015.04.005.

[32] M.I. Betancur, K. Motoki, C. Spence, C. Velasco, Factors influencing the choice of beer: a review, Food Res. Int. 137 (2020) 109367, https://doi.org/10.1016/J.foodres.2020.109367.
[33] S. Pai, F. Brennan, A. Janik, T. Correia, L. Costabello, Unsupervised customer segmentation with knowledge graph embeddings, WWW 2022 - Companion Proc. Web Conf. (2022) 157–161, https://doi.org/10.1145/3487553.3524224, 2022.
[34] E.K. Clemons, G. Gao, L.M. Hitt, When Online Reviews Meet Hyperdifferentiation: a Study of the Craft Beer Industry, Routledge, 2014, pp. 149–171, https://doi.org/10.2753/mis0742-1222230207.
[35] J. McAuley, J. Leskovec, D. Jurafsky, Learning attitudes and attributes from multi-aspect reviews, Proc. IEEE Int. Conf. Data Min. ICDM (2012) 1020–1025, https://doi.org/10.1109/ICDM.2012.110.
[36] A.C. da Costa Fulgêncio, G.A.P. Resende, M.C.F. Teixeira, B.G. Botelho, M. M. Sena, Determination of alcohol content in beers of different styles based on portable near-infrared spectroscopy and multivariate calibration, Food Anal. Methods 1 (2021) 1–10, https://doi.org/10.1007/S12161-021-02126-W, 2021.
[37] S. Grassi, J.M. Amigo, C.B. Lyndgaard, R. Foschino, E. Casiraghi, Beer fermentation: monitoring of process parameters by FT-NIR and multivariate data analysis, Food Chem. 155 (2014) 279–286, https://doi.org/10.1016/j.foodchem.2014.01.060.
[38] S. Engelhard, H.-G. Löhmannsröben, F. Schael, Quantifying ethanol content of beer using interpretive near-infrared spectroscopy, Appl. Spectrosc. 58 (2004) 1205–1209, https://doi.org/10.1366/0003702042336000.
[39] H. Li, Y. Takahashi, M. Kumagai, K. Fujiwara, R. Kikuchi, N. Yoshimura, T. Amano, J. Lin, N. Ogawa, A chemometrics approach for distinguishing between beers using near infrared spectroscopy, J. Near Infrared Spectrosc. 17 (2009) 69–76, https://doi.org/10.1255/jnirs.830.
[40] V. Giovenzana, R. Beghi, R. Guidetti, Rapid evaluation of craft beer quality during fermentation process by vis/NIR spectroscopy, J. Food Eng. 142 (2014) 80–86, https://doi.org/10.1016/j.jfoodeng.2014.06.017.
[41] O. Klein, A. Roth, F. Dornuf, O. Schöller, W. Mäntele, The good vibrations of beer. The use of infrared and UV/Vis spectroscopy and chemometry for the quantitative analysis of beverages, Zeitschrift Für Naturforsch. B. 67 (2012) 1005–1015, https://doi.org/10.5560/znb.2012-0166.
[42] A. Biancolillo, R. Bucci, A.D.A.L. Magrì, A.D.A.L. Magrì, F. Marini, Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication, Anal. Chim. Acta 820 (2014) 23–31, https://doi.org/10.1016/j.aca.2014.02.024.
[43] M. Vasas, F. Tang, E. Hatzakis, Application of NMR and chemometrics for the profiling and classification of ale and lager American craft beer, Foods 10 (2021) 807, https://doi.org/10.3390/foods10040807.
[44] M. Kaufmann, K.J. Schwarz, A. Dallmann, T. Kuballa, M. Bergmann, 1H NMR spectroscopic discrimination of different beer styles combined with a chemical shift-based quantification approach, Eur. Food Res. Technol. 1 (2021) 1–11, https://doi.org/10.1007/S00217-021-03914-8, 2021.
[45] A. Palmioli, D. Alberici, C. Ciaramelli, C. Airoldi, Metabolomic profiling of beers: combining 1H NMR spectroscopy and chemometric approaches to discriminate craft and industrial products, Food Chem. 327 (2020) 127025, https://doi.org/10.1016/j.foodchem.2020.127025.
