Consumer Acceptance of Beer: An Automated Sentiment Analysis Approach

Ellise A. Canty

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Food Science and Technology

Jacob Lahne, Chair
Amanda Stewart,
Renee Boyer

June 24, 2022
Blacksburg, Virginia

Keywords: sentiment analysis, beer, consumer acceptance, free comment data

Consumer Acceptance of Beer: An Automated Sentiment Analysis Approach

Ellise A. Canty

ABSTRACT

Selecting the correct methodology to better understand how consumers perceive food products is a challenging task for the food industry and sensory researchers alike. Free comment tasks (FC) utilize the advantages of open-ended questions to generate intuitive comments from untrained consumers to help identify and describe sensory attributes of products. However, FC data is typically analyzed using text analysis done by hand and is very cumbersome to organize and interpret. There is a growing need and interest to add to the library of data analysis tools used to understand FC data and consumer acceptance studies. Sentiment analysis is an opinion mining tool commonly used in marketing and computer science that extracts the emotional valence of the author from an unstructured text in the form of a sentiment score. A few studies in sensory evaluation use lexicon-based sentiment analysis which has many drawbacks: it is time-consuming, requires a large amount of data and dictionaries need to be tailored for food. We used a deep learning sentiment analysis approach to analyze and predict consumer sentiment/acceptance. The research objectives of this study are 1) to explore quicker and automated methods of sentiment analysis to better understand and predict consumer acceptance, and 2) to examine the advantages and disadvantages of sentiment analysis as a data analysis tool in sensory evaluation. We avoided the pitfalls of creating a sentiment lexicon by using online beer reviews to train a word embedding model where all of the relevant words in the review are converted into vectors. We used the distance and similarity (clustering) of the vectors to determine taste/flavor attributes that correspond to negative and positive sentiment. Next, to validate and test our model we gathered FC data in a consumer acceptance study. Panelists (N=68) were presented with six beers, one at a time and were instructed to taste and smell before leaving comments. We performed sentiment analysis on the FC data, and we compared our deep learning sentiment analysis model with three other pre-existing sentiment analysis models: SentimentR, VADER, and Liu and Hu opinion lexicon. Our deep learning sentiment analysis model had the highest accuracy (69%) and precision rate (73%). Overall, our findings provide an early look into the advantages and disadvantages of sentiment analysis applied to FC data in sensory evaluation.

Consumer Acceptance of Beer: An Automated Sentiment Analysis Approach

Ellise A. Canty

GENERAL AUDIENCE ABSTRACT

It can be a challenging task for the food industry and sensory researchers to select the correct methodology to better understand how consumers perceive food products. One method is Free comment tasks (FC) which use open-ended questions to generate comments from consumers. However, FC data is typically analyzed using text analysis done by hand and is very cumbersome to organize and interpret. This thesis is interested in investigating the application of data analysis methods from computer science on FC data. Sentiment analysis is an opinion mining tool commonly used in marketing and computer science that finds the emotional tone of the author from a text in the form of a sentiment score. First, we created a deep learning sentiment analysis model which uses algorithms to find useful patterns in the text that indicate positive and negative sentiment with minimal human intervention. We were interested if there were any advantages in creating a model so we compared our model to three widely used sentiment analysis models: VADER, Liu Hu and Sentiment R. Next, to test our sentiment analysis model and the three others we gathered FC data in a consumer acceptance study. Panelists (N=68) were presented with six beers, one at a time and were instructed to taste and smell before leaving comments. The research objectives of this study are 1) to explore quicker and automated data analysis methods to understand FC data, and 2) to examine the advantages and disadvantages of data analysis tools from computer science in sensory evaluation. Our deep learning sentiment analysis model had the highest accuracy (69%) and precision rate (73%). Overall, our findings provide an early look into the advantages and disadvantages of sentiment analysis applied to FC data in sensory evaluation.

Acknowledgements

This thesis would not have been possible if not for the insight and emotional support of my advisors Dr. Jacob Lahne, Dr. Amanda Stewart and my committee member, Dr. Renee Boyer. I am also immensely grateful for Postdoctoral Associate Dr. Marlon Ac-Pangan who provided insight and expertise in the VT sensory lab which greatly assisted my research. Lastly, I would like to extend my gratitude to my family, and friends whose generosity and guidance are with me in whatever I pursue.

Table of Contents

Chapter 1: Introduction *1*

Chapter 2: Literature Review *3*

2.1. Sensory Food Science **3**

2.2. Consumer Acceptance **3**

2.3. Free comments **4**

2.4. Big Data **5**

2.5. Sentiment Analysis **8**

2.6. Deep learning method **11**

2.7. Survey of SA in sensory evaluation **11**

References **13**

Chapter 3: Experimental *19*

3.1 Introduction **19**

3.2 Materials and Methods **20**

3.2.1. Sentiment Analysis 20

3.2.2. SentimentR, VADER, and Liu Hu Sentiment Analysis Models 21

3.2.3. LIME Model 22

3.2.4. Free Comment Task 23

3.2.4.1. Samples 23

3.2.4.2. Panelists 23

3.2.4.3. Free Comment Task and Questionnaire Procedure 24

3.3. Statistical Analysis **24**

3.4. Results **25**

3.4.1. Linear Regressions 25

3.4.2. Accuracy, Recall, and Precision Tests 26

3.4.3. ANOVAs 27

3.5. Discussion **27**

References 32

Chapter 4: Conclusions and Future Work *35*

Figures, Tables, and Appendix *36*

Figures **36**

Tables **41**

Appendix **47**

1.Pre-Screener 47

2.Data Collection Template 48

Chapter 1: Introduction

Sensory food science is concerned with understanding and analyzing the way we perceive and our food through the primary five senses. Sensory evaluation is the method of analysis in sensory food science. Sensory evaluation is defined as a *scientific method used to evoke, measure, analyze, and interpret those responses to products as perceived through the senses of sight, smell, touch, taste, and hearing* (see Lawless & Heymann, 2010). The need for sensory evaluation in the food industry was driven by the challenge of attempting to analytically measure human responses to foods, with minimal interference from the potentially biasing effects of brand identity (Lawless & Heymann, 2010). The central principle or guiding "central dogma" for all sensory evaluation is that the chosen methodology should be appropriate to meet the objectives of the test.

A free comment (FC) task is used to understand consumer perceptions through the direct analysis of their written words. These methods use open-ended questions to generate intuitive comments from consumers to help identify and describe sensory attributes of products (Symoneaux et al., 2012). FC data is typically analyzed using text analysis done by hand or with the support of coding software such as R software (R core team, 2013) to create contingency tables which are analyzed using chi-squared (Fonseca et al., 2016a; Mahieu et al., 2020). These current techniques can be very labor intensive and time-consuming. First, to generate data, appropriate trained or untrained subjects need to be selected, and samples must be purchased to conduct a free comment task. Secondly, the data must go through several transformations before it is ready to be analyzed.

Recently, sensory scientists have begun exploring alternative methods that further address the drawbacks. There is a growing repository of online food reviews which can be considered "big data." These online reviews may have the same potential for analysis as free comment data collected from a traditional sensory study. Additionally, sentiment analysis could be used to analyze free comment data. Sentiment analysis is an opinion mining tool used to identify opinions or tone of a given text. This method is frequently used on short unstructured texts like tweets from Twitter and online reviews from interactive forums. There are many different approaches to sentiment analysis. We created a sentiment analysis model using a deep learning approach. Deep learning (DL) is within the umbrella of AI, but it is more closely related to machine learning. In deep learning methods an algorithm learns to recognize novel patterns or useful representations in the data. There is a focus on learning "successive layers of increasing meaningful representations" (Chollet & Allaire 2018).

First, we developed a deep learning sentiment analysis model using a template provided by 'Deep Learning with R' (Chollet & Allaire 2018). We trained our deep learning sentiment analysis model using previously collected online beer reviews and compared our model to three others: VADER, Liu Hu and SentimentR. Our training data set was a subset of online beer reviews collected by McAuley (2012) from the beer-rating website RateBeer.com. The dataset we have was collected from Apr 2000 - Nov 2011 and consists of 2,924,127 reviews, 40,213 users of which 4,798 of them had more than 50 reviews. We obtained the data from Stanford University SNAP library (accessed here

https://snap.stanford.edu/about.html). Second, to help validate and test sentiment analysis as an effective data analysis method we conducted a consumer acceptance study in our sensory lab to generate FC data. Our samples were 6 beers: three from our training data set and three locally recommended beers. Panelists were presented with one sample at a time and were instructed to smell and taste the sample prior to answering the questions. Panelists also gave comments and rated all of the samples.

This study will explore the use of sentiment analysis for free comment data and compare the results to more traditional data collection methods. Specifically, the research objectives of this study are 1) to explore quicker and automated methods of sentiment analysis to better understand and predict consumer acceptance, and 2) to examine the advantages and disadvantages of sentiment analysis as a data analysis tool in sensory evaluation. Sentiment analysis is a new and underutilized data analysis method in a sensory scientist's toolkit. These findings have the potential to help easily organize and interpret the somewhat informal and ambiguous sensory language untrained consumers use when rating and discussing sensory characteristics of food. These results may help form a preliminary framework of the effectiveness of sentiment analysis in future free comment tasks and consumer acceptance studies. This knowledge may be beneficial in learning the advantages and disadvantages of sentiment analysis in sensory evaluation and identifying consumer preference and acceptance.

Chapter 2: Literature Review

2.1. Sensory Food Science

Sensory food science is a multidisciplinary science within the umbrella of food sciences. Sensory food science analyzes human sensory perceptions and affective responses to foods. Sensory evaluation method of analysis in sensory food science. Sensory evaluation is defined as a *scientific method used to evoke, measure, analyze, and interpret those responses to products as perceived through the senses of sight, smell, touch, taste, and hearing* (see Lawless & Heymann, 2010). Sensory evaluation arose to address the need to measure human responses to foods and reduce the potentially biasing effects of brand identity and other influences on consumer perception (Lawless & Heymann, 2010). The scope of sensory evaluation has grown to help meet the demands of modern product development and intense competition within the food industry which requires a comprehensive understanding of the sensory aspects of foods, and sensory techniques. The central principle or guiding "central dogma" for all sensory evaluation is that the chosen methodology should be appropriate to meet the objectives of the test.

Sensory tests can be widely divided into two classes: analytical and affective. Analytical sensory tests such as discrimination and descriptive methods and affective or hedonic tests which include assessing consumer liking or preferences (Lawless & Heymann, 2010). Analytical sensory tests require a more "analytical" mindset. The job of the panelists is to focus on specific aspects of the product. For example, a discrimination test is used to determine whether any perceptible differences exist between two types of products. It asks the question "are these products different in any way?" Descriptive tests are used to quantify the intensity of a sensory characteristics in a product. It asks the question: "how does one sensory characteristic differ between these products?' Panelist are screened for sensory acuity and have to receive training (sometimes extensively) prior to testing (Lawless & Heymann, 2010).

In contrast, affective tests ask participants to also consider their personal feelings and perspectives on the product. For example, an affective test is used to quantitatively determine the degree of liking or disliking of a product. It asks the question: "how well are these products liked or which products are preferred?" Analysis is usually based on a hedonic scale (Lawless & Heymann, 2010).

2.2. Consumer Acceptance

Consumer acceptance is a prominent concern for the modern food industry (Prescott et al., 2014). Consumer acceptance is utilized to scale the degree of acceptability of foods. A consumer acceptance test has the ability to tell the researcher whether a food product is strongly liked or disliked (Lawless & Heymann, 2010a).

Many affective tests have been developed to gain consumer insight by identifying whether the consumers like the product, prefer it over another product, or find the product acceptable based on its perceived sensory characteristics. Affective testing can be further divided into consumer preference and consumer acceptance. Consumer acceptance tests

which are relevant to this thesis, are methods for scaling the degree of acceptability of foods (Lawless & Heymann, 2010). Early consumer acceptance evolved from a need to understand consumer liking of food products. In the early 1940s U.S solders were forgoing their food rations and it was negatively impacting their performance. In 1944, the Army Quartermaster Subsistence Research and Development Laboratory in Chicago, Illinois, established the first Food Acceptance Research Branch for the purpose of providing reliable and valid prediction of the acceptability of various food products and rations (Peryam & Girardot, 1952). One of the major contributions to sensory science from the Food Acceptance Branch was the development of the nine-point hedonic scale. The development of the scale began in 1949 by Peryam and Norman Girardot (Peryam & Girardot, 1952; Meiselman & Schutz, 2013).

Scales are typical analytical tools in acceptability tests, but the nine-point hedonic scale is one of the most widely used. This nine-point hedonic scale revolutionized the way sensory scientists conduct consumer acceptance studies by providing a standardized scale that is easy for consumers to understand (Lawless & Heymann, 2010). Currently, the nine-point has been translated into many different languages and is still a commonly used scale in consumer acceptance studies in the world (Meiselman & Schutz, 2013). The scale can be used vertically or horizontally and responses on this scale are usually numbers from 1 to 9. 1 represents dislike extremely and 9 represents like extremely. The hedonic scale assumes consumer preferences exist on a unidimensional trajectory and that preferences can be binaurally categorized into like and dislike (Lawless & Heymann, 2010).

The nine-point scale is a great tool to use to understand consumers' global like and dislike of a product. For example, the nine-point scale would be an appropriate tool to help determine which apple juice is more acceptable to consumers. However, the scale doesn't tell the researcher the consumers' reaction to the specific sensory characteristics of the apple juice such as: sweetness, sourness, aroma etc. Alternative acceptability scales to the traditional nine-point scale were developed to give the researcher more information on which attributes are liked and disliked (Lawless & Heymann, 2010). For example, the just-about-right (JAR) scale combines the measurements of attribute intensity and consumer acceptability to determine if an attributes intensity is at an optimal level (Moskowitz et al., 2008). The JAR scale is bipolar, the opposite ends of the scale usually represent the highest intensity of a specific attribute such as "Too little" and "Too much." The midpoint indicates the ideal acceptability of the attribute and can be labeled as "just right" or "just about right" (Lawless & Heymann, 2010b; Rothman & Parker, 2009).

2.3. Free comments

Selecting the correct methodology to better understand how consumers perceive food products is a challenging task for the food industry and sensory researchers alike. For example, a company wants to substitute an ingredient for a cheaper alternative, but still maintain the signature taste and quality known and liked by consumers. A consumer preference test may be useful in determining which formulation is preferred but a discriminatory test would be used to differentiate whether there is a discernable

difference between the formulations. In this example a discriminatory test would be best suited to address the concerns of the company. Consumer input is an important aspect in the development, marketing, quality control and potential reformulation of existing products (Lawless & Heymann, 2010).

