ERASMUS UNIVERSITEIT ROTTERDAM

Uncovering factors influencing the low adoption of sustainable cleaning products

A Text Analytics Approach

Erasmus School of Economics
Master thesis: Data Science and Marketing Analytics

Paula Alonso Campos
Student Number: 657458

Supervisor: Prof. Vahe Avagyan
Second supervisor: Prof. Michiel van Crombrugge

July 19, 2023

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Vahe Avagyan, for his genuine support, help and Wednesday's feedback meetings during the process of the thesis. I am also grateful to my professors during this year for their guidance and expertise, which enabled me to acquire knowledge beyond measure.

Additionally, this journey would not have been possible without the invaluable financial support from Group NN, who also offered students like me the valuable opportunity of having a mentor.

Lastly, I would like to extend my heartfelt gratitude to my family and friends, especially my parents, sister, boyfriend, and grandparents. Their constant belief in me, love and unwavering support have always kept my spirits and motivation high.

This journey would not have been possible without them, and I am forever grateful.

Agradecimientos

Quiero expresar mi más profunda gratitud a mi supervisor, el Prof. Vahe Avagyan, por su genuino apoyo, ayuda y reuniones de retroalimentación cada miércoles durante el proceso de la tesis. También agradezco a mis profesores durante este año su orientación y experiencia, las cuales me han permitido adquirir un conocimiento incalculable.

Además, este camino no habría sido posible sin el increíble apoyo financiero de Group NN, quienes también brindaron a estudiantes como yo la valiosa oportunidad de contar con un mentor.

Por último, me gustaría extender mi más sincero agradecimiento a mi familia y amigos, especialmente a mis padres, hermana, pareja y abuelos. Su constante creencia en mí, amor y apoyo inquebrantable siempre han mantenido altos mi espíritu y motivación.

Sin ellos, este camino no habría sido posible y estoy eternamente agradecida.

Abstract

This master's thesis explores the resistance to adopting sustainable cleaning products and uncovers crucial determinants influencing customers' attitudes and purchase decisions. This study analyses customer reviews from Amazon and compares sustainable and non-sustainable cleaning alternatives, which provides a comprehensive understanding of consumer preferences and satisfaction beyond the star rating. The findings reveal that pre-purchase expectations, product quality and advertisement viridity are crucial in building trust and loyalty. Among the provided recommendations, sustainable companies should ensure accurate and transparent advertisements to prevent negative word-of-mouth. Product quality can be addressed by improving packaging, preventing leakage, developing soft fragrance variety, and exploring disinfectant ingredients and odour controllers. By enhancing product quality and brand image, companies can increase consumer satisfaction and gain a competitive advantage, leading to increased adoption rates of sustainable cleaning products.

Keywords: customer satisfaction, unstructured data, user-generated content, brand recognition, sustainable consumption, Random Forest, penalised regression.

Table of Contents

1. Introduction

In recent years, environmental deterioration and climate change have given rise to concepts such as eco-friendly, eco-innovation, environmentally conscious or green consumerism. As consumers and companies have recognized the consequences of their choices on the environment, sustainability has become a major concern and focus area. For this reason, one of the most important issues nowadays for many companies and governments is to understand customers' perceptions of sustainable products as well as promote more sustainable consumption. This is crucial to motivate a growing demand for green products, which could prevent or decrease ecological deterioration. However, companies face the challenge of understanding customers' perceptions to develop effective marketing campaigns, make strategic decisions and improve or maintain their brand's reputation or image.

To address this challenge, unstructured text analytics can be a powerful tool. Unstructured data comprises textual data such as open-ended survey responses, social media conversations, and customers' reviews, but it can also be non-textual such as images, video, and audio data. Unstructured text analytics offers the possibility to study this information by applying methodologies such as Natural Language Processing (NLP), which has been proven to provide valuable insights for new product development and brand reputation understanding. The first enables companies to stay competitive as well as meet customers' expectations and needs (Markham et al., 2015) and the second helps them to comprehend their SWOT - Strengths, Weaknesses, Opportunities and Threats - and develop a successful strategy for the future (Krawczyk & Xiang, 2016). Authors such as Balducci & Marinova, (2018) or Boegershausen et al., (2022) have also discussed the key role of unstructured web data in various industries and its reshaping business practices potential. However, they highlight an underutilization of available data by many companies to drive decisions.