[46] L. Mannina, F. Marini, R. Antiochia, S. Cesa, A. Magrì, D. Capitani, A.P. Sobolev, Tracing the origin of beer samples by NMR and chemometrics: trappist beers as a case study, Electrophoresis 37 (2016) 2710–2719, https://doi.org/10.1002/elps.201600082.
[47] L.A. da Silva, D.L. Flumignan, A.G. Tininis, H.R. Pezza, L. Pezza, Discrimination of Brazilian lager beer by 1H NMR spectroscopy combined with chemometrics, Food Chem. 272 (2019) 488–493, https://doi.org/10.1016/j.foodchem.2018.08.077.
[48] T. Winograd, Understanding natural language, Cogn. Psychol. 3 (1972) 1–191, https://doi.org/10.1016/0010-0285(72)90002-3.
[49] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (Almost) from scratch, J. Mach. Learn. Res. 12 (2011) 2493–2537. http://www.jmlr.org/papers/v12/collobert11a.html.
[50] R.E. Banchs, Text Mining with MATLAB®, Springer New York, New York, NY, 2013 https://doi.org/10.1007/978-1-4614-4151-9.
[51] A. Hotho, A. Nürnberger, G. Paaß, F. Ais, A brief survey of text mining, in: http://www.crisp-dm.org/Process/index.htm, 2005.
[52] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. (2004), https://doi.org/10.1073/pnas.0307752101.
[53] D.M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (2009) 515–534, https://doi.org/10.1093/biostatistics/kxp008.
[54] A.K. Smilde, I. Måge, T. Naes, T. Hankemeier, M.A. Lips, H.A.L. Kiers, E. Acar, R. Bro, Common and distinct components in data fusion, J. Chemom. 31 (2017) e2900, https://doi.org/10.1002/cem.2900.
[55] D. Fox, A.W. Sahin, D.P. De Schutter, E.K. Arendt, Mouthfeel of beer: development of tribology method and correlation with sensory data from an online database, J. Am. Soc. Brew. Chem. 80 (2021) 112–127, https://doi.org/10.1080/03610470.2021.1938430.
[56] C. Price, BrewFinder – an Interactive Flavor Map Informed by Users, Springer, Cham, 2018, pp. 342–354, https://doi.org/10.1007/978-3-319-91521-0_25.
[57] C. Price, Georgia brew finder, (n.d.). http://www.gabrewfinder.com/(accessed July 23, 2025).
[58] M. Radovanović, M. Ivanović, Text mining: approaches and applications, Novi Sad J. Math. 38 (2008) 227–234. https://www.emis.de/journals/NSJOM/Papers/38_3/NSJOM_38_3_227_234.pdf.
[59] B. Pang, L. Lee, Opinion mining and sentiment analysis, found, Trends® Inf. Retr. 2 (2008) 1–135, https://doi.org/10.1561/1500000011.
[60] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: a survey, Ain Shams Eng. J. 5 (2014) 1093–1113, https://doi.org/10.1016/j.asej.2014.04.011.
[61] T.K. Landauer, P.W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse Process. 25 (1998) 259–284, https://doi.org/10.1080/01638539809545028.
[62] B. Braun, R. Timpe, Text Based Rating Predictions from Beer and Wine Reviews, 2015.
[63] N. Diakopoulos, D. Elgesem, A. Salway, A. Zhang, K. Hofland, Compare clouds: visualizing text corpora to compare media frames, in: Proc. IUI Work. Vis. Text Anal., 2015, pp. 193–202.
[64] A.W. Rivadeneira, D.M. Gruen, M.J. Muller, D.R. Millen, Getting our head in the clouds, in: Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI '07, ACM Press, New York, New York, USA, 2007, p. 995, https://doi.org/10.1145/1240624.1240775.
[65] A. Rajaraman, J.D. Ullman, Data mining, in: Min. Massive Datasets, Cambridge University Press, Cambridge, 2011, pp. 1–17, https://doi.org/10.1017/CBO9781139058452.002.