In recent years, consumer-oriented methods have evolved to meet the demand of the modern food industry and the expectations of the consumer. For example, previous studies have attempted to understand consumer perceptions through the direct analysis of their written words. These methods use open-ended questions to generate intuitive comments from untrained consumers to help identify and describe sensory attributes of products (Symoneaux et al., 2012).

Free comment (FC) utilizes the advantages of open-ended questions. In a FC task consumer are instructed to comment and describe the food product without heavy oversight from the researcher (ten Kleij and Musters, 2003). FC has been used to investigate consumer acceptance, and the hedonic value of a wide variety of food and beverage products (Fonseca et al., 2016; Lahne et al., 2014; Luc et al., 2020; Mahieu et al., 2020; Symoneaux et al., 2012). Furthermore, a FC task does not require a trained panel (ten Kleij & Musters, 2003), but depending on the aim of the study a FC task can be effectively used in describing products with both trained and untrained consumers (Lahne et al., 2014; Mahieu et al., 2020; ten Kleij & Musters, 2003).

FC data is typically analyzed using text analysis done by hand or with the support of coding software such as R software (R core team, 2013). The data is first cleaned for spelling, grammar errors, and irrelevant words. Then, the data is lemmatized by grouping together words with the same root or lemma. For example, if "plays, playing, and played" were all mentioned in one review those words would be reduced to their root "play." Once the data is transformed into sensory descriptors and terms, the words are grouped together usually by hedonic value (sensory attributes that indicate the products were liked or disliked) or by similarity (words describing the same sensory characteristics). Lastly, these groups are used to create a contingency table and are analyzed using chi-squared (Fonseca et al., 2016a; Mahieu et al., 2020).

2.4. Big Data

The growing prevalence of online grocery shopping, food delivery services and the rise of social media reflect emerging changes in the consumer dining experience (Fonseca et al., 2016; Lahne et al., 2014; Luc et al., 2020; Mahieu et al., 2020; Snuggs & McGregor, 2021; Symoneaux et al., 2012; Z. Zhang et al., 2010). The internet has evolved into a diverse medium for food acquisition, consumer spending habits, acceptance of food products and food-related social interactions. There is an expanding repository of "big data" available on social media, food delivery apps, and online grocery stores. Online customer reviews may be valuable, but unfortunately, they are a widely untapped resource for understanding food related consumer sentiment in sensory research. For the purposes of this study, online customer reviews are broadly defined as a person or customer's thoughts/opinions and experiences in an online review or comment section.

For example, food delivery apps such as UberEats, GrubHub, and Yelp.com allow its users to make ratings and leave text reviews about the hedonic value of the food as well as the quality of restaurant service and delivery. UberEats hosts an impressive 66 million annual users in 2020 (*Uber Eats Revenue and Usage Statistics (2021)*, 2020). Grubhub hosts 31.4 million annual users in 2020 (*Grubhub Revenue and Usage Statistics (2021)*, 2020).

Recently, there is a growing interest to explore this "big data" as a research tool in sensory and consumer science. For example, researchers have started to explore the viability of new data sources such as food-related social media (Carr et al., 2015; Hamilton & Lahne, 2020; Miller et al., 2021; Tao et al., 2020; Tian et al., 2021; Vidal et al., 2015).Tian et al., (2021) extracted 175,879 text-based Yelp restaurant reviews and performed a lexicon-based sentiment analysis to measure food-related consumer sentiments and affective responses from online review data. They found that online review data can provide comprehensive information on consumer behavior research. Vidal et al., (2015) contributed to the evaluation of Twitter as a research tool on food-related consumer behavior, by retrieving a subset 16,000 tweets relating to "what people say when they tweet about different eating situations." The study revealed that generally the tweets about eating situations included information which can contribute to our understanding of food choice decisions and how eating patterns are shaped. Using, a subset of 1,125,458 restaurant reviews, extracted from the Yelp.com open dataset, which is composed of over 8 million reviews, Ashgar (2016) created sixteen prediction models to determine the best model for predicting the ratings from reviews.

These studies aren't exclusive to just text data. Researchers are also looking at visual symbols such as emojis and other images, as tokens of sentiment (Jaeger et al., 2021; Schouteten et al., 2018, 2019; Swaney-Stueve et al., 2018; Vidal et al., 2016, 2020). Schouteten et al., (2018) evaluated the use of emojis in assessing children's emotional evaluation of food products. The children were tasked with selecting any emoji from a set of 33 that, related to consuming a biscuit. They panelists selected between 1 and 18 emojis that related to consuming a specific sample. On average, 10% of the emojis were used for each sample. The authors concluded that emojis can be used to obtain discriminating emotional profiles between food samples of the same food product category and may help to indicate consumer acceptance in children. In a follow-up study Schouteten et al., (2019) tackled the question "would a standardized or a product-specific emoji list perform better at evaluating the consumer acceptance of children?" Researchers compared the product-specific list of emojis to a standard set of 33 emojis and found that the product-specific list was able to better discriminate between product samples.

Researchers in marketing, computer science and business disciplines have utilized online reviews as an important data source that reflects customers' experiences and the evaluation of products. For example, online grocery shoppers that utilize Amazon Fresh or Target have the option of leaving ratings and reviews for the food and beverages they purchase. Fang & Zhang (2015) extracted over 5.1 million product reviews from Amazon.com in 2014 and proposed an algorithm to identify negative phrases in their sentiment analysis to address the sentiment polarity categorization problem. It was found

that the algorithm was able to identify 21,586 different phrases with total occurrence of over 0.68 million, each of which has a negative prefix.

McAuley et al., (2012), collected online written reviews from the beer-rating websites BeerAdvocate.com and RateBeer.com which allow users to rate beers using a five-aspect rating system. They introduced a new sentiment analysis model, the Preference and Attribute Learning from Labeled Groundtruth and Explicit Ratings or PALE LAGER. The PALE LAGER can predict ratings based on the words the describe multiple aspects of the product and the associated sentiment. Additionally, the PALE LAGER model was able to successfully predict ratings and outperform simple Support Vector Regression baselines. Using the same dataset, researchers developed more statistical models to better understand rating dimensions (McAuley & Leskovec, 2013).

There are many advantages of using food-related online review data to examine consumer dining sentiments. First, there is a large amount of data that is readily available. For example, RateBeer.com (2020) is one of the largest beer review forums which catalogues user-driven surveys and reports of beers from around the world. The forum regularly has 20,000 unique daily visitors and millions of members that participate in the forum each month. RateBeer.com began in May 2000 and now has an impressive library of 9,529,839 reviews. Likewise, Yelp.com allows users to participate in discussions based on local restaurants, food, and service. As of December 2020, Yelp states they have over 224 million reviews and 31 million monthly desktop users (*Yelp - Company - Fast Facts*, n.d.). Additionally, Amazon.com one of the most widely used e-commerce platforms which allows users to rate the acceptability of food products. For example, a dataset collected by Kaggle consists of over ~500,000 reviews of fine foods from amazon from October 1999 to October 2012 (*Amazon Fine Food Reviews*, n.d.; McAuley & Leskovec, 2013).

Secondly, as a qualitative research tool, mining for online reviews is drastically less cost prohibitive than traditional focus groups in the U.S. and Europe. The average trained sensory panel requires a large number of participants and several hours to properly train them to recognize the target sensory attributes. Screening, buying and prepping samples can cost several thousands of dollars. If a company has multiple products in very different food categories, this process could become very expensive and time-consuming (Meiselman & Schutz, 2013).

Thirdly, online reviews are typically conducted in a natural eating environment that the consumer is familiar and comfortable with in a specific physical (ex: at home, dining hall) and social context (ex: formal family dinner, eating with peers) (de Graaf et al., 2005; De Wijk et al., 2019; Meiselman, 1992), unlike lab tests or central-location tests (CLT) which are carried out in a standardized location under controlled conditions. Furthermore, CLT conditions differ from natural eating environments in the amount of food consumed. In a CLT the consumer is given a small portion-controlled sample whereas in natural eating environment consumers usually eat *ad libitum* and in larger portions. Also, in a natural eating environment consumers control when they choose to eat and in what manner they eat their food. In contrast, in a CLT the timing and

presentation of the samples are out of the consumers control as well as how the sample is prepared and given to them to eat (Boutrolle et al., 2007; Meiselman, 1992). Given the sterile environment of a CLT, it has been assumed that a natural eating environment would produce more relevant hedonic data and insight on consumer acceptance (Niimi et al., 2022). For example, data extracted from a natural social media environment was found to contain more detailed descriptions of food products compared to the focus group (Carr et al., 2015). Additionally, (Kozlowska et al., 2003) compared hedonic tests of apple juice conducted in an CLT environment and in a natural eating environment in a university common room and they found that the hedonic tests from the natural eating environment had higher scores.

2.5. Sentiment Analysis

Sentiment analysis (SA) is "the computational study of people's opinions, appraisals, attitudes, and emotions toward entities and their aspects." The entities can represent products, services, organizations, individuals, events, etc. The aspects are components of the entities which commonly found in reviews (L. Zhang & Liu, 2014). Early sentiment analysis began in the late 1950s by Osgood et al., (1957) in series of pioneering studies called "the measurement of meaning." The authors were interested in how to measure meaning or affective responses in a given text and developed a technique called semantic differential to address the problem (Osgood et al., 1957). Currently, SA utilizes approaches from natural language processing (NLP) which is a subfield of artificial intelligence (AI) concerned with the analysis and interpretation of human languages.

The goal of SA is to identify the authors' opinions or tone about specific topics within an unstructured text (Medhat et al., 2014). The sentiment is commonly measured by a value called polarity. Polarity can be classified as positive, negative or neutral and the value of the polarity influences the overall tone of the text or sentiment. In relation to this thesis the tone is represented by a number called a "sentiment score" (L. Zhang et al., 2014). When the emotional tone present in a text is positive, the coordinating sentiment score will be positive/higher. When the emotional tone present in a text is negative, the coordinating sentiment score will be negative/lower. SA techniques provide a useful method to capture and measure individual opinions and sensory experiences which may contribute towards consumers' hedonic liking and acceptance (L. Zhang et al., 2014; Medhat et al., 2014).

SA techniques can be divided into two main approaches categorized into lexicon-based methods and machine learning methods (L. Zhang et al., 2014; Medhat et al., 2014; Taboada et al., 2011; Chen & Zimbra, 2010).

Lexicon-based methods use a pre-defined list of words with sentiment scores attached to give sentiment scores to text, whereas machine-learning methods use a variety of optimization techniques to derive sentiment from the texts themselves. This approach recognizes sentiment using premade or manually curated sentiment dictionaries. A given sentiment dictionary will contain information about the sentiment or polarity of the words of a given language and assign weights or valence to the words and grammatical rules in

the text. Usually, a dictionary will assign a score for each word (Liu & Zhang, 2012; Visalli et al., 2020). A popular English sentiment analysis dictionary is SentiWordNet which assigns three sentiment scores: positive, negative and objective (Esuli, 2019/2021).The advantage of a lexicon-based SA is it's a relatively simple model, and computationally efficient. Thus, they are frequently used to solve general sentiment analysis problems (Medhat et al., 2014). However, lexicon-based methods depend on human labeling or manual intervention, which can be labor intensive and time-consuming (H. Zhang et al., 2014; Singhal et al., 2018). The disadvantage of relying on a lexicon is that words can have multiple meanings depending on the context and the meaning and sentiment of a word in one domain may not transfer over to another domain. Furthermore, words that are generally considered 'objective' or lacking sentiment can carry sentiment in certain contexts. This is a significant challenge when using pre-made dictionaries that are not specifically related to the characteristics of the intended product.

The second approach, is machine learning (ML) based SA which is defined by Chollet & Allaire (2018) "as an algorithm trained to search for useful representations of the input data, within a predefined space of possibilities using a guidance feedback signal in order to make predictions based on previously established representations to find other known models." ML is a subdiscipline of AI which has the main aim of automating intellectual tasks usually reserved for humans (Shieber, 2004). Additionally, some of the earlier work on modern machine learning was by Turing (1950) in his remarkable paper the "Computing Machinery and Intelligence." Turing proposed the question "can machines think?" and other key concepts that dramatically shifted the landscape of AI and ML. The question "can machines think?" raises many others such as: "can machines learn? can a computer extrapolate beyond what it is initially programed to perform?" These questions lead to the development of what is now understood as ML.

In traditional programming the algorithm is given input data and a set of rules (or a program) which are instructions on how to process and interpret the data. Then the algorithm produces answers based on the rules it was fed. In ML the algorithm is given input data and the expected answers from the data. From this information the algorithm produces the rules, and these can be used on novel sets of data (Chollet & Allaire, 2018). In order to derive the rules from the data the ML program must learn useful representations of the input data to better predict the expected outcome. A representation is simply a translation or different way to look at the data. For example, many digital streaming platforms use ML algorithms to recommend shows the user is likely to watch or purchase. The task of the ML algorithm is to predict what the user will want to watch based on their previous viewing history. What the user chooses to watch, how they rate what they watch, what they avoid watching and the library of content available are all a part of the input data. One way to represent the data may be to organize the shows based on their titles. Content with similar titles could be clustered together. For example, two shows titled "American horror story" and "American dream" would share a cluster. However, the title of a show doesn't necessary directly relate to the topic or give a lot of information about the show. A more useful way to represent this data may be to categorize the users previously watched shows and content library based on genres (comedy, horror, anime etc.). The "learning" in ML is the automatic process of finding

the most useful representation of input data. If a user watches several shows tagged with "horror" that data will be clustered with other similar shows also tagged horror. The degree of similarity increases if the user positively rates "horror" shows. Based on this representation of the user's data the ML algorithm predicts the user will watch another show tagged "horror" over the other content available. If the user follows the recommendation provided by ML algorithm that serves as a positive feedback signal and the algorithm will make more predictions using data with the "horror" tag. ML algorithms remain a useful tool to curate content to the users' preferences and to understand users opinions and feelings about specific content.