Berger et al. (2020) provided an overview of how text analysis can generate marketing insights and they differentiated between text used for prediction or understanding. The authors provided an overview of methodologies, metrics, and challenges when using text data and highlight the importance of accessing what people say and how it is said. Furthermore, several studies have applied text mining in the marketing field. Archak et al. (2011) used text mining techniques to include review data in a consumer choice

model and study customer preferences and future sales. Similarly, Chen & Lee (2023) combined physician rating data from Yelp with data from Medicare to study whether ratings are correlated with physicians' quality and whether online ratings influence patients' choices. Radojevic et al. (2017) collected survey data including hotels customers reviews from TripAdvisor and used a multilevel modeling framework to study factors affecting customer satisfaction. A slightly different approach was taken by Mishra (2022), who performed Latent Dirichlet Allocation on mined Twitter data to study customer experience related to online shopping. Moreover, Ghose et al. (2012) made also use of user-generated content (UGC) on social media to help design ranking systems for hotels on travel search engines to improve user satisfaction. Similarly, UGC was also studied in relation to stock market performance (Tirunillai & Tellis, 2012). Finally, Lee & Bradlow (2011) utilized online customer reviews to create a methodology for automated marketing research.

As can be observed, existent literature has examined how online consumer reviews may be a substitute for traditional word-of-mouth recommendations. The fact that potential buyers seek out information from reviews emphasises the critical role of online opinions in defining consumer behaviour (Chevalier & Mayzlin, 2006; Zhu & Zhang, 2010; Kostyra et al., 2016; Netzer et al., 2012). Additionally, there is another factor that influences consumers' purchase decisions, which is brand reputation. Brand reputation has been proven to influence customer loyalty (Mazurek, 2019; Oppong & Caesar, 2023), which can lead to positive perceptions among consumers and, therefore, to a potential increase in sales.

However, although different industries have been inspected, the role of unstructured data in understanding the sustainable cleaning products domain remains unexplored. Given the rising popularity of sustainable alternatives, it is crucial to investigate which factors inhibit their adoption, particularly within the cleaning sector. Despite the growing demand for sustainable alternatives, the adoption of these products is still relatively low compared to conventional products (Long, 2018). Previous studies in this field have also focused on individual features or products, overlooking a comprehensive analysis of customer satisfaction and the underlying reasons for it.

Customer satisfaction plays a critical role in shaping customer purchase decisions since it encircles not only the overall customer happiness with a product but also those factors that make the difference between a positive and a negative experience.

In general, even when people are not sure about their motives or expressions, the language they use acts as an explanation and a fingerprint (Pennebaker, 2011).

The structure of this research paper is as follows. The next section is going to delve into the problem that is going to be investigated, which is followed by a discussion of the relevance of the topic for society, management, and academia. Subsequently, a theoretical framework is presented, where relevant theories and previous studies are explored. The next section provides an overview of the data, collection technique and initial cleaning. The methodology section describes further cleaning steps by introducing NLP techniques as well as the methods used for data visualization and customer satisfaction predictions. The next section presents the results of visualization and model predictions, which is followed by a discussion addressing the research questions, managerial implications, and academic contributions. Finally, the study limitations, a bibliography and an appendix for further reference conclude this analysis.

2. Problem definition & objective of this research

Where is the resistance to adopting sustainable cleaning products coming from? What is stopping consumers to embrace sustainable cleaning options in their everyday lives? And, most importantly, how can this resistance be scaled down?

Cleaning products, particularly, have a significant impact on the environment since they usually contain ingredients such as surfactants, solvents, colourants, or fragrances that can be harmful to both environment and human health. For this reason, helping to switch to sustainable alternatives is crucial. Therefore, the purpose of this study is to understand the preferences, attitudes, and sentiments of customers concerning sustainable cleaning products by comparing their reviews to reviews of non-sustainable options from Amazon.