[66] W. Zhang, T. Yoshida, X. Tang, A comparative study of TF*IDF, LSI and multi-words for text classification, Expert Syst. Appl. 38 (2011) 2758–2765, https://doi.org/10.1016/j.eswa.2010.08.066.
[67] M. Van Zaanen, P. Kanters, Automatic mood classification using tf*idf based on lyrics, in: ISMIR, 2010, pp. 75–80. http://www.crayonroom.com/. (Accessed 15 February 2019).
[68] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831, https://doi.org/10.1039/C3AY41907J.
[69] S.C. Rutan, A. de Juan, R. Tauler, Introduction to multivariate curve resolution, in: Compr. Chemom., Elsevier, 2009, pp. 249–259, https://doi.org/10.1016/B978-044452701-1.00046-6.
[70] A. Cutler, L. Breiman, Archetypal analysis, Technometrics 36 (1994) 338–347, https://doi.org/10.1080/00401706.1994.10485840.
[71] M. Mørup, L.K. Hansen, Archetypal analysis for machine learning and data mining, Neurocomputing 80 (2012) 54–63, https://doi.org/10.1016/j.neucom.2011.06.033.
[72] R. Bro, E.E. Papalexakis, E. Acar, N.D. Sidiropoulos, Coclustering-a useful tool for chemometrics, J. Chemom. (2012), https://doi.org/10.1002/cem.1424.
[73] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022. http://www.jmlr.org/papers/v3/blei03a.html.
[74] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791, https://doi.org/10.1038/44565.
[75] N. Cavallini, F. Savorani, R. Bro, M. Cocchi, Fused adjacency matrices to enhance information extraction: the beer benchmark, Anal. Chim. Acta (2019), https://doi.org/10.1016/j.aca.2019.02.023.
[76] K. Varmuza, P. Filzmoser, Calibration, in: Introd. to Multivar. Stat. Anal. Chemom., CRC Press, 2009, https://doi.org/10.1201/9781420059496.ch4.
[77] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130, https://doi.org/10.1016/S0169-7439(01)00155-1.
[78] Y. Goldberg, O. Levy, word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. http://arxiv.org/abs/1402.3722, 2014.
[79] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. http://arxiv.org/abs/1301.3781, 2013.
[80] I. Måge, PCA-GCA on MATLAB file exchange, (n.d.). https://it.mathworks.com/matlabcentral/fileexchange/171089-pca-gca (accessed July 23, 2025).
[81] I. Måge, A.K. Smilde, F.M. van der Kloet, Performance of methods that separate common and distinct variation in multiple data blocks, J. Chemom. 33 (2019) e3085, https://doi.org/10.1002/CEM.3085.
[82] M.A. Drake, G.V. Civille, Flavor lexicons, Compr. Rev. Food Sci. Food Saf. 2 (2003) 33–40, https://doi.org/10.1111/j.1541-4337.2003.tb00013.x.
[83] L.J.R. Lawless, G.V. Civille, Developing lexicons: a review, J. Sens. Stud. 28 (2013) 270–281, https://doi.org/10.1111/joss.12050.
[84] M.C. Meilgaard, C.E. Dalgliesh, J.F. Clapperton, Beer flavour terminology, J. Inst. Brew. 85 (1979) 38–42, https://doi.org/10.1002/j.2050-0416.1979.tb06826.x.
[85] V. Daems, F. Delvaux, Multivariate analysis of descriptive sensory data on 40 commercial beers, Food Qual. Prefer. 8 (1997) 373–380, https://doi.org/10.1016/S0950-3293(97)00012-8.
[86] A. Kabakoff, Beer flavor wheel, (n.d.). https://www.beerflavorwheel.com/(accessed July 23, 2025).
[87] C. Schönberger, T. Kostelecky, 125th anniversary review: the role of hops in brewing, J. Inst. Brew. 117 (2011) 259–267, https://doi.org/10.1002/j.2050-0416.2011.tb00471.x.