The growing availability of "big data" related to the consumer preference and acceptance of food products has great potential for ML sentiment analysis which usually requires a large set of input data. As previously outlined above, it's reasonable to suggest there is a lot of interesting and useful information on consumer acceptance that is housed in "big data" online reviews or microblogs. For instance, online reviews of food have the potential to predict which foods the consumer like and dislike. A good example of this is the e-commerce platform Amazon.com. Amazon.com allows users to rate products using a 5-point scale and give free comments on the products acceptability (*Amazon Fine Food Reviews*, n.d). If a user searches "organic peanut butter" on Amazon.com when they click on an option, there will be information about the product and a "you might also like" section which recommends similar products to a user. These recommendations may be made by grouping together products with similar attributes (nut products) and branding. In this case it would be other peanut butters or peanut based products with "green" marketing. Amazon.com also has a "frequently bought together" section which shows the user other people who bought the product they are interested in also bought these other related products. To make these suggestions Amazon.com may use a ML algorithm that records users who bought the organic peanut butter, then within a certain time frame the other products these users purchased. A useful representation of this data may be a similarity cluster analysis between the organic peanut butter and the products purchased later. Products that are highly clustered together could be good suggestions for the "frequently bought together" section. However, the user purchasing these products does not give us any indication of their satisfaction of these products. A better prediction would only include highly clustered products with positive reviews. The short text format of online food review data has yielded successful prediction models for food products and bettered our understanding of consumer acceptance.

Machine learning based sentiment analysis requires a training dataset or labeled data. Labeled data and classes are pieces of data that have been tagged with one or more labels identifying certain properties or characteristics about that set of data. For example, a labeled training dataset for SA may consists of words that are labeled positive, negative or neutral and have a score based on their label. Essentially, the training data is a set of examples of unstructured text (the input) and the coordinating sentiment scores (the desired outputs). Additionally, by analyzing the training data the algorithm can find the best representation of the data in order to infer a set of rules which can be applied to new data. Lastly, the predicted sentiment scores (output) are compared against the observed sentiment scores (input). The difference between the predicted and observed sentiment

scores is the test error metric. Depending on how large the test error, the researcher may choose to refine the model by looking for other data representations and repeat the process until the researcher is satisfied with the accuracy. (Agarwal & Mittal, 2016; Ahmad et al., 2017; Samal et al., 2017).

2.6. Deep learning method

Deep learning (DL) is within the umbrella of AI, but it is more closely related to ML. Similar to ML, in DL the algorithm learns new data representations. However, there is a focus on learning "successive layers of increasing meaningful representations" (Chollet & Allaire 2018). The numbers of data representation layers are referred to as depth. A typical deep learning model can have hundreds of different layers that are automatically learned from the training data. These layered data representations are usually learned via neural networks. Neural networks are machine learning models loosely based on biological neural networks however, there is no evidence to suggest that biological neurons function like neural networks or that neural networks act as a model of the brain. A better interpretation is that neural networks are "structured in literal data representation layers that estimate functions and depend on a large quantity of inputs/features (Chollet & Allaire 2018; Minaee et al., 2021). Another way to visualize neural networks is to think of them as a water purifier. The original data set or murky water goes through a series of filters or transformations to abstract useful information from the data. Each filter or layer is different and gives more information as the layers progress (Chollet & Allaire 2018). In sentiment analysis the DL approach utilizes multiple neural network models to map useful data representations or layers of text into a vector space. Then, the algorithm makes predictions and learns the best data representation using a feedback signal. In summation, DL can receive input from unlabeled data, learn the features (hidden layers) from the input with minimal intervention from the programmer (Chollet & Allaire 2018; Minaee et al., 2021).

Recently, deep learning methods have played an increasing role in addressing NLP tasks without some of the disadvantages of machine learning. For example, ML relies on the researcher manually engineering layers or useful data representations. ML algorithms will not work directly on unstructured text that has not been labeled. Manually, extracting text features and creating a training document can be very time-consuming. Also, a strong dependence on domain knowledge for designing features makes this method difficult to generalize to new data. In contrast, a DL algorithm automates this step and can learn the features or layers from the text without the labor and time considerations required of ML (Liang et al., 2017; Minaee et al., 2021).

2.7. Survey of SA in sensory evaluation

There is growing interest in using SA in sensory science. "Big data" like social media and food-related reviews have been a primary target of researchers. International Flavors and Fragrances - Sensory & Consumer Insights (SCI) research team was interested in whether, data extracted from social media platforms such as Twitter would give a return on investment. As well as weather the insights derived are reliable when compared to

traditional sensory science techniques. In 2007 they conducted focus groups in the U.S. which consisted of 3 different consumer groups about coffee products. They compared the comments from the focus groups with popular social media searches on coffee and found that almost all of the attributes identified from the focus groups were also relevant in social media search results. The authors highlighted the validity of social media output and recommended more sensory product research professionals to include social media data in their future consumer research frameworks (Carr et al., 2015).

In a similar study, SA was used to analyze short text Twitter comments to reveal insights on food-related consumer behavior (Vidal et al., 2015). Researchers explored the question "what do people say when they tweet about different eating situations?" They searched for tweets with the words: *breakfast*, *lunch*, *dinner* and *snack* and a total of 69,961 tweets were retrieved, of which 48,746 corresponded to original tweets and were subject to an automated word analysis and manual content analysis. The tweets contained valuable information which foods was consumed, when, where, with whom, and why. These results confirmed to the researchers the potential of Twitter data as a source of information about food choice decisions. Overall, the authors found that SA and food related "big data" have promising potential as a tool in sensory science, but not without limitations such as high labor costs and lack of emotional expressions in the text.

To overcome the limitations of processing social media data, focus shifted to analyzing free comment data (Luc et al., 2020). A lexicon-based SA approach was used to analyze and check the consistency of data from a free comment and JAR task on seven samples of perfume. The researchers manually created a sentiment dictionary tailored to perfumes and the other products in their study. The authors concluded that SA was an effective way to capture the emotional valence of the consumer from free comment and JAR data. However, they disliked the results from the lexicon-based approach because it relies on sentiment dictionaries, which are time-consuming to create, and specific for each product space. Recently, Visalli et al. (2020) used the Microsoft Text Analytics API a ML sentiment analysis model on free comment data collected from home use studies. The researchers found that the sentiment scores were significantly correlated with liking. Overall, the ML based approach was able to overcome many of the shortcomings of the lexicon-based approach in previous similar studies. The researchers recommended that if a sentiment score is to be considered an indirect measurement of liking further exploration is needed, especially in cases where the sensory difference between products is small.

References

Agarwal, B., & Mittal, N. (2016). *Machine learning approach for sentiment analysis* (pp. 21–45). Springer.

Ahmad, M., Aftab, S., Muhammad, S. S., & Ahmad, S. (2017). Machine learning techniques for sentiment analysis: A review. *International journal of multidisciplinary sciences and engineering*, *8*(3), 27.

Allaire, J. J. (2018). *Deep Learning with R*. Simon and Schuster.

Al-Shabi, M. A. (2020). *Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining*. 7.

Amazon Fine Food Reviews. (n.d.). Retrieved September 15, 2021, from https://kaggle.com/snap/amazon-fine-food-reviews

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, *77*, 236–246. https://doi.org/10.1016/j.eswa.2017.02.002

Asghar, N. (2016). Yelp Dataset Challenge: Review Rating Prediction. *ArXiv:1605.05362 [Cs]*. http://arxiv.org/abs/1605.05362

Boutrolle, I., Delarue, J., Arranz, D., Rogeaux, M., & Köster, E. P. (2007). Central location test vs. home use test: Contrasting results depending on product type. *Food Quality and Preference*, *18*(3), 490–499. https://doi.org/10.1016/j.foodqual.2006.06.003

Bunn, A., & Korpela, M. (n.d.). *An Introduction to dplR*. 16.

Carr, J., Decreton, L., Qin, W., Rojas, B., Rossochacki, T., & Yang, Y. wen. (2015). Social media in product development. *Food Quality and Preference*, *40*, 354–364. https://doi.org/10.1016/j.foodqual.2014.04.001

Chapter 00065—Sensory Science | Elsevier Enhanced Reader. (n.d.). https://doi.org/10.1016/B978-0-444-52512-3.00065-6

Chen, H., & Zimbra, D. (2010). AI and opinion mining. *IEEE Intelligent Systems*, *25*(3), 74–80.

Chollet, F., & Allaire, J. J. (2018). *Deep Learning mit R und Keras: Das Praxis-Handbuch von den Entwicklern von Keras und RStudio*. MITP-Verlags GmbH & Co. KG.

Collobert, R. (2011). Deep learning for efficient discriminative parsing. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 224–232.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

de Graaf, C., Cardello, A. V., Matthew Kramer, F., Lesher, L. L., Meiselman, H. L., & Schutz, H. G. (2005). A comparison between liking ratings obtained under laboratory and field conditions: The role of choice. *Appetite*, *44*(1), 15–22. https://doi.org/10.1016/j.appet.2003.06.002

De Wijk, R. A., Kaneko, D., Dijksterhuis, G. B., van Zoggel, M., Schiona, I., Visalli, M., & Zandstra, E. H. (2019). Food perception and emotion measured over time in-lab and

in-home. *Food Quality and Preference*, *75*, 170–178. https://doi.org/10.1016/j.foodqual.2019.02.019
DoorDash Revenue and Usage Statistics (2021). (2020, September 14). Business of Apps. https://www.businessofapps.com/data/doordash-statistics/
Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38*(1), 188–230. https://doi.org/10.1002/aris.1440380105
Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 281–285.
Esuli, A. (2021). *SentiWordNet*. https://github.com/aesuli/SentiWordNet/blob/f2c813a62ecb933aba0634932d1d838b408d1ef3/papers/LREC06.pdf (Original work published 2019)
Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, *2*(1), 1–14.
Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & González-Castaño, F. J. (2016). Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, *58*, 57–75.
Fonseca, F. G., Esmerino, E. A., Tavares Filho, E. R., Ferraz, J. P., da Cruz, A. G., & Bolini, H. M. (2016). Novel and successful free comments method for sensory characterization of chocolate ice cream: A comparative study between pivot profile and comment analysis. *Journal of Dairy Science*, *99*(5), 3408–3420.
Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. *2014 Seventh International Conference on Contemporary Computing (IC3)*, 437–442.
Grubhub Revenue and Usage Statistics (2021). (2020, September 11). Business of Apps. https://www.businessofapps.com/data/grubhub-statistics/
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.
Heng, Y., Gao, Z., Jiang, Y., & Chen, X. (2018). Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach. *Journal of Retailing and Consumer Services*, *42*, 161–168. https://doi.org/10.1016/j.jretconser.2018.02.006
Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. 168–177.
Hutto, C. J. (2021). *Cjhutto/vaderSentiment* [Python]. https://github.com/cjhutto/vaderSentiment (Original work published 2014)
Jaeger, S. R., Vidal, L., & Ares, G. (2021). Should emoji replace emotion words in questionnaire-based food-related consumer research? *Food Quality and Preference*, *92*, 104121. https://doi.org/10.1016/j.foodqual.2020.104121
Khoo, C. S. G., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, *44*(4), 491–511. https://doi.org/10.1177/0165551517703514
Kozlowska, K., Jeruszka, M., Matuszewska, I., Roszkowski, W., Barylko-Pikielna, N., & Brzozowska, A. (2003). Hedonic tests in different locations as predictors of apple juice consumption at home in elderly and young subjects. *Food Quality and Preference*, *14*(8), 653–661. https://doi.org/10.1016/S0950-3293(02)00207-0

Lahne, J., Trubek, A. B., & Pelchat, M. L. (2014). Consumer sensory perception of cheese depends on context: A study using comment analysis and linear mixed models. *Food Quality and Preference*, *32*, 184–197.
Lawless, H. T., & Heymann, H. (2010a). Acceptance Testing. In H. T. Lawless & H. Heymann (Eds.), *Sensory Evaluation of Food: Principles and Practices* (pp. 325–347). Springer. https://doi.org/10.1007/978-1-4419-6488-5_14
Lawless, H. T., & Heymann, H. (2010b). *Sensory evaluation of food: Principles and practices* (Vol. 2). Springer.
Liang, H., Sun, X., Sun, Y., & Gao, Y. (2017). Text feature extraction based on deep learning: A review. *EURASIP Journal on Wireless Communications and Networking*, *2017*(1), 1–12.
Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415–463). Springer.
Luc, A., Lê, S., & Philippe, M. (2020). Nudging consumers for relevant data using Free JAR profiling: An application to product development. *Food Quality and Preference*, *79*, 103751. https://doi.org/10.1016/j.foodqual.2019.103751
MacFie, H. J., Bratchell, N., GREENHOFF, K., & Vallis, L. V. (1989). Designs to balance the effect of order of presentation and first-order carry-over effects in hall tests. *Journal of Sensory Studies*, *4*(2), 129–148.
Mahieu, B., Visalli, M., Thomas, A., & Schlich, P. (2020). Free-comment outperformed check-all-that-apply in the sensory characterisation of wines with consumers at home. *Food Quality and Preference*, *84*, 103937.
McAuley, J. J., & Leskovec, J. (2013). From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13*, 897–908. https://doi.org/10.1145/2488388.2488466
McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. *Proceedings of the 7th ACM Conference on Recommender Systems*, 165–172. https://doi.org/10.1145/2507157.2507163
McAuley, J., Leskovec, J., & Jurafsky, D. (2012). *Learning attitudes and attributes from multi-aspect reviews*. 1020–1025.
Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011
Meiselman, H. L. (1992). Methodology and theory in human eating research. *Appetite*, *19*(1), 49–55. https://doi.org/10.1016/0195-6663(92)90235-X
Meiselman, H. L., & Schutz, H. G. (2003). History of food acceptance research in the US Army. *Appetite*, *40*(3), 199–216. https://doi.org/10.1016/S0195-6663(03)00007-2
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.
Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.
Miller, C., Hamilton, L., & Lahne, J. (2021). Sensory Descriptor Analysis of Whisky Lexicons through the Use of Deep Learning. *Foods*, *10*(7), 1633.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning–based Text Classification: A Comprehensive Review. *ACM Computing Surveys (CSUR)*, *54*(3), 1–40.
Moskowitz, H. R., Muñoz, A. M., & Jr, M. C. G. (2008). *Viewpoints and Controversies in Sensory Science and Consumer Product Testing*. John Wiley & Sons.
Muir, D. D., & Hunter, E. A. (1991). Sensory evaluation of Cheddar cheese: Order of tasting and carryover effects. *Food Quality and Preference*, *3*(3), 141–145.
Niimi, J., Collier, E. S., Oberrauter, L.-M., Sörensen, V., Norman, C., Normann, A., Bendtsen, M., & Bergman, P. (2022). Sample discrimination through profiling with rate all that apply (RATA) using consumers is similar between home use test (HUT) and central location test (CLT). *Food Quality and Preference*, *95*, 104377. https://doi.org/10.1016/j.foodqual.2021.104377
Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The Measurement of Meaning*. University of Illinois Press.
Peryam, D., & Girardot, N. (n.d.). I952. Advanced taste-test method. *Food Eng*, *24*, 58.
Prescott, J., Hayes, J. E., & Byrnes, N. K. (2014). Sensory Science. In *Encyclopedia of Agriculture and Food Systems* (pp. 80–101). Elsevier. https://doi.org/10.1016/B978-0-444-52512-3.00065-6
Qiu, J., Liu, C., Li, Y., & Lin, Z. (2018). Leveraging sentiment analysis at the aspects level to predict ratings of reviews. *Information Sciences*, *451–452*, 295–309. https://doi.org/10.1016/j.ins.2018.04.009
RateBeer.com—Statistics. (n.d.). Retrieved September 15, 2021, from https://www.ratebeer.com/Stats.asp
Rinker, T. (2021). *Sentimentr* [R]. https://github.com/trinker/sentimentr (Original work published 2015)
Rojas-Barahona, L. M. (2016). Deep learning for sentiment analysis. *Language and Linguistics Compass*, *10*(12), 701–719. https://doi.org/10.1111/lnc3.12228
Rothman, L., & Parker, M. (2009). Just-about-right (JAR) scales. *West Conshohocken, PA: ASTM International*.
Samal, B., Behera, A. K., & Panda, M. (2017). Performance analysis of supervised machine learning techniques for sentiment analysis. *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)*, 128–133.
Schouteten, J. J., Verwaeren, J., Gellynck, X., & Almli, V. L. (2019). Comparing a standardized to a product-specific emoji list for evaluating food products by children. *Food Quality and Preference*, *72*, 86–97. https://doi.org/10.1016/j.foodqual.2018.09.007
Schouteten, J. J., Verwaeren, J., Lagast, S., Gellynck, X., & De Steur, H. (2018). Emoji as a tool for measuring children's emotions when tasting food. *Food Quality and Preference*, *68*, 322–331. https://doi.org/10.1016/j.foodqual.2018.03.005
Sharif, M. K., Butt, M. S., Sharif, H. R., & Nasir, M. (2017). Sensory evaluation and consumer acceptability. *Handbook of Food Science and Technology*, 361–386.
Shieber, S. M. (2004). *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. MIT Press.
Singhal, S., Maheshwari, S., & Meena, M. (2018). Survey of challenges in sentiment analysis. In *Recent Findings in Intelligent Computing Techniques* (pp. 229–238). Springer.