As aforementioned, existent literature extracted info for a specific product or a specific feature, however, I decided to broaden the scope and focus on the mining and analysis of a diverse array of products inside the cleaning sector. To do this, a control group of non-sustainable alternatives is analysed as well to make comparisons between both options. This allows for uncovering differences in strengths and weaknesses as well as

factors that influence customers' decisions, which have not been explored before. The reason is that, by analysing multiple products, a more comprehensive understanding of consumer preferences regarding environmentally friendly alternatives can be achieved. Additionally, comparing sustainable and non-sustainable alternatives helps discover differences in strengths and weaknesses between the two options as well as factors that influence customers' decisions. Uncovering these factors is fundamental when addressing barriers to adoption and will not be possible without studying the competitor.

To accomplish this, a combination of methods is applied. First, Multidimensional Scaling (MDS) is executed, which helps to inspect and interpret the data. Sentiment Analysis is performed both on a review and a sentence level by applying the polarity function. Then, topics extracted with Latent Dirichlet Allocation (LDA), Principal Component Analysis (PCA) factors, most frequent words in the data, bi-grams, emotions, and word embeddings are combined to make predictions about customers' satisfaction. This approach was not considered before in the existing literature and it is a powerful tool to investigate data and customers' satisfaction beyond star rating, which provides limited information. Therefore, this study can provide insights into how sustainable companies can enhance brand recognition, which will help to differentiate their products and create a competitive advantage. Furthermore, by identifying those factors that resonate with customers, these companies can tailor their brand messaging, product development and communication strategies.

Consequently, the following research question and contributions are proposed:

What are the key determinants that positively influence customers' decision-making regarding the selection of sustainable cleaning products? Conversely, what are the factors that contribute to the relatively low adoption rates of these alternative options?

- Enhancing brand and product recognition: *How could these companies improve brand reputation and differentiate their products to gain a competitive advantage over non-sustainable ones?*
- Actionable insights for refining and optimizing marketing efforts: *What are the factors that can be derived from this analysis to refine and optimize marketing efforts for companies producing sustainable cleaning products? Furthermore, how can these factors be addressed and resolved to enhance their marketing strategies effectively?*

2.1. Relevance of the topic

2.1.1. Society's contribution

The proposed research question aims to address the existing problem of low adoption rates of sustainable cleaning products. Despite the growing demand for sustainable alternatives, a significant proportion of customers end up not fully converting into this sustainable consumption, that is, not becoming loyal or long-term customers. This poses a threat to the environment and human health since traditional cleaning products contain harmful chemicals that can result in air and water pollution as well as long-term effects on public health. Contrary, sustainable options consider safer production materials and more environmentally respectful and clean practices that embrace products' life-cycle perspectives (Long, 2018). By understanding the motives of low adoption rates and promoting sustainable cleaning alternatives, this study can contribute to public health improvement and environmental sustainability.

2.1.2. Management contribution

Sustainable companies usually face challenges in gaining market share over conventional ones. Sustainable cleaning product producers might struggle to compete with established brands that produce conventional options, which limits the innovation and development of more responsible alternatives. Consequently, the transition to a more sustainable feature will slow down. By using sentiment analysis and advanced machine learning techniques, this study provides actionable insights to help companies optimise their marketing efforts and further understand customer satisfaction. In the end, this can lead to increased adoption and long-term usage of sustainable cleaning products, which will improve their market share. Recommendations to build trust and loyalty regarding product quality improvement, product development and advertisement transparency are provided.

2.1.3. Academia contribution

First, this study addresses the gap in understanding the low adoption rates of sustainable cleaning products, as well as the barriers and motivations for sustainable behaviour. This contributes to the existing literature in the field of sustainable consumption by providing a deeper understanding of the factors affecting customer choices. Second, this

combination of machine learning, text mining and text analytics methods, including web scraping, Multidimensional Scaling, Sentiment Analysis, Latent Dirichlet Allocation, Principal Component Analysis, and word embeddings, contributes to the field of Natural Language Processing and Data Analytics. The use of web scraping and powerful text preprocessing techniques to extract and clean the data from Amazon reviews also presents a valuable dataset for analysis and contributes to the existing literature on text mining (Chen & Lee, 2023; Kostyra et al., 2016; Netzer et al., 2012; Rust et al., 2021; Zhu & Zhang, 2010).