[88] N. Rettberg, M. Biendl, L.A. Garbe, Hop aroma and hoppy beer flavor: Chemical backgrounds and analytical tools—A review, J. Am. Soc. Brew. Chem. 76 (2018) 1–20, https://doi.org/10.1080/03610470.2017.1402574.
[89] S.R. Palamand, J.M. Aldenhoff, Bitter tasting compounds of beer. Chemistry and taste properties of some hop resin compounds, J. Agric. Food Chem. 21 (1973) 535–543, https://doi.org/10.1021/jf60188a005.
[90] J. Kidrič, I.J. Košir, Characterization of the chemical composition of beverages by NMR spectroscopy, in: Mod. Magn. Reson., Springer Netherlands, Dordrecht, 2008, pp. 1597–1603, https://doi.org/10.1007/1-4020-3910-7_177.
[91] M.T. Ayseli, Y. İpek Ayseli, Flavors of the future: health benefits of flavor precursors and volatile compounds in plant foods, Trends Food Sci. Technol. 48 (2016) 69–77, https://doi.org/10.1016/j.tifs.2015.11.005.
[92] J. Zhou, L. Chan, S. Zhou, Trigonelline: a plant alkaloid with therapeutic potential for diabetes and central nervous system disease, Curr. Med. Chem. 19 (2012) 3523–3531, https://doi.org/10.2174/092986712801323171.

[93] N. Mohamadi, F. Sharififar, M. Pournamdari, M. Ansari, A review on biosynthesis, analytical techniques, and pharmacological activities of trigonelline as a plant alkaloid, J. Diet. Suppl. 15 (2018) 207–222, https://doi.org/10.1080/19390211.2017.1329244.
[94] K. Carbone, V. Macchioni, G. Petrella, D.O. Cicero, Exploring the potential of microwaves and ultrasounds in the green extraction of bioactive compounds from Humulus lupulus for the food and pharmaceutical industry, Ind. Crops Prod. 156 (2020) 112888, https://doi.org/10.1016/j.indcrop.2020.112888.
[95] A.R. Spevacek, K.H. Benson, C.W. Bamforth, C.M. Slupsky, Beer metabolomics: molecular details of the brewing process and the differential effects of late and dry hopping on yeast purine metabolism, J. Inst. Brew. 122 (2016) 21–28, https://doi.org/10.1002/jib.291.
[96] "Helping Hand" beer by Mikkeller, (n.d.). https://untappd.com/b/mikkeller-helping-hand/818743 (accessed July 23, 2025).
[97] S. Ohtake, Y.J. Wang, Trehalose: current use and future applications, J. Pharm. Sci. 100 (2011) 2020–2053, https://doi.org/10.1002/jps.22458.
[98] A.J. Buglass, M. Mckay, C.G. Lee, Beer, in: Handb. Alcohol. Beverages, John Wiley & Sons, Ltd, Chichester, UK, 2010, pp. 132–210, https://doi.org/10.1002/9780470976524.ch9.
[99] E. Pretsch, P. Bühlmann, M. Badertscher, UV/Vis spectroscopy, in: Struct. Determ. Org. Compd., Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 1–20, https://doi.org/10.1007/978-3-540-93810-1_9.
[100] "Booze" - definition from Urban dictionary, (n.d.). https://www.urbandictionary.com/define.php?term=Booze.
[101] N. Whittle, H. Eldridge, J. Bartley, G. Organ, Identification of the polyphenols in barley and beer by HPLC/MS and HPLC/electrochemical detection, J. Inst. Brew. 105 (1999) 89–99, https://doi.org/10.1002/j.2050-0416.1999.tb00011.x.
[102] M.G. Moshonas, P.E. Shaw, Analysis of flavor constituents from lemon and lime essence, J. Agric. Food Chem. 20 (1972) 1029–1030, https://doi.org/10.1021/jf60183a019.
[103] C. Sánchez-Estébanez, S. Ferrero, C.M. Alvarez, F. Villafañe, I. Caballero, C. A. Blanco, Nuclear magnetic resonance methodology for the analysis of regular and non-alcoholic lager beers, Food Anal. Methods 11 (2018) 11–22, https://doi.org/10.1007/s12161-017-0953-8.