Snuggs, S., & McGregor, S. (2021). Food & meal decision making in lockdown: How and who has Covid-19 affected? *Food Quality and Preference*, *89*, 104145.
Swaney-Stueve, M., Jepsen, T., & Deubler, G. (2018). The emoji scale: A facial scale for the 21st century. *Food Quality and Preference*, *68*, 183–190. https://doi.org/10.1016/j.foodqual.2018.03.002
Symoneaux, R., Galmarini, M., & Mehinagic, E. (2012). Comment analysis of consumer's likes and dislikes as an alternative tool to preference mapping. A case study on apples. *Food Quality and Preference*, *24*(1), 59–66.
Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.
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.
Team, R. C. (2013). *R: A language and environment for statistical computing*.
ten Kleij, F., & Musters, P. A. (2003). Text analysis of open-ended survey responses: A complementary method to preference mapping. *Food Quality and Preference*, *14*(1), 43–52.
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
Turing, A. M., & Haugeland, J. (1950). *Computing machinery and intelligence*. MIT Press Cambridge, MA.
Uber Eats Revenue and Usage Statistics (2021). (2020, August 25). Business of Apps. https://www.businessofapps.com/data/uber-eats-statistics/
Vidal, L., Ares, G., Blond, M. L., Jin, D., & Jaeger, S. R. (2020). Emoji in open-ended questions: A novel use in product research with consumers. *Journal of Sensory Studies*, *35*(6), e12610. https://doi.org/10.1111/joss.12610
Vidal, L., Ares, G., & Jaeger, S. R. (2016). Use of emoticon and emoji in tweets for food-related emotional expression. *Food Quality and Preference*, *49*, 119–128. https://doi.org/10.1016/j.foodqual.2015.12.002
Vidal, L., Ares, G., Machín, L., & Jaeger, S. R. (2015). Using Twitter data for food-related consumer research: A case study on "what people say when tweeting about different eating situations." *Food Quality and Preference*, *45*, 58–69.
Visalli, M., Mahieu, B., Thomas, A., & Schlich, P. (2020). Automated sentiment analysis of Free-Comment: An indirect liking measurement? *Food Quality and Preference*, *82*, 103888. https://doi.org/10.1016/j.foodqual.2020.103888
Yelp—Company—Fast Facts. (n.d.). Retrieved September 15, 2021, from https://www.yelp-press.com/company/fast-facts/default.aspx
Zhang, H., Gan, W., & Jiang, B. (2014). Machine Learning and Lexicon Based Methods for Sentiment Classification: A Survey. - *2014 11th Web Information System and Application Conference*, 262–265. https://doi.org/10.1109/WISA.2014.55
Zhang, L., & Liu, B. (2014). Aspect and Entity Extraction for Opinion Mining. In W. W. Chu (Ed.), *Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities* (pp. 1–40). Springer. https://doi.org/10.1007/978-3-642-40837-3_1

Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management*, *29*(4), 694–700. https://doi.org/10.1016/j.ijhm.2010.02.002
(N.d.).

Chapter 3: Experimental

3.1 Introduction

There are many factors to consider when developing a food product such as nutritional benefits, high quality packaging and rigorous food safety testing. However, if the sensory quality of a food or beverage is not high, the probability that it will perform well with consumers is low (Lawless & Heymann, 2010). A free comment task, which is a sensory evaluation test used to generate intuitive comments from untrained consumers, may help identify and describe sensory attributes of products to help predict sensory quality (Symoneaux et al., 2012). However, traditional sensory evaluation methods to analyze free comment data can be time consuming and cumbersome (Fonseca et al., 2016a; Mahieu et al., 2020). The aim of the current study is to examine the potential utility of sentiment analysis (SA) as a new and underutilized data analysis method that can be added to the toolkit of a sensory scientist.

SA is computer science tool that is used to identify an authors' opinions or tone about specific topics within an unstructured text (L. Zhang & Liu, 2014; Medhat et al., 2014). The sentiment is commonly measured by sentiment score. Every model has different parameters, but generally a higher sentiment score indicates a positive tone and a negative sentiment score indicates a negative tone (L. Zhang et al., 2014). There are two types of sentiment analysis: 1) lexicon-based methods, which use sentiment dictionaries containing a pre-defined list of words with sentiment scores attached to determine the sentiment score of a text, and 2) machine-learning (ML) methods, which use a variety of optimization models that find and recognize useful patterns to derive sentiment from a text (Chollet & Allaire 2018; Minaee et al., 2021). Deep learning (DL) is a subset of ML but there is a stronger focus on learning increasingly complex layers of useful patterns or representations in the data using neural networks (Chollet & Allaire 2018)

The benefit of using a DL sentiment analysis method is that it has the potential to quickly recognize key features and semantic relationships in the text that relate to positive and negative sentiment with minimal human intervention (Liang et al., 2017; Minaee et al., 2021). A drawback of this method is that it requires a large amount of input data to train and fine tune the model (Shieber, 2004). Gathering a large volume of high-quality sensory data can be very hard to acquire and expensive to produce. One alternative is online food reviews. Online food reviews can be defined as a person or customer's thoughts/opinions and experiences in an online review or comment section (Snuggs & McGregor, 2021). An online review is structured similarly to free comment data: there is an implied open-ended question where consumers give their thoughts and feelings about a why they dislike or like a product. Currently, there is a growing repository of online food reviews available. For example, Yelp.com allows users to participate in discussions based on local restaurants, food, and service. As of December 2020, Yelp has over 224 million reviews and 31 million monthly desktop users (*Yelp - Company - Fast Facts*, n.d.). Recently, researchers have started to explore the viability of new data sources such as online food reviews as a research tool in sensory and consumer science (Carr et al., 2015; Hamilton & Lahne, 2020; Miller et al., 2021; Tao et al., 2020; Tian et al., 2021;

Vidal et al., 2015). Ashgar (2016) retrieved a subset of 1,125,458 restaurant reviews, extracted from the Yelp.com open dataset, and created sixteen prediction models to determine the best model for predicting ratings from reviews.

Typically, when performing ML sentiment analysis, the researcher will gather a large data set that they are interested in classifying. A small subset of this data will be extracted and this will become the training dataset (Agarwal & Mittal, 2016; Ahmad et al., 2017; Samal et al., 2017). This process is important because it would be very difficult to manually collect millions of free comment data from sensory studies. The large data requirement presents a significant obstacle to sensory scientists who are interested in using or creating a ML/DL sentiment analysis method. To circumvent this obstacle, in the current study we used a large dataset of online food reviews to train the DL sentiment analysis model we created, and we then tested our model on a smaller scale using free comment data we generated from a consumer acceptance study.

The purpose of the current project was to evaluate how sentiment analysis performs as a data analysis tool in a sensory food science study. There were two phases to our research design. In phase 1, we developed a sentiment analysis model using a deep learning convolution neural network (CNN) method. To train and fine tune our model, we used online beer reviews from Ratebeer.com. In phase two, we recruited a consumer acceptance panel to generate free comment data that we used to validate and test our sentiment analysis model and three other popular models. To generate FC data, we conducted a consumer acceptance study. Our samples were 6 beers: three from our training data set and three locally recommended beers. Panelists were presented with one sample at a time and were instructed to smell and taste the sample before answering the questions, providing their comments, and rating the samples using the traditional 9-pt scale. The specific research objectives of this study were: 1) to explore quicker and automated methods of sentiment analysis to better understand and predict consumer acceptance, and 2) to examine the advantages and disadvantages of sentiment analysis as a data analysis tool in sensory evaluation.

3.2 Materials and Methods

3.2.1. Sentiment Analysis

All data analysis was carried out using SPSS statistics software, Microsoft Excel and R programing software. We created a deep learning sentiment analysis model using the methodology outlined in 'Deep learning for R' guidebook (Chollet & Allaire, 2018). In deep learning sentiment analysis, the algorithm learns to recognize patterns in words, sentences and paragraphs.

First, we preprocessed the online beer review data (McAuley et al., 2012) from our test dataset into useful representations by word embedding or vectorizing the text into numeric tensors. Word vector representations or word embeddings convert text into numbers while maintaining important sematic representations. Word vectors use a ML model to map the text into a low-dimensional vectors. Vectors are usually an array of

numbers used to store data in an organized fashion (Minaee et al., 2021; Rojas-Barahona, 2016). For example, a vector could be v=(0,1,2,3). This would be a 1x4-demensional vector because there are four elements in the array and each of these elements could refer to a layer.

Second, we used a subset of online beer reviews collected by McAuley (2012) from one of the largest beer-rating websites - RateBeer.com (Table 1). The dataset we used was collected from Apr 2000 - Nov 2011 and consists of 2,924,127 reviews, 40,213 users of which 4,798 had more than 50 reviews. We obtained the data (with permission) from Stanford University SNAP library (accessed here https://snap.stanford.edu/about.html).

Third, we created a multilayer neural network using one-dimensional convolutions for sequence processing (Chollet & Allaire, 2018). Convolutional neural networks (CNN) also known as convents are a type of deep-learning model (See figure 1). A CNN has hidden layers called convolutional layers which detect patterns in the text and find useful representations of the data. The CNN will extract short segments of text or a patch, in our case 3-words or less from the word vectors. Each convolutional layer has a filter, and these filters help the network to learn local patterns in a sequence. When the CNN receives an input, the filter will slide over each patch or 3-word segment and apply an identical transformation to the input or entire sentence. This sliding is also referred to as convolving. The output of this process creates a matrix which is called a 3D tensor or input feature map. The output feature map is composed of filters – filters are responsible for encoding specific aspects of the input data such as: semantic signals or emotions (e.g., happiness, disgust, sadness etc.) (Chollet & Allaire, 2018).

In a CNN, a pattern learned from one 3-word segment of a sentence can be learned in a different position in the same sentence or a different part of the text in any context. In contrast, a neural network would need to learn the same pattern in a 3-word segment again if it was in a different position or context in the sentence. CNNs can be very advantageous for those working with smaller datasets because they require less training data to learn useful representations of data because it has the ability to generalize what it has learned (Chollet & Allaire, 2018).

3.2.2. SentimentR, VADER, and Liu Hu Sentiment Analysis Models

Additionally, we compared our DL model to three other pre-existing sentiment analysis models: SentimentR, VADER, and Liu and Hu opinion lexicon. We selected these three models to use a benchmark comparison to our model. SentimentR, VADER and Liu Hu are three lexicon-based models that are available for free for research or scientific purposes. We are interested to see if there are advantages to creating a sentiment analysis model with a template versus using a model that is popular and widely available.

SentimentR is a package for R software, and it uses a dictionary-based approach that considers valence shifters. Valence shifters are words that change the meaning of polarized words. For example, “I do like it” has an overall positive polarity, but if a negative valence shifter is added – “I do *not* like it” the overall polarity of the sentence

becomes negative (Rinker, 2015/2021). VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis package is a lexicon and rule-based sentiment analysis that is specifically attuned to sentiments expressed in social media. The sentiment scores are scaled from -1 (negative feeling) to 1 (positive feeling), with 0 being neutral or indeterminate. The scores are calculated by summing up the intensity for each world in the text, negative, neutral and positive. The compound score is then computed by normalizing the 3 scores (Hutto & Gilbert, 2014). The Liu and Hu opinion lexicon module is lexicon-based sentiment analysis. There is no specific range however, the sentiment score is the difference between the sum of positive and sum of negative words, normalized by the length of the document and multiplied by a 100. The Liu and Hu opinion lexicon module counts the number of positive, negative and neutral words in the text and organizes them based on which polarity is most frequently represented. The lexicons are a list of positive and negative words. It considers any word that does not appear in the lexicon neutral (Hu & Liu, 2004).