Moreover, considering a wider and more diverse range of products within the cleaning sector as well as making a comparison between sustainable and non-sustainable options, fills a gap in the literature, which often focuses on a specific product or feature and does not provide this comparison (Archak et al., 2011; Kostyra et al., 2016; Lee & Bradlow, 2011; Radojevic et al., 2017). Finally, this research contributes to the field of marketing and business management due to its practical implications and offered guidance for companies producing sustainable cleaning products seeking to effectively position their sustainable products.

In summary, this research is relevant to society since it can contribute to environmental sustainability and public health improvement. It is significant for companies as it can provide actionable insights to refine and optimise marketing strategies, which could potentially improve the competitiveness of sustainable companies. Lastly, it is relevant for academia as it contributes to the understanding of consumer behaviour towards sustainable products while contributing to the literature on web scraping, green marketing, relationship marketing and sustainability.

3. Theoretical background

Having discussed the problem that is going to be addressed and the implications of this research paper, a solid foundation and understanding of the subject matter is provided in this section. The aim of exploring previous studies and existing theories is to shed light on the fundamental principles and perspectives that guide this paper.

3.1.Web Scraping and text mining of unstructured data for marketing.

As of April 2023, the number of internet users is 5.18 billion users while the number of social media users remains quite close, 4.8 users. This represents an important proportion of the total world population, which is 8.03 billion (DataReportal, 2023). This massive digitalisation leads to a huge volume of web data that can be collected for marketing purposes. One of the main tools is web scraping, which is the process of creating software that automatically extracts information from multiple web pages. This technology presents a powerful advantage given the time and effort saved compared to performing it manually.

However, efforts have been made to study the validity of data sets generated through application programming interfaces (APIs) and web scraping (Boegershausen et al., 2022) considering the data sources, the data collection and the data extraction. The authors highlight the importance of applying the best practices to ensure reliability. Regarding legal and ethical considerations such as intellectual property rights, data protection or user privacy, efforts should be made to identify which information is essential to achieve the research goal. Moreover, it is important to note how these challenges have shaped the current technology. For example, data extracted from websites has helped the development of automated text analysis (Berger et al., 2020) as well as the analysis of multimedia content (Liu et al., 2020).

Balducci & Marinova (2018) define unstructured data (UD) as information that lacks a predefined data model or organised structure such as images, videos, voice recordings or text. They also discuss how UD has increasingly impacted various industries and business practices as well as its relevance for marketing. As they mention, this type of data can identify underlying patterns and trends that would not be apparent from structured data. However, despite its potential, many firms are not fully and effectively utilizing it. It is highlighted the importance of marketing researchers to take advantage of computational advancements in machine learning (ML) and artificial intelligence (AI) to analyse and extract insights from UD.

Berger et al. (2020) also highlight the power of converting raw text data into actionable marketing insights. They make an important contribution by underscoring the potential of text analysis to help unite different fields of marketing such as consumer behaviour, advertising, branding, or marketing research. Additionally, it can also help unite

qualitative and quantitative research such as identifying themes in the data and calculating the intensity of them. It provides tools and approaches, both qualitative and quantitative, that can be applied across all these subfields, and this allows marketers to produce integrated and relevant insights into the marketing field.

Finally, Albalawi et al. (2020) focus on studying different methods to perform topic modeling, which allows to extraction of valuable information from the text-based analysis that can be later used to inform marketing decisions. They conclude that Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) generate the most valuable results for short-text analysis.

3.2. User-Generated Content (UGC) potential

Numerous studies have extensively analysed the potential of user-generated content (UGC) as a form of new word-of-mouth recommendations. Zhu & Zhang (2010) found that many users access online reviews before making offline purchases. Their findings suggest that reviews have a greater impact on less popular games whose players have extensive internet experience. Therefore, companies might prefer to allocate efforts to motivate internet-experienced users to leave reviews specifically on niche products. This highlights the influential role of online reviews in shaping consumer behaviour, as potential buyers first look for information on the internet (Chevalier & Mayzlin, 2006).

Similarly, Netzer et al. (2012) prove that UGC offers reliable and representative insights about the market structure by comparing the insights generated by traditional sales and survey-based data with those from UGC. Additionally, Kostyra et al. (2016) use a choice-based conjoint experiment to investigate the impact of Online Customer Reviews (OCRs) on consumer decision-making. Their findings indicate that positive reviews have a stronger impact on brand and product features, while negative reviews have a stronger impact on price perceptions. Consequently, they accentuate the relevance for companies to gain a comprehensive understanding of the underlying motivations behind positive or negative consumer reviews, which is the purpose of my study.

In general, some studies focus on review ratings to drive insights for marketing (Chen & Lee, 2023; Sridhar & Srinivasan, 2012) while others focus more on the content of reviews (Archak et al., 2011; Lee & Bradlow, 2011; Tirunillai & Tellis, 2012). Other

authors, contrary, decide to acquire knowledge from social media (Ghose et al., 2012; Mishra, 2022; Rust et al., 2021). Nevertheless, all approaches provide information that will not be gathered without applying text analysis.

3.3. Marketing theories about consumer behaviour and satisfaction

Within the field of marketing, various theories have helped to understand consumer attitudes and behaviours, especially, those related to customer satisfaction and sustainability. In this section, relevant marketing theories such as the Theory of Planned Behaviour (TPB) and Value-Belief-Norm (VBN) are going to be explored since they provide treasured insights into customer decision-making processes as well as help explain the adoption rates of sustainable products.

First, the Theory of Planned Behaviour, first documented by Ajzen (1985), suggests that attitudes, subjective norms, and perceived behavioural control are key determinants of behaviour. Similarly, the Value-Belief-Norm theory (Stern et al., 1999) proposes that people's values, beliefs and norms guide their behaviour. In the context of sustainable cleaning products, the TPB is going to serve as the basis to understand those factors influencing an individual's decision-making when accepting or rejecting these products. Some factors might include social pressure from family and friends or how easy or difficult it is to incorporate these products into their lives. The VBN theory helps to understand the impact of people's values and beliefs about sustainability, environmental protection, and health on their behaviour towards these products.

Fielding et al. (2008) propose that social identity, which refers to an individual's sense of belonging to a group, integrated with the TPB can influence the engagement and adoption of sustainable agricultural practices, which impacted the water quality.

On the other side, green marketing theory has also been instrumental in encouraging sustainable consumption, specifically, by targeting environmentally conscious customers (Peattie & Charter, 2002) with eco-labelling, environmental communication, and green product design (Laheri et al., 2014). However, the adoption rates of sustainable products remain low (Carrington et al., 2010; Long, 2018). The concept and evolution of green marketing and its potential to motivate the adoption of sustainable alternatives is also explored by Laheri et al. (2014), emphasizing its importance in today's environmentally conscious world.

In addition to these theories, Ellis & Brown (2017) are pioneers of what is known as growth marketing. The main objective is to drive rapid and sustainable business growth through innovative and data-driven strategies. By doing this, barriers to adoption, consumer behaviour and preferences can be identified. Moreover, it allows the development of targeted and informed and effective marketing interventions to increase the adoption rates of sustainable cleaning products. A marketing strategy should involve utilizing data-driven techniques, experimentation, continuous testing and optimization and innovative strategies to retain and acquire customers.

Furthermore, the expectation disconfirmation theory posits that only when the service or product exceeds the customer's pre-purchase expectations, customer satisfaction is achieved (Oliver, 1977, 1980, 2014). This theory provides valuable insights into understanding consumer satisfaction with sustainable cleaning products.

Existing literature has already demonstrated the importance of product quality to determine customer preferences and satisfaction, market share, brand performance and long-term brand success (Jacobson & Aaker, 1987; Tellis & Johnson, 2007; Tirunillai & Tellis, 2012). Therefore, it is important to assess the quality dimensions of sustainable cleaning products to identify areas of improvement and improve customer satisfaction. Ghose et al. (2012) and Radojevic et al. (2017) also allocated their resources to improve customer satisfaction in the hotel sector. Ghose et al. (2012) designed ranking systems for hotels on travel search engines and Radojevic et al. (2017) used a multilevel modeling framework to study factors that influence customers' reviews of hotels.