Our sentiment analysis model, SentimentR, VADER, and Liu and Hu models were applied to the free comments from this study, and we collected the sentiment scores. We used the free comment data to help test and validate our sentiment analysis models. The means and standard deviations of the ratings and sentiment scores were computed. To better facilitate comparisons for the ratings (1-9) and the sentiment analysis models which all had different senstiment score ranges all the data was normalized. The ratings ranged from 1-9: where 1 represents “dislike extremely” and 9 represents “like extremely,” and the sentiment scores across the four different models were normalized to a range of 0.00 – 1.00. This range was divided into thirds where: 0.00-0.33 represented negative score range, 0.33-0.66 represented the neutral score range and 0.66-1.00 represented the positive score range. We also removed any outliers from the sentiment scores for all four models.

A linear regression was used to analyze how well the sentiment scores generated from our model predict the ratings from the traditional nine-point scale. A linear regression can be used to fit a predictive model to the observed data from the free comment task to the values of explanatory variables or the sentiment scores from our model. Accuracy, precision and recall tests for each sentiment analysis model. These tests reveal how well our model preformed correctly predicting positive and negative sentiment from the free comment data. Lastly, between subjects ANOVAs to compare preference differences between the six beer samples. These tests were performed to investigate if the ratings and/or the sentiment analysis models could detect significant differences between the sentiment related to the different beer samples from the free comment study.

3.2.3. LIME Model

LIME was presented in 2016 and it was created to explain any black box model by creating a local approximation using the inputs and outputs of the predictive model (Ribeiro, Singh, Guestrin et al., 2016). LIME is an acronym which represents “local interpretable model agnostic explanations.” The CNN model used in this study does not have a linear relationship or a simple explanation for how the model makes a prediction

or gives a sentiment score based on the input. Instead, the model learns complex patterns from the input data or free comments. If we were interested in learning why the sentiment analysis model gave a positive or negative sentiment score - it would be very difficult to summarize the sentiment analysis model's decision boundaries in one explanation. LIME focuses on the local area of the individual prediction which allows us to generate an easy explanation in that local region without having to take the entire model into account. It works on many different data types such as: text, tabular, image, graph data. LIME provides explanations which help to improve the acceptance of a predictive algorithm. The explanations are locally faithful, meaning the explanation given will be the most relevant to a specific input and the area around it. LIME works by creating a local approximation of a complex model for a specific input. In this thesis, we used the LIME to help explain which words or phrases in the free comment responses weigh heavily in the CNN model's decision to predict the sentiment of the text. By examining these key words and phrases we can better understand the patterns the sentiment analysis model is learning about the free comment responses.

3.2.4. Free Comment Task

3.2.4.1. Samples

Six beers were selected for this study. Previous studies that also carried out free comment tasks had a similar number of samples (Luc et al., 2020; Visalli et al., 2020). The beers selected were representative of our test dataset or were popular beers, readily available in the area (Blacksburg, VA, USA). Three of the beers selected represented the extremes of our test dataset: beers that were rated the highest and lowest. The beers that received high reviews from our test dataset were largely stouts, porters and sour beers and the beers that were generally negatively reviewed were American light lager beers. In consideration of our test demographic who were in their 20s and/or were college students, we also selected three locally available beers. Just Pressed, All Day IPA and Budweiser were recommended by local liquor store staff as popular beers that were highly requested and well-liked by customers (see Table 2).

All the samples were presented in 20 mL opaque, black, stemmed glasses labeled with randomized 3-digit codes in one session. The samples were served, one by one and presented in a Williams Latin square design to avoid carryover effects (MacFie et al., 1989; Muir & Hunter, 1991). Samples were kept in the refrigerator and poured 15 minutes prior to panelists arrival. Beer samples were decanted into sealed Nalgene bottles to preserve carbonation after opening. All beer samples, open cans of beer, and the contents in the Nalgene bottles were discarded after three hours and a new can was opened to reduce carbonation loss and sensory characteristics of the beer.

3.2.4.2. Panelists

Panelists (N = 68) were recruited from Virginia Tech and the surrounding Blacksburg, Virginia community. Panelists were recruited through the university e-mail listservs and word of mouth. All recruitment, scheduling and data collection was managed by

Compusense Cloud sensory management software (Guelph, ON, Canada). Prior to beginning the study, all potential panelists completed a brief screening survey to determine their age, any allergies to beer and its ingredients, weekly consumption of beer, overall opinion about beer and eligibility. All panelists were untrained, required to be 21 years or older, have previously consumed beer, and report no allergies to alcohol or fermented beverages. Panelists who completed and qualified the screening survey (See Appendix) were directed to schedule a 30-minute session to participate in the study. This study was approved by the Virginia Tech Institutional Review Board, protocol 21-890.

3.2.4.3. Free Comment Task and Questionnaire Procedure

First, panelists (N=68) were presented with one sample at a time and were instructed to smell and taste the sample before answering the questions and leaving comments. They repeated this procedure for all six samples. After tasting a sample the panelist had to complete three questions before they received the next sample. Panelists were asked to "Describe the overall sensory characteristics and your personal experience of the beer," to collect free comment data. Next, panelists were asked to assess the overall acceptance of the samples. The acceptance test used two scales. The panelists were asked "Overall, did you like this beer?" and were given the option to check "Yes" or "No". Lasty, the panelists were given the traditional nine-point scale (Peryam & Girardot, 1952). Responses were collected through the sensory software Compusense Cloud sensory management software (Guelph, ON, Canada) then analyzed.

Panelists were instructed to expectorate after each tasting and were provided water and saltine crackers *ad libitum* to cleanse their palate after each tasting (Sharif et al., 2017). Data collection was carried out in individual sensory booths with standard conditions such as sample coding, random order of serving samples, consistent sample preparation and serving, light conditions and room temperature further outlined in Lawless and Heymann (2010).

3.3. Statistical Analysis

We used the accuracy formula used by Al-Shabi (2020) in a similar study to compare our sentiment model with the other three models.

$$Accuracy = \frac{Tp \ + \ Tn}{P \ + \ N}$$

Where:
Tp = the number of positive sentiments detected correctly
Tn = the number of negative sentiments detected correctly
P = the total number of positively rated reviews
N = the total number of negatively rated reviews

We also calculated precision and recall (see Table 3.). Precision is the ratio of correctly predicted positive reviews to the total number of predicted positive reviews. Precision

answers the question: "out of all of the reviews the model labeled as positive, how many of those actually received a positive rating?" A very high precision percentage indicates a low false positive rate. The formula for precision is as follows:

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

Recall is the ratio of correctly predicted positive reviews to the number of all the reviews. Recall answers the question: "out of all of the reviews that received a high rating, how many did the model label as positive?" The formula for recall is as follows:

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

The results section will cover statistical analysis performed on the FC data gathered from our consumer acceptance study and the sentiment scores calculated our model, VADER, Liu HU and SentimentR. The means and standard deviations for the ratings and sentiment scores were calculated (Table 4). The tests we performed include:

(1) simple linear regressions were performed to see if there was a relationship between the sentiment scores and ratings (1-9) collected using the traditional hedonic scale. We were interested in how well the sentiment scores predicted the observed ratings.

(2) Accuracy, precision and recall tests for each sentiment analysis model. These tests reveal how well our model preformed correctly predicting positive and negative sentiment from the free comment data.

(3) Between subjects ANOVAs to compare preference differences between the six beer samples. These tests were performed to investigate if the ratings and/or the sentiment analysis models could detect significant differences between the sentiment related to the different beer samples from the free comment study.

3.4. Results

3.4.1. Linear

Simple linear regressions were performed to see if there was a relationship between the sentiment scores and ratings (1-9) collected using the traditional hedonic scale. The sentiment scores were set as the independent variable and the ratings were set as the dependent variable (Table 5). We are interested in whether the sentiment scores calculated by four different models: CNN, VADER, Liu Hu, and SentimentR were good predictors of the ratings.

The results of the regression revealed the CNN model was a significant predictor of the ratings ($F(1, 525) = 164.78$ $p<0.001$), with an r^2 of 0.24 (Figure 2). It was found that sentiment scores significantly predicted the ratings ($\beta = 0.34$, $p<.001$). The final predicted

model was: predicted ratings = 0.46 + (0.35 * sentiment score). The sentiment scores calculated by the CNN model accounted for the highest percentage of 24% of the variance in the ratings (1-9).

The results of the regression revealed the SentimentR model was a significant predictor of the ratings ($F(1, 525)= 120.79$, $p<0.001$), with an r^2 of 0.19 (Figure 3). It was found that sentiment scores significantly predicted the ratings ($\sqrt{6} = 0.11$, $p<.001$) The final predicted model was: predicted ratings = 0.70 + (0.11 * sentiment score). The SentimentR model had the second highest coefficient of determination where 19% of the variance in the ratings (1-9) is accounted for in the sentiment scores.

The results of the regression revealed both the VADER model ($F(1, 525)= 75.66$, $p<0.001$) and the Liu Hu model $F(1, 525)= 78.91$, $p<0.001$), significantly predicted the hedonic scale ratings (1-9). Both models had very weak coefficient of determination r^2 of 0.13 and 13% of the variance in the ratings (1-9) is accounted for in the sentiment scores (Figure 4 and 5).

3.4.2. Accuracy, Recall, and Precision Tests

Accuracy tests were performed to determine how accurately our model predicts negative and positive ratings and how the CNN model compares to other popular sentiment analysis models that are readily available. The accuracy test revealed how accurately each module detects polarity using the accuracy formula. The results showed that the best performance in terms of overall accuracy was achieved by our CNN model with an accuracy rate of 69%. The VADER model was the next most accurate with an accuracy rate of 47%. Liu Hu and SentimentR had similar accuracy rates at the lower end of the spectrum 38% and 39% respectively. Overall, all the models were better at detecting highly negative sentiment (1-3 ratings). The CNN model was the best at detecting negative sentiment with 85% negative accuracy, followed by VADER with 54% negative accuracy. The CNN model also had the highest positive accuracy with 50% positive accuracy. Overall, Vader (38%), Liu Hu (34%), and Vader (29%) had a similar weaker performance at detecting positive sentiment.

The precision tests were performed to determine the percentage of false positives made by the models. The results demonstrated that the CNN model had the highest precision at 73%, followed by the VADER model at 40%. The SentimentR and Liu Hu had the lowest precision at 32% and 30% respectively.

Recall tests were performed to determine the percentage of correctly identified positive reviews out of the data labeled positive. Results revealed that the CNN model (50%) had the highest percent recall. Overall, VADER (38%), SentimentR (34%), and Liu Hu (29%) had a low recall with the Liu Hu model having the weakest performance (Table 6).

3.4.3. ANOVAs

In a traditional consumer acceptance study, we would want know if some of the beer samples were significantly rated higher than the other beer samples. In other words: “are there any samples that were significantly liked more and which sensory attributes contributed to that liking?”. To answer this question, one of the statistical tests we would perform is a between-subjects ANOVA using the ratings (1-9) from the 9pt-scale.

A between-subjects ANOVA was performed to determine if the ratings were able to find significant differences in how the samples were liked. The results of the ANOVA revealed that the ratings did not have a significant effect for sample name. The ratings were not able to discern specific differences in how the six beer samples were rated or liked (Table 7).

A between-subjects ANOVAs were performed to compare preference differences between the six beer samples using the sentiment scores calculated by the four models: our deep learning CNN model, SentimentR, VADER, and Liu Hu models. The results of the ANOVA revealed that the CNN model had a significant sample effect [$F(5, 335.12) = 26.98$, $p < .001$] for sample name and [$F(63, 315) = 1.47$, $p = .019$] for panelist name (Table 7). The CNN model was able to find significant differences between the 6 beer samples. Pairwise comparisons indicate that the sample All Day IPA, Just Pressed and Speedway were significantly preferred over Budweiser. Just Pressed and Speedway were significantly preferred over Coors Light and Miller Light. Overall, All Day IPA and Just pressed were the most preferred and Budweiser and Miller Light were the least preferred (Table 8).

The SentimentR model [$F(63,315) = 1.82$, $p = <.001$], the VADER model [$F(63, 315) = 2.13$, $p = <.001$] and the Liu Hu model [$F(63,315) = 1.95$, $p = <.001$] had a significant effect for panelist name. There was no significant effect for sample name and there were no significant pairwise comparisons (Table 7).

3.5. Discussion

3.5.1. Introduction

Our first objective was to explore quicker and automated methods of sentiment analysis to better understand and potentially predict consumer acceptance. Overall, all the simple linear regressions revealed that the sentiment scores calculated by CNN, VADER, Liu Hu and SentimentR models positively correlated with the ratings. However, these were very weak correlations. The strongest correlation accounted for 24% of the variance in the ratings and the weakest correlations accounted for 13% of the variance in the ratings. As the ratings become more positive (7-9) the sentiment analysis models became worse predictors of positive sentiment and the correlation became weaker. Interestingly, none of the models were very good predicting the rating but, our CNN model did outperform the

three industry standard sentiment analysis models. Additionally, the between-subjects ANOVA demonstrated that the CNN model was able to find specific differences in how the beer samples were liked. This could not be accomplished using the ratings from the traditional 9pt-scale or the other three lexicon-based models. The lexicon-based models may have had better performance if they were equipped with tailored sentiment dictionaries. Regardless of the fact we believe these results indicate there is substantial merit in creating a sentiment analysis model in comparison to using a non-customized sentiment analysis model available online.

The accuracy, precision and recall tests reveal that all of the models performed better at identifying the negative ratings. We suggest the coefficient of determinations were weaker because the sentiment analysis models were poor predictors of positive sentiment and stronger predictors of negative sentiment. A closer examination of the raw data demonstrates that there were more negative and neutral ratings than positive ratings. For example, 227 of the ratings were from 1-5 and the remaining 181 - were extremely positive from 6-9. Furthermore, there is a bias in the data towards negatively rated reviews which may have impacted the accuracy of negatively rated reviews in comparison to positively rated reviews. To better visualize this discrepancy, we will compare how VADER, the best performing lexicon-based sentiment analysis model and CNN (our deep learning sentiment analysis) model evaluate the four reviews selected from our consumer acceptance study.

3.5.2. Comparison of VADER and CNN Models

To facilitate a better comparison, the sentiment scores for VADER and CNN were normalized. The ranges for the sentiment scores were different across all four models. We normalized the sentiment scores calculated by the VADER and CNN model to 0.00 – 1.00. With 0.00 representing extremely negative sentiment and 1.00 representing extremely positive sentiment. There were four reviews selected that represent when the models calculated a false positive result (the panelist gave the sample a negative rating, but the SA model predicted positive sentiment) and false negative result (the panelist gave the sample a positive rating but the SA model predicted negative sentiment). To determine which words or phrases influenced the predicted sentiment scores for the CNN model we used the LIME models explanations (see Figure 6 and 7). In the case of the VADER model we inferred based on the context of the review which words and phrases would be labeled negative or positive in its sentiment dictionaries.