Finally, relationship marketing theory (Gordon et al., 1998) can also be employed to build long-lasting relationships with customers who have adopted sustainable products. By fostering trust and effective communication, companies can encourage loyalty as customers will be more satisfied and willing to continue using those products. However, it is fundamental to recognise customer needs and preferences when developing effective relationship marketing strategies.

In summary, customer satisfaction is crucial for business since it directly influences customer loyalty, positive word-of-mouth, and long-term success.

3.4. Consumer behaviour and their response to sustainability interventions

As described by Kurz et al. (2015), habits are behaviours that become automatic due to repeated exposure to a context and they can potentially obstruct the goal of achieving a sustainable behaviour change. Gonzalez-Arcos et al. (2021) try to understand which factors generate customers' resistance to sustainability interventions. Among their recommendations they mention the relevance of communicating from a shared and societal coalition perspective, that is, not placing only the responsibility on customers. Additionally, they highlight the importance of actively monitoring consumer reactions to these interventions including social media sentiment.

Additionally, factors influencing consumers to be more sustainable have been also studied (White et al., 2019) such as social influences, habit formation, tangibility or individual characteristics. Contrarily, motives for why people do not adopt sustainable products such as lack of awareness, high costs, lack of trust and lack of access to them have also been mentioned (Wijekoon & Sabri, 2021).

3.5. Brand reputation, product recognition and consumer satisfaction

As stated by (Aaker, 1996) brands are developed to create a unique identity, build customer loyalty and differentiate products from competitors. In general, unstructured text analytics can be a powerful tool for marketing since it can provide valuable insights for example for product development and brand reputation understanding (Markham et al., 2015; Krawczyk & Xiang, 2016). First, to meet customer needs effectively, product managers need to identify what preferences lead to product adoption, which can be done by analysing for example user-generated content (UGC). This will help companies seeking new markets or customers, however, unstructured text analytics can also help companies gain competitive advantage by identifying new suppliers, mapping market reactions and discovering new trends in the industry (Markham et al., 2015). Second, Krawczyk & Xiang (2016) demonstrated that UGC can also provide an understanding of the hotel industry market structure by creating perceptual maps from the most frequent words people used on reviews of a travel agency. By understanding brand reputation, companies can identify areas where their image can be improved or develop marketing strategies based on their target market. In summary, companies can identify weak points and areas of improvement as well as their strengths concerning

competitors, which is crucial to make effective and strategic marketing decisions (Krawczyk & Xiang, 2016). Therefore, understanding a brand's reputation is crucial in developing marketing campaigns and maintaining a positive brand image.

For sustainable companies, product and brand recognition are especially important to differentiate their products from conventional ones and create a competitive advantage. This can attract environmentally conscious consumers who might be willing to pay a slightly higher price for sustainable alternatives. Additionally, following previous theories, by consistently delivering high-quality sustainable products, these companies can build trust and loyalty as well as a positive reputation and brand image that resonates with individuals.

This is also supported by Mazurek (2019), who found that a positive brand reputation can create trust, loyalty, and positive attitudes among consumers, leading to higher sales and revenue. Conversely, a negative brand reputation can lead to a sales decrease and a loss of consumer trust. She put on the spotlight the potential of managing and monitoring companies' online presence as well as building and maintaining a positive brand reputation to stay competitive. Olsen et al. (2014) found that an effective way to impact brand attitude is green new product introductions, which also depends on the number of green messages, the type of product and the source credibility.

In summary, product recognition and brand reputation are crucial marketing theories that can be applied to sustainable products and companies and big data is a powerful tool to understand brand perceptions and customers' preferences (Tirunillai & Tellis, 2014).

4. Data

In this study, data will be web scrapped from Amazon reviews of sustainable cleaning products and non-sustainable cleaning products using R programming language. Amazon was selected as the ideal platform for analysing consumer behaviour for being one of the largest online retailers containing a vast collection of products. Additionally, Amazon provides a special classification of Climate Pledge Friendly products.

The data was extracted for different cleaning products such as multi-purpose cleaners, dish soap, laundry or dishwasher detergent or cleaning wipes. Additionally, different

brands products from small and big businesses, products at different price points and different formats such as individual products or packs were considered to ensure that the analysis is not skewed. Moreover, to select a control group of non-sustainable cleaning products, the same approach was considered.