The panelist wrote a free comment review for this beer sample and gave it a rating of 1. This is the unabridged and unedited review:

"This one smelled nice, but it was absolutely distinguishing. I guess smells can be deceiving too. It smelled like it would have a fruit taste like maybe apricot or passion fruit, but it was so bitter I couldn't decipher any of that when I tasted it. Overall, did not like this beer."

The VADER model detected very positive sentiment from this review and calculated a normalized sentiment score of 0.85. However, the panelist gave the review above a very negative score. This is an example of a false positive result from the VADER model. By analyzing this review, we can infer that there were several words and phrases the VADER model may have categorized as positive and contributed to the positive sentiment score. For example, the panelist used the word "distinguishing" which generally has a positive connotation. However, in this context, we believe the panelist intended to use the word disgusting which has a more negative connotation. Furthermore, "smelled nice," "apricot," and "passion fruit" may have positively influenced the sentiment score.

The CNN model calculated a very negative normalized sentiment score of 0.17 that better captured the intent of the panelist. Furthermore, model accurately predicted "distinguishing" to have a negative connotation and it was a key contributor in the CNN model's final results. The words "smelled," "guess," and "not" carried the most weight towards the negative sentiment score (Figure 6 and 7).

In the second example the panelist gave this beer sample a rating of 8. The review is as follows:

"Dark, bitter, coffee, not sweet, no alcohol smell. Overall, this was a surprise to taste since it was so different from the others. Initially, I thought I didn't like it, but once I realized it was a different type of beer I could evaluate it differently. Very enjoyable. Overall, would love a full glass."

The VADER model detected negative sentiment from this review and calculated a normalized sentiment score of 0.29 whereas the panelist classified this review as positive – this is an example of a false negative result. To a human observer the last two sentences give the overall review a positive tone and indicates that the panelist had a positive disposition towards the beer. However, we suggest the words "dark, bitter, and not sweet" were considered negative by VADER's sentiment dictionaries and that heavily influenced the negative sentiment score. Something of interest to note, the panelist used the word bitter in a negative context, but, bitter is usually considered a positive adjective when reviewing beer. This panelist may not be a typical consumer of beer. Additionally, bitter is widely considered to have a negative connotation and we believe it was negatively labeled in the VADER models sentiment dictionaries.

The CNN model calculated an extremely positive normalized sentiment score of 0.98 and was correct in its prediction, the review does represent positive sentiment. The words "enjoyable," and "very" were the strongest contributors to the positive sentiment result. Interestingly, the words "dark, bitter, and not sweet" were not key words that influenced the CNN model's decision. We believe CNN model was able to accurately recognize strong positive sentiment because it detected the importance of the adjectives "enjoyable," and "very" in the last two sentences which were overwhelming positive in a somewhat negative leaning review (See Figure 6 and 7).

In the third example, the panelist rated the beer sample 1. The review is as follows:

"I smell hints of what seems to be liquid smoke or some Smokey source. Taste very woody and Smokey. Caught some notes of coffee and chocolate on the back end."

The VADER model calculated a normalized sentiment score of 0.50. The VADER model predicted a neutral sentiment from this review but it doesn't completely capture the strong negative sentiment in the lowly rated review. We suggest the words "smokey," and woody had a very negative connotation and the words "chocolate," and "coffee" had a positive connotation. If these words were weighted similarly it may account for the VADER model's neutral prediction.

In contrast, the CNN model predicted a false positive result: it calculated a very positive normalized sentiment score of 0.84 but the panelist rated this beer very negativily. However, The words "source" and "chocolate" were the strongest contributors to the positive sentiment result and, CNN detected the word "smokey" had a positive connotation. The CNN model may have learned that "smoke" and "smokey" were positive attributes from the training data set, however, this panelist intended "smokey" as a negative beer attribute.

In the fourth and last example, the panelist gave the beer a very high rating of 7. The panelist stated the following:

"The smell is of apple or ripe pear. The beer tastes like a watered down cider with a bit of bitterness or hoppy flavor at the end. This drink would be light and refreshing."

The VADER model calculated a normalized sentiment score of 0.47. It is difficult to pinpoint which words or phrases influenced the VADER model's evaluation in the last review. We suggest the CNN and VADER models both detected that "watered down" was a very negative phrase. However, we believe the VADER model identified "refreshing" and "light" (which CNN detected as negative) as positive words.

The CNN model predicted a false negative result: it calculated a very negative normalized sentiment score 0.06. The CNN model tagged "watered," and "cider" as having a negative connotation and they had the strongest weights (See Figure 6 and 7). "Refreshing" had the next strongest weight and was tagged as positive. However, approximately 2/3 of the key words identified by CNN had a negative connotation and we believe the majority influenced the CNN model's very negative prediction (See figure 6).

3.5.3. Advantages and Disadvantages of Sentiment Analysis

Our second objective was to examine the advantages and disadvantages of sentiment analysis as a data analysis tool in sensory evaluation. One advantage is there are sentiment analysis tools available for novice programmers or sensory scientists who are unexperienced with coding. In the preliminary stages of our data analysis and data

collection we used a website called orangedataming.com. The Orangedatamining application allows one to run VADER, Liu Hu and other sentiment analysis models on free comment data or short reviews like twitter comments. It is a free and easily accessible tool to use and it requires no prior knowledge with any programming languages. Secondly, sentiment analysis allows the sensory scientist to process a large volume of data very quickly. We were able to process 408 reviews with four different sentiment analysis models within a short period of time. Lastly, the deep learning CNN model we created was able to find differences in liking where the ratings did not. In a conventional consumer acceptance study we would compare the ratings to determine if there were any significant differences between the way the beers were liked (Fonseca et al., 2016a; Mahieu et al., 2020). In the between-subjects ANOVA the ratings had no significant effect for sample name. In contrast our model was able to distinguish specific differences between how the samples were liked using the sentiment scores. For example, All Day IPA and Just pressed beers were the most preferred and Budweiser and Miller Light beers were the least preferred. This could be very advantageous to sensory scientists who are interested analyzing specific dimensions of liking within one product space.

One of the disadvantages is that for optimal results a sensory scientist may need a large training data set or lexicon. If one uses a machine learning sentiment analysis model a training data set is required for the algorithm to find the best representation of the data in order to infer a set of rules which can be applied to new data. Our training data set was collected from online beer reviews. The dataset we used was collected from Apr 2000 - Nov 2011 and consists of 2,924,127 reviews, 40,213 users of which 4,798 of them had more than 50 reviews. We were able to easily collect high volume and high-quality data because we took advantage of online food reviews and blogs. If a sensory scientist did not have that option, curating a large enough training data set may be more difficult. Lexicon based sentiment analysis models are available with pre-made dictionaries. However, a lot of the adjectives that we use in relation to sensory characteristics and liking food are perceived differently in non-food related scenarios. A tailored dictionary to the food product of interest is very time consuming and labor intensive to create but, it is more likely to generate better results.

References

Agarwal, B., & Mittal, N. (2016). *Machine learning approach for sentiment analysis* (pp. 21–45). Springer.

Ahmad, M., Aftab, S., Muhammad, S. S., & Ahmad, S. (2017). Machine learning techniques for sentiment analysis: A review. *International journal of multidisciplinary sciences and engineering*, *8*(3), 27.

Al-Shabi, M. A. (2020). *Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining*. 7.

Carr, J., Decreton, L., Qin, W., Rojas, B., Rossochacki, T., & Yang, Y. wen. (2015). Social media in product development. *Food Quality and Preference*, *40*, 354–364. https://doi.org/10.1016/j.foodqual.2014.04.001

Chollet, F., & Allaire, J. J. (2018). *Deep Learning mit R und Keras: Das Praxis-Handbuch von den Entwicklern von Keras und RStudio*. MITP-Verlags GmbH & Co. KG.

Fonseca, F. G., Esmerino, E. A., Tavares Filho, E. R., Ferraz, J. P., da Cruz, A. G., & Bolini, H. M. (2016). Novel and successful free comments method for sensory characterization of chocolate ice cream: A comparative study between pivot profile and comment analysis. *Journal of Dairy Science*, *99*(5), 3408–3420.

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.

Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. 168–177.

Hutto, C. J. (2021). *Cjhutto/vaderSentiment* [Python]. https://github.com/cjhutto/vaderSentiment (Original work published 2014)

Lawless, H. T., & Heymann, H. (2010a). Acceptance Testing. In H. T. Lawless & H. Heymann (Eds.), *Sensory Evaluation of Food: Principles and Practices* (pp. 325–347). Springer. https://doi.org/10.1007/978-1-4419-6488-5_14

Lawless, H. T., & Heymann, H. (2010b). *Sensory evaluation of food: Principles and practices* (Vol. 2). Springer.

Liang, H., Sun, X., Sun, Y., & Gao, Y. (2017). Text feature extraction based on deep learning: A review. *EURASIP Journal on Wireless Communications and Networking*, *2017*(1), 1–12.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415–463). Springer.

MacFie, H. J., Bratchell, N., GREENHOFF, K., & Vallis, L. V. (1989). Designs to balance the effect of order of presentation and first-order carry-over effects in hall tests. *Journal of Sensory Studies*, *4*(2), 129–148.

Mahieu, B., Visalli, M., Thomas, A., & Schlich, P. (2020). Free-comment outperformed check-all-that-apply in the sensory characterisation of wines with consumers at home. *Food Quality and Preference*, *84*, 103937.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD '16*). Association for Computing Machinery, New York, NY, USA, 1135–1144. DOI:https://doi.org/10.1145/2939672.2939778

McAuley, J. J., & Leskovec, J. (2013). From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13*, 897–908. https://doi.org/10.1145/2488388.2488466

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. *Proceedings of the 7th ACM Conference on Recommender Systems*, 165–172. https://doi.org/10.1145/2507157.2507163

McAuley, J., Leskovec, J., & Jurafsky, D. (2012). *Learning attitudes and attributes from multi-aspect reviews*. 1020–1025.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Miller, C., Hamilton, L., & Lahne, J. (2021). Sensory Descriptor Analysis of Whisky Lexicons through the Use of Deep Learning. *Foods*, *10*(7), 1633.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning–based Text Classification: A Comprehensive Review. *ACM Computing Surveys (CSUR)*, *54*(3), 1–40.

Muir, D. D., & Hunter, E. A. (1991). Sensory evaluation of Cheddar cheese: Order of tasting and carryover effects. *Food Quality and Preference*, *3*(3), 141–145.

Peryam, D., & Girardot, N. (n.d.). I952. Advanced taste-test method. *Food Eng*, *24*, 58.

Prescott, J., Hayes, J. E., & Byrnes, N. K. (2014). Sensory Science. In *Encyclopedia of Agriculture and Food Systems* (pp. 80–101). Elsevier. https://doi.org/10.1016/B978-0-444-52512-3.00065-6

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Rinker, T. (2021). *Sentimentr* [R]. https://github.com/trinker/sentimentr (Original work published 2015)

Samal, B., Behera, A. K., & Panda, M. (2017). Performance analysis of supervised machine learning techniques for sentiment analysis. *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)*, 128–133.

Shieber, S. M. (2004). *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. MIT Press.

Snuggs, S., & McGregor, S. (2021). Food & meal decision making in lockdown: How and who has Covid-19 affected? *Food Quality and Preference*, *89*, 104145.

Symoneaux, R., Galmarini, M., & Mehinagic, E. (2012). Comment analysis of consumer's likes and dislikes as an alternative tool to preference mapping. A case study on apples. *Food Quality and Preference*, *24*(1), 59–66.

Visalli, M., Mahieu, B., Thomas, A., & Schlich, P. (2020). Automated sentiment analysis of Free-Comment: An indirect liking measurement? *Food Quality and Preference*, *82*, 103888. https://doi.org/10.1016/j.foodqual.2020.103888

Yelp—Company—Fast Facts. (n.d.). Retrieved September 15, 2021, from https://www.yelp-press.com/company/fast-facts/default.aspx

Zhang, L., & Liu, B. (2014). Aspect and Entity Extraction for Opinion Mining. In W. W. Chu (Ed.), *Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities* (pp. 1–40). Springer. https://doi.org/10.1007/978-3-642-40837-3_1

Chapter 4: Conclusions and Future Work

Overall, the findings from this study demonstrate there are many advantages to using sentiment analysis in sensory evaluation study. We encourage more sensory scientists to consider using sentiment analysis to gain multiple layers of interpretation to their free comment data. All the sentiment analysis models we tested had weak correlations. The simple linear regression we used may not have been the best model to fit the data. The models were generally able to detect extremely positive or extremely negative sentiment however, the traditional 9pt scale does not align well with this binary classification. This does raise the question: "is there a significant difference between a 6 rating (slightly positive) and a 8 rating (very positive) in the likelihood to purchase that product or the important sensory characteristics that drive liking?" Furthermore, our deep learning CNN model was superior at detecting specific differences in liking between the six beer samples where the ratings could not. Using the sentiment scores, we were able to find instances where the panelists liked some of the beer samples over others. We believe this suggests there is more in-depth information gained about consumer acceptance from free comment data in comparison to relatively simple scale. Additionally, the sentiment analysis models performed better on negative reviews than positive reviews. We believe there are benefits to creating a SA model in comparison to using a pre-made SA model. The results from this study highlight the many advantages and disadvantages to using sentiment analysis in sensory evaluation. We suggest sensory scientists with a large amount of data at their disposal and some interest or aptitude for programming languages would have the easiest time learning how to use sentiment analysis and would reap the most benefits in their data analysis.

Future opportunities to build upon this work include facilitating a better comparison between the sentiment analysis models. Our machine learning model was trained using online beer reviews and three of the beer samples we used in our free comment data collection task came from this test data set. In contrast, the three lexicon-based models: VADER, Liu Hu and SentimentR used premade generic dictionaries. We believe a beer-based lexicon or dictionary would have greatly improved the performance of these sentiment analysis models and may have enabled a better comparison. To add further data to support our exploration of the advantages and disadvantages of sentiment analysis, a comparison between traditional free comment analysis carried out by hand and the sentiment analysis models would have been an interesting way to analyze accuracy with a benchmark standard data analysis method. We also suggest using an untrained human evaluator with no context to predict what the panelist rated each review and compare their predictions to the predicted sentiment scores. We would present this evaluator with the free comment reviews and ask them to judge the emotional intent of the author using the traditional 9pt scale. The ratings from the untrained evaluators could be compared to the original reviews and the sentiment scores.