The web scraped data contains, for each product, the variable rating indicating the number of stars each customer gave to the product, the title of the review, the content of the review, the location, and the date of the review. It is important to mention that, given that web scraping Amazon reviews is subject to legal and ethical considerations regarding data privacy, no personal information or sensitive consumer info such as usernames, user IDs, email addresses or profile photos were collected. After that, all the reviews for sustainable products were combined into a single data set named "*sustainable*" and all the reviews for non-sustainable products were combined into the "*sustainable*" data set.

For both data sets, data cleaning was performed to ensure their reliability and appropriateness for subsequent analysis. By identifying discrepancies in the data sets, the overall data quality and integrity of analysis can be enhanced. Next, the steps taken to pre-process the data are going to be discussed.

Duplicated entries were identified and removed to avoid redundancy and ensure data integrity. Missing values, reviews without review content or rating, were removed to retain only those reviews containing all the information. They amount to 1843 entries in "*sustainable*" and 107 in "*NONsustainable*". Moreover, variables were converted to the appropriate format. The rating was presented as "*4.0 out of 5 stars*", therefore, everything but the rating was deleted resulting in variable rating being converted to numerical format (e.g.: *5, 4, 3, 2 or 1*). Date and location were the same variables (e.g.: *Reviewed in the United States ᴜs on May 27, 2023*), therefore, it was split into two variables and converted to date and character format, respectively.

Additionally, the language of the reviews was analysed, and only English reviews were considered. Most of the selected reviews were written in the United States (95.6% in "*sustainable*", 98.9% in "*NONsustainable*"), however, English reviews were written in other countries as well, such as Canada, Germany, Spain, Mexico, United Kingdom, Australia, Japan, Italy, The Netherlands, Singapore, and France and they were also considered since they can also provide relevant information for the proposed analysis.

For this reason, the country was not selected as an indicator for the language and a package to detect the language was utilised. Country data was also pre-processed to unify the nomenclature since many of them were present as a single country name (e.g.: "*United States")* but also accompanied by their country code (e.g.: "*United States us*").

The resultant data sets consist of 19916 reviews for "*sustainable*" and 20725 reviews for "*NONsustainable*", where each review is represented in one row. To provide an idea of how this data looks like, an extract of five rows in the data for sustainable cleaning products can be seen in Table 1 in the Appendix. Then, further cleaning steps using Natural Language Processing (NLP) are explained in the next section.

Finally, to predict customer satisfaction, the variable rating is converted to a binary outcome, where the possible outcomes are "happy" for a rating of 4 or 5 stars or "not happy" for a rating equal to or lower than 3 stars. Based on this context, a rating of 3 stars is in the "not happy" group since this might indicate that customers' expectations were not completely met, and that the product might be classified as average or mediocre. Therefore, customers might not be extremely dissatisfied, however, they might be less likely to change habits and become sustainable loyal customers, failing to achieve the purpose of higher adoption rates. Moreover, Amazon classifies reviews of 1, 2 or 3 stars as "critical" and reviews of 4 and 5 stars as "positive". Additionally, since the distribution of rating is positively skewed, this helps lighten the class imbalance for the binary prediction models. Then, to find which factors lead to the low adoption rates of sustainable cleaning products, 145 variables are extracted from the data to capture the relevant information. These are 20 PCA factors, 20 LDA topics, 10 extracted emotions, 50 most frequent words in the data after cleaning, 14 most frequent bigrams or pairs of words, 30 word embeddings and the number of words in a review. Therefore, these methods are going to be explained next along with the methods used for interpretation.

5. Methodology

This section encompasses additional cleaning steps, an explanation of data visualization and interpretation techniques such as Sentiment Analysis and Multidimensional Scaling and, finally, an explanation of predictive modeling techniques utilised to help answer the research questions.

5.1. Further cleaning the data with NLP

To perform the subsequent analysis, Natural Language Processing (NLP) techniques will be applied to further clean the data. First, both data sets are duplicated to clean the data differently for Sentiment Analysis than for the other methods.