Figures, Tables, and Appendix

Figures

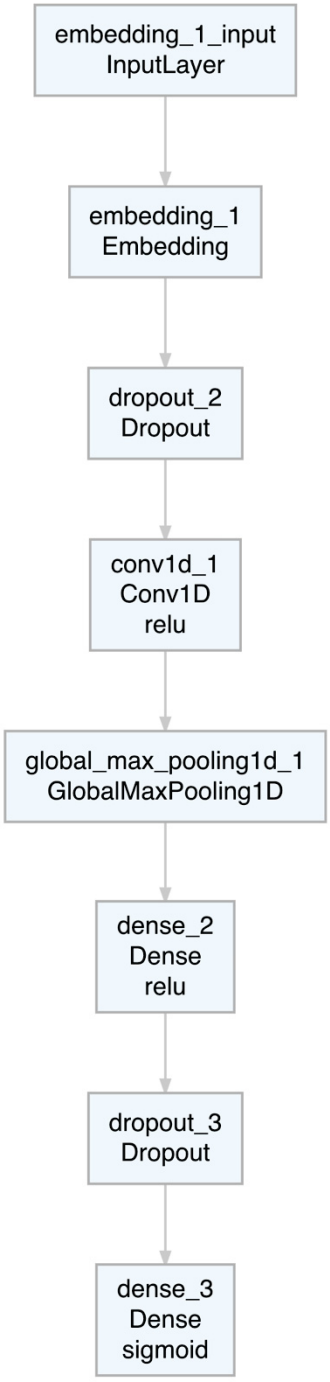


Figure 1. Flow chart of the multilayer neural network we created using one-dimensional convolutions for sequence processing The Convolutional neural networks (CNN) has hidden layers called convolutional layers which detect patterns in the text and find useful

representations of the data The output of this process creates a matrix which is called a 3D tensor or input feature map. The output feature map is composed of filters – filters are responsible for encoding specific aspects of the input data such as: semantic signals or emotions (e.g. happiness, disgust, sadness etc.).

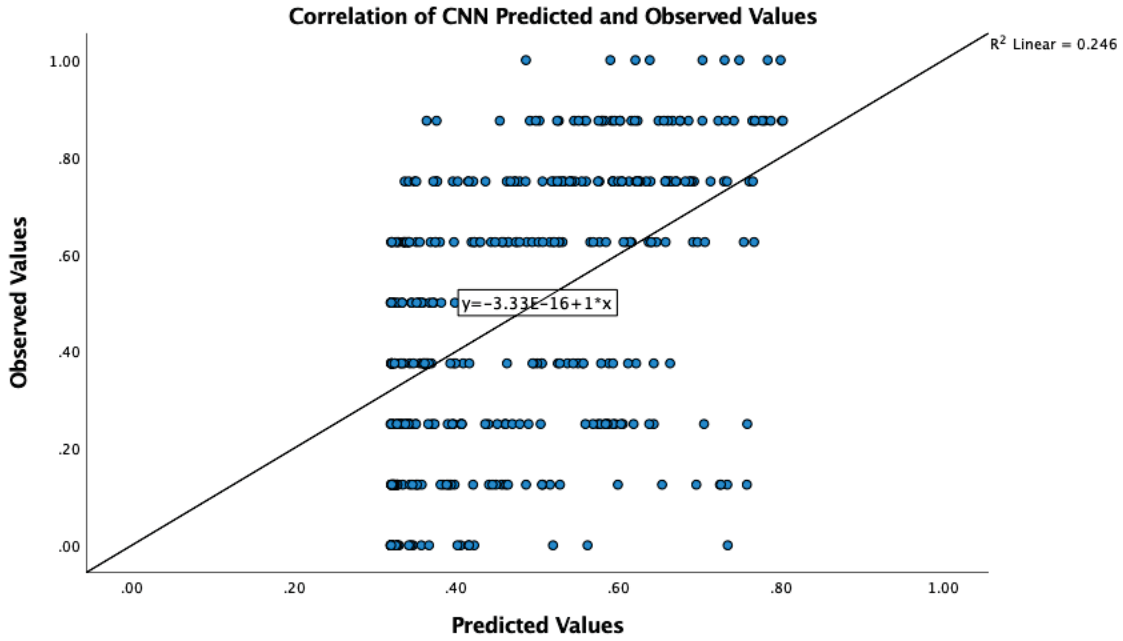


Figure 2. Scatterplot of the predicted ratings or observed values calculated from the simple linear regression of the convolutional neural network (CNN) model sentiment scores (X) and the ratings (Y). The rating and sentiment scores were normalized from 0.00-1.00.

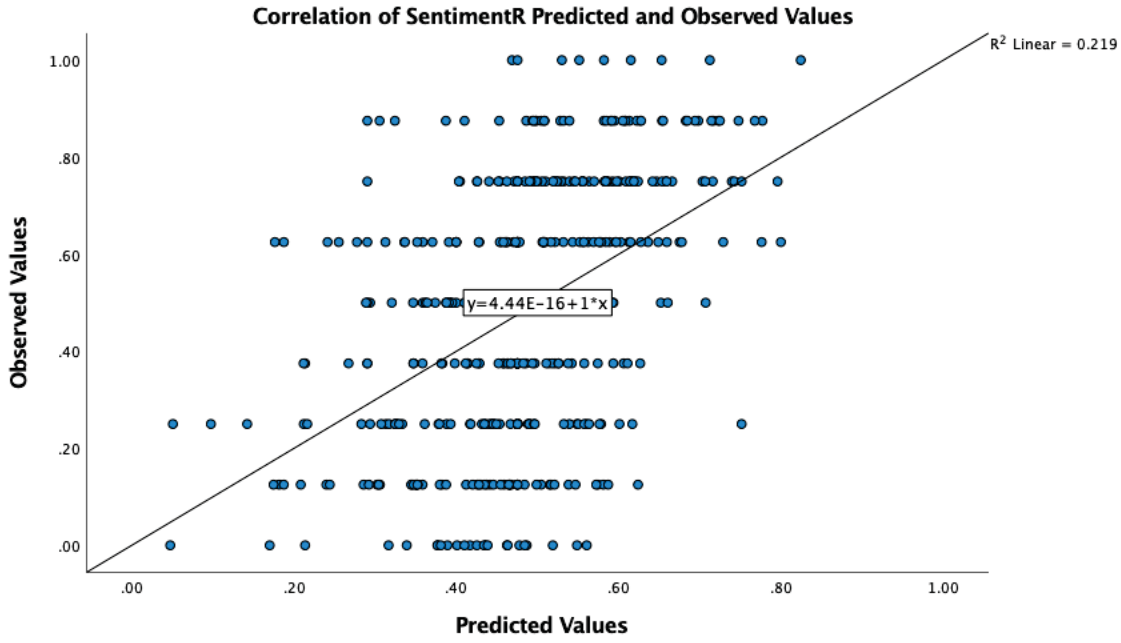


Figure 3. Scatterplot of the predicted ratings or observed values calculated from the simple linear regression of the SentimentR model sentiment scores (X) and the ratings (Y). The rating and sentiment scores were normalized from 0.00-1.00.

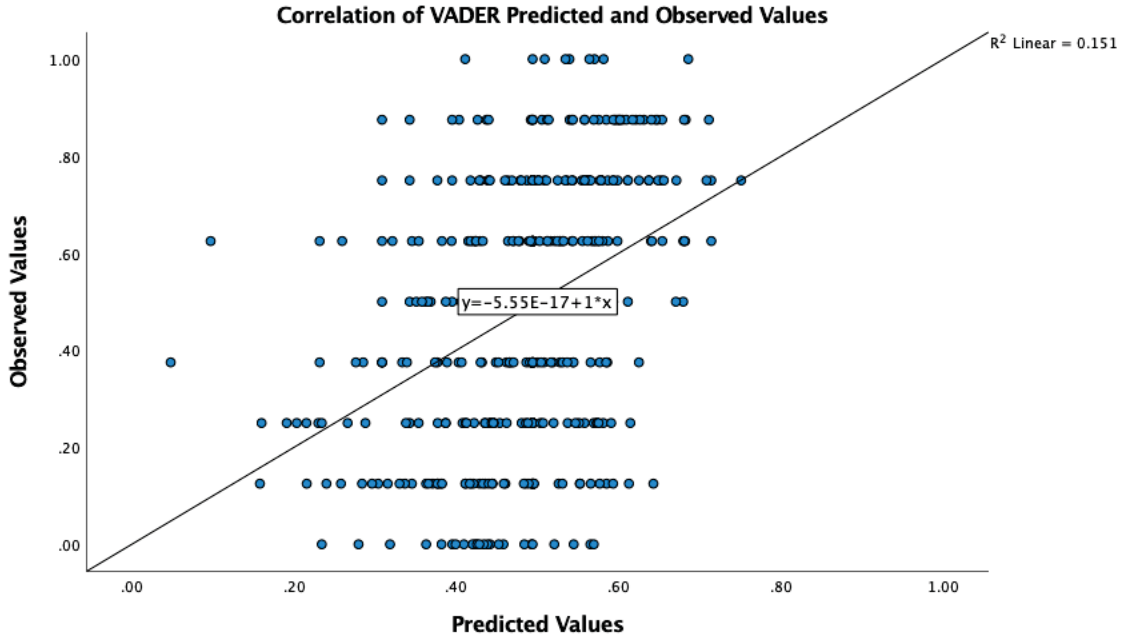


Figure 4. Scatterplot of the predicted ratings or observed values calculated from the simple linear regression of the VADER model sentiment scores (X) and the ratings (Y). The rating and sentiment scores were normalized from 0.00-1.00.

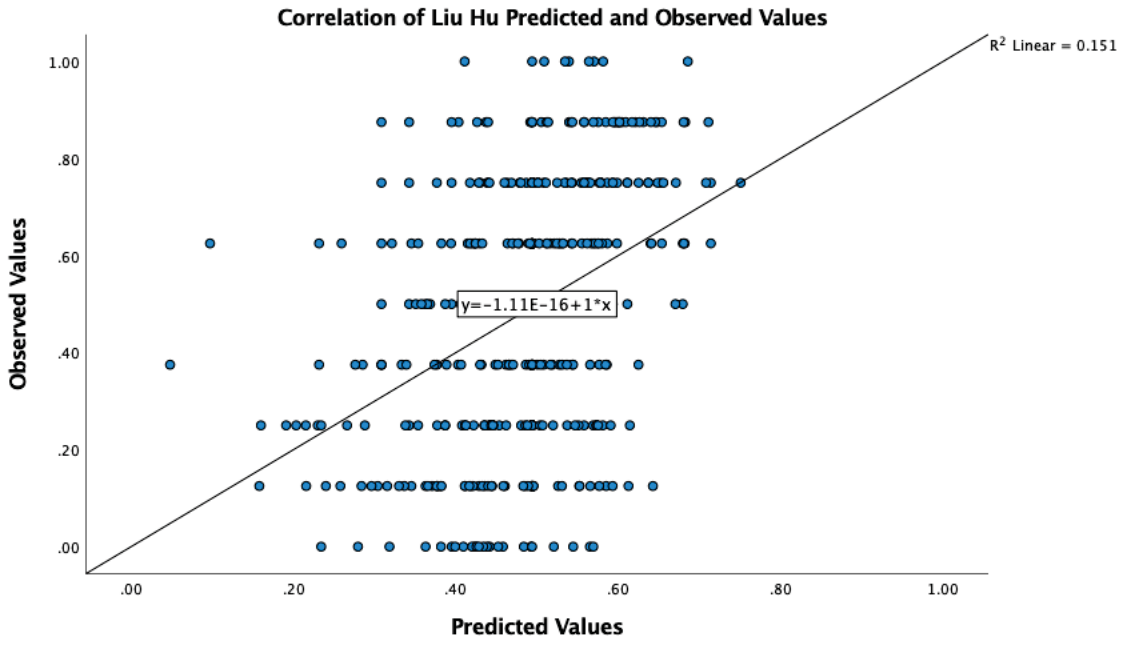


Figure 5. Scatterplot of the predicted ratings or observed values calculated from the simple linear regression of the LiuHu model sentiment scores (X) and the ratings (Y). The rating and sentiment scores were normalized from 0.00-1.00.

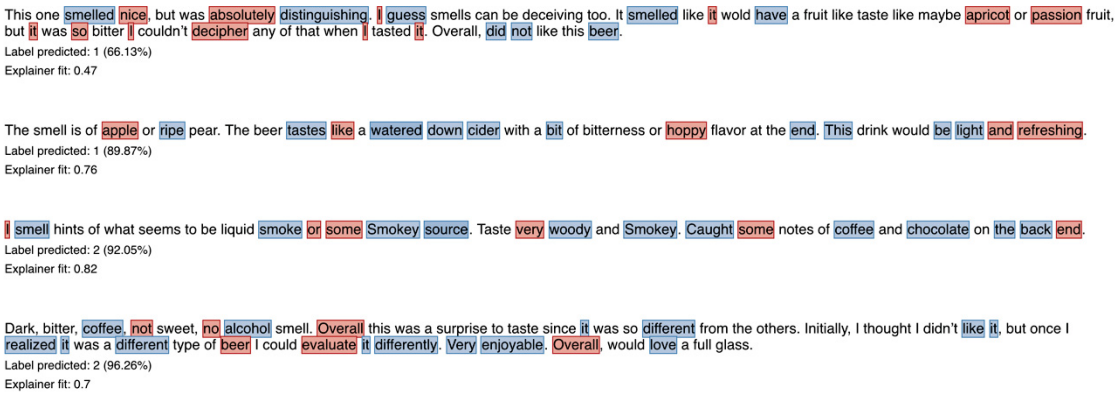


Figure 6. Visual representation of the LIME Model interpreting the results from the CNN our deep learning model. Four examples were chosen. The label predicted represents positive (1) and negative (2). The blue highlight represents what the LIME model indicates as a word or phrase that contributes to the label or results. The red highlight represents a word or phrase that contradicts the label or results.

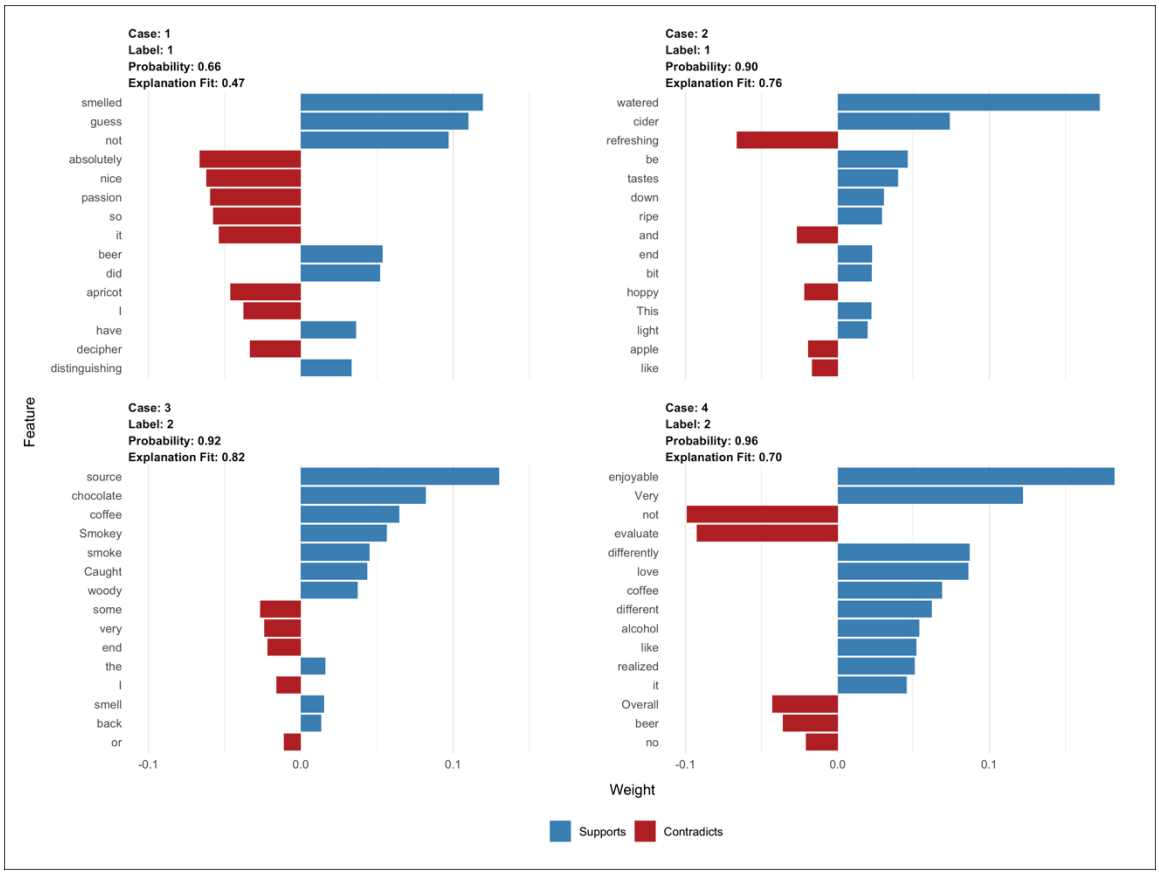


Figure 7. Graphical representation of the LIME Model interpreting the results from the CNN our deep learning model. Four examples were chosen. The label predicted represents positive (1) and negative (2). The blue highlight represents what the LIME model indicates as a word or phrase that contributes to the label or results. The red highlight represents a word or phrase that contradicts the label or result. The weight represents the strength of how much the word influenced the label or result.

Tables

Table 1. Example of the data from the Ratebeer data set which includes the name of the beer, beer style, overall score and review text.

Beer Name	Beer Style	Overall Score	Review Text
John Harvards Simcoe IPA	India Pale Ale	13/20	On tap at the Springfield, PA location. Poured a deep and cloudy orange (almost a copper) color with a small sized off white head. Aromas or oranges and all around citric. Tastes of oranges, light caramel and a very light grapefruit finish. I too would not believe the 80+ IBUs - I found this one to have a very light bitterness with a medium sweetness to it. Light lacing left on the glass.
Lucky Bucket IPA	India Pale Ale	15/20	Dripping with resinous hops, this IPA packs a surprisingly bitter punch, wringing every last west coast IBU from the recipe. The malts give a grainy, meek background, one barely noticeable in the face of such massive piney massiveness. Good nose, even with a hint of metallic ick. The texture isnt that great--the beer seems overcarbonated and spiky, too harsh when coupled with a sandpaper-dry finish. Good but one is plenty.

Table 2. List of beer sample names used in free comment data collection study, beer style, ABV%, Price and reason for selection.

Beer Name	Beer Style	ABV %	Price	Reason for Selection
Just Pressed	Sour ale	4.6	1, 12-oz can for $2.15	Recommended by local liquor store
All Day IPA from Founders brewery	American India pale ale	4.7	1, 12-oz can for $1.75	Recommended by local liquor store
AleSmith Speedway	Imperial Stout	12	1, 16-oz can for $4.50	Selected from test dataset
Miller Lite	Pilsner Lager	4.2	24, 12-oz cans for $16.00	Selected from test dataset
Coors Light	American light lager	4.2	24, 12-oz cans for $16.00	Selected from test dataset
Budweiser	American Lager	5	24, 12-oz cans for $16.00	Recommended by local liquor store

Table 3. Confusion matrix

		Predicated Values		
		Positive	Negative	Precision TP/(TP + FP)
Actual Values	Positive	True Positive (TP)	False Negative (FN) *Type II Error*	
	Negative	False Positive (FP) *Type I Error*	True Negative (TN)	
		Recall TP/(TP+FN)		Accuracy (TP+TN)/(TP+FP+TN+FN)

Table 4. Descriptive statistics of CNN (our model), Liu Hu model, the ratings from the traditional 9-pt scale, SentimentR model, and VADER model

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
CNN	408	0.001	0.983	0.33146	0.013746	0.277664
Liuhu	408	-1.699	1.581	-0.04386	0.020577	0.415629
Rating	408	1	9	4.86	0.109	2.201
SentimentR	408	-1.114	1.265	0.0123	0.012846	0.259476
Vader	408	-0.911	0.967	0.18888	0.023097	0.466528
Valid N (listwise)	408					

Table 5. Simple Linear Regressions of the sentiment scores calculated by the four sentiment analysis models (CNN, SentimentR, Vader, Liu Hu) as the independent variable and the ratings from the traditional 9-pt scale (1-9) as the dependent variable

		Unstandardized	Coefficients	Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	3.548	0.148		24.052	<.001
	CNN	3.944	0.341	0.496	11.555	<.001
2	(Constant)	4.808	0.097		49.509	<.001
	SentimentR	3.887	0.374	0.458	10.386	<.001
3	(Constant)	4.446	0.105		42.514	<.001
	Vader	2.169	0.208	0.46	10.429	<.001
4	(Constant)	4.946	0.101		48.969	<.001
	Liuhu	2.07	0.242	0.391	8.553	<.001

Table 6. accuracy, precision and recall calculations. N represents the total number of negatively rated reviews. P represents the total number of positively rated reviews. Tp represents the # of positive sentiments detected correctly. Tn represents the # of negative sentiments detected correctly.

Model		N	P	Tn	Negative Accuracy	Tp	Positive Accuracy	Total	Accuracy %	Precision	Recall
1	CNN	227	181	193	85%	90	50%	283	69%	73%	50%
2	Vader	227	181	122	54%	69	38%	191	47%	40%	38%
3	Sentiment R	227	181	97	43%	61	34%	158	39%	32%	34%
4	Liu Hu	227	181	102	45%	53	29%	155	38%	30%	29%

Table 7. Between subjects ANOVAs

Tests of Between-Subjects Effects

Dependent Variable	Source		Type III Sum of Squares	df	Mean Square	F	Sig.
CNN	Sample_Name	Hypothesis	7.32	5	1.46	27	<.001*
		Error	18.18	335.12	0.05		
	Panelist_Name	Hypothesis	5.01	63	0.08	1.5	0.019*
		Error	17.1	315	0.05		
Liuhu	Sample_Name	Hypothesis	1.55	5	0.3	2.07	0.069
		Error	49.31	329.37	0.15		
	Panelist_Name	Hypothesis	18.67	63	0.3	1.95	<.001*
		Error	47.98	315	0.15		
SentimentR	Sample_Name	Hypothesis	0.4	5	0.08	1.33	0.252
		Error	19.78	330.98	0.06		
	Panelist_Name	Hypothesis	6.93	63	0.11	1.82	<.001*
		Error	19.1	315	0.06		
Vader	Sample_Name	Hypothesis	2.2	5	0.44	2.42	0.036
		Error	61.43	337.55	0.18		
	Panelist_Name	Hypothesis	24.16	63	0.18	2.13	<.001*
		Error	56.7	315	0.18		
Rating	Sample_Name	Hypothesis	43.363	5	8.67	1.92	0.09
		Error	1495.22	331.21	4.51		
	Panelist_Name	Hypothesis	409.91	63	6.51	1.43	0.27
		Error	1431.96	314.09	4.56		

Table 8. Pairwise comparisons from the between-subjects ANOVA on the CNN model. The table displays significant mean differences between the sentiment scores of the six beer samples.

Dependent Variable: CNN

(I) Sample Name	(J) Sample Name	Mean Difference (I-J)	95% Confidence Interval for Difference[b] Std. Error	Sig.[b]
All Day IPA	Budweiser	0.254	0.041	<.001
	Coors Light	0.161	0.041	0.009
	Miller Light	0.149	0.041	0.018
	Speedway	-0.145	0.041	0.024
Budweiser	All Day IPA	-0.254	0.041	<.001
	Just Pressed	-0.302	0.041	<.001
	Speedway	-0.399	0.041	<.001
Coors Light	All Day IPA	-0.161	0.041	0.009
	Just Pressed	-0.208	0.041	<.001
	Speedway	-0.306	0.041	<.001
Just Pressed	Budweiser	0.302	0.041	<.001
	Coors Light	0.208	0.041	<.001
	Miller Light	0.196	0.041	<.001
Miller Light	Just Pressed	-0.196	0.041	<.001
	Speedway	-0.294	0.041	<.001
Speedway	All Day IPA	0.145	0.041	0.024
	Budweiser	0.399	0.041	<.001
	Coors Light	0.306	0.041	<.001
	Miller Light	0.294	0.041	<.001

Based on estimated marginal means

* The mean difference is significant at the .05 level

[b]Adjustment for multiple comparisons: Bonferroni

Appendix

1.Pre-Screener

Q1 Thanks for your interest in participating in research to investigate the degree consumers dislike or like the selection of beers provided. This research will also help further our understanding of what beer attributes are important to consumers and how that impact liking. To apply for the research study, complete this survey. If you qualify, you will be contacted with further information to sign up for a time to complete the study at the Virginia Tech Sensory Evaluation Lab in HABB1. For more information or questions, please contact Elyse Canty (ecanty@vt.edu), or Jacob Lahne, (jlahne@vt.edu).

Q2 What is your age?
*Please note, if you are under 21 we cannot accept your participation due to federal law. You will be required to present a valid, government-issued ID before the start of the focus group session.

- ○ under 21 (1)
- ○ 21+ (2)

Q5 Have you ever consumed beer?

- ○ Yes (4)
- ○ No (5)

Q3 Do you have an allergy to beer or beer stuff?

- ○ Yes (1)
- ○ No (2)

Q4 Do you self report as currently being pregnant?

GOVERNMENT WARNING: (1) According to the Surgeon General, women should not

drink alcoholic beverages during pregnancy because of the risk of birth defects. (2) Consumption of alcoholic beverages impairs your ability to drive a car or operate machinery, and may cause health problems.

○ Yes (1)

○ No (2)

Q5 Do you consider yourself a beer enthusiast?

○ Yes (1)

○ No (2)

Q6 How many beers do you consume in a week?

2.Data Collection Template

Instructions Screen Text:

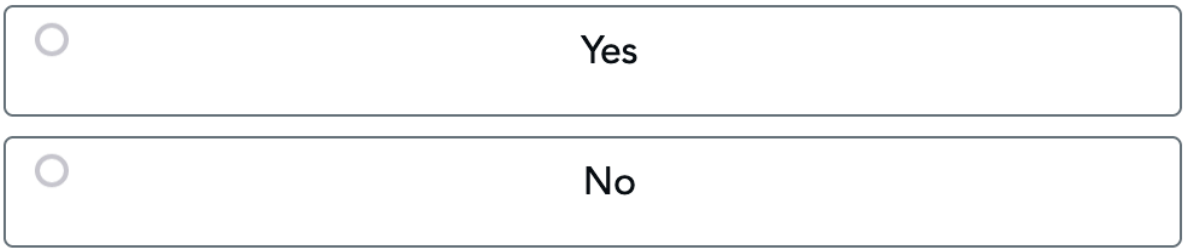
- You will be asked to evaluate and give comments on each beer sample, (2) rate your overall liking by checking "yes" or "no" and (3) rate your liking of the beer sample on a 9-point scale.
- You will be presented with one beer sample at a time, crackers, water, and a napkin in front of you on a tray. We will replenish your water and crackers as needed, when you pass back your tray. Please taste each beer and complete each task. After tasting the sample, please expectorate (spit out) after tasting the sample into the spit-cup.
- After you have completed all three tasks for a sample, please pass your tray through for your next sample. You will be given 60 seconds in between samples. In this time please rinse your mouth water and/or eat part of a cracker to cleanse your palate.

HIT NEXT WHEN COMPLETED.

Task 1: Free Comment Task
-Instructions Screen Text:
Please lift off the watch glass, smell and taste the sample.
In your own words describe "What are your overall impressions and feelings about the sample?"
Please, spit out the samples into the spit-cup after tasting.
When you are finished, pass the glass through the hood and wait for the next sample.

HIT NEXT WHEN COMPLETED.

Task 2: Overall Liking
- Instructions Screen Text:
Overall, did you like this beer?



HIT NEXT WHEN COMPLETED.

Task 3: 9-Point Hedonic Scale

- **Instructions:** How would you rate your overall liking of this beer sample? Please spit out samples into the spit cup after tasting.
- **Scale:** Sample 9-Point Hedonic scale:

Dislike Extremely	Dislike very much	Dislike Moderately	Dislike Slightly	Neither like nor dislike	Like Slightly	Like Moderately	Like Very Much	Like Extremely

HIT NEXT WHEN COMPLETED.

BREAK SCREEN (shown for 60 seconds between each of the six samples):

Text: Please pass your tray back for your next sample. During this break please drink water and/or eat part of a cracker to cleanse your palate. Please select next after 60 seconds and when you have received your next sample.

** After 60 seconds panelist can select next**

End Screen:
Text: Thank you for completing this test!