In general, data was transformed to lowercase and to reduce noise, non-alphabetic signs, excess punctuation, accented characters, numbers, and extra spaces were removed using regular expressions. Regular expressions are a powerful tool used to match a pattern or character combinations in strings.

Additionally, this review data contained emojis, which are images, these were removed. However, emoticons such as *:), :-), or :o)* were replaced by the description "*happy smile*" and emoticons such as *:(, :-(, or :o(* were replaced by "*sad smile*". Then, stop words (e.g.: *"the", "of", "and"*) were also removed since they contain little value for analysis.

As the last cleaning step, stemming was performed to remove all the inflexions from words. This technique reduces every word to its root and only keeps the meaningful part of the word. This process was carried out using Porter's algorithm (Porter, 2006), which consists of several phases, each of which differentiates which endings can be eliminated. For example, *"likeable", "like"* and *"likes"* will be transformed into *"like"*, however, *"table"* or *"tables"* will not be transformed into *"t"*. In the same way, *"replacement"* will be converted to *"replac",* but *"cement"* will not change to *"c"*.

As it can be observed, these rules also consider the length of the word to analyse if it is a reasonable approach to simply delete the suffix (e.g.: -able, -ement). This process will increase the accuracy since it reduces the number of unique words.

On the other side, the duplicated data set was cleaned by only replacing emojis with their meaning or description. In addition, stop words were not deleted since they could affect the overall sentiment of a review or sentence, and consequently, are very informative (Pennebaker, 2011). Finally, stemming was not applied since this could lead to loss of information and introduce noise.

Finally, in Table 1 it can be observed that after stemming and without stop words, 9373 unique terms were present in the "*sustainable*" data set and 9480 in "*NONsustainable*".

Those most frequent ones were “clean”, “smell”, “product”, “love” and “cloth” for “*sustainable*” and “clean, “product”, “smell”, “love” and “wipe” for “*NONsustainable*”. The distribution of word frequency was quite skewed, which can be seen in Figure 1.

Table 1

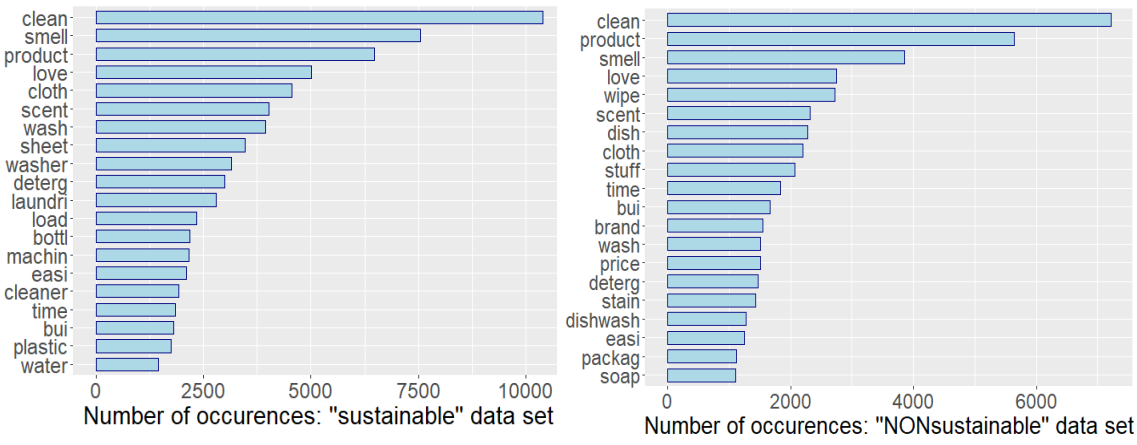
Descriptive statistics for each data set

	sustainable	*NONsustainable*
Average number of words per review	35	28
Average number of words per review after cleaning	8	6
Unique words after stemming	9585	9468
Number of infrequent words (< 1%)	9370	9283
Share of infrequent words	97.8%	98.0%
Unique words after filtering (< 1%)	215	185

Those more frequent terms are kept since they contain valuable information, however, infrequent words are filtered so that terms occurring in less than 1% of the reviews are eliminated due to the little information they provide. Low-frequency terms sum up to 9151 in “*sustainable*” and 9295 in “*NONsustainable*”, which represents 97.6% and 98.0% of the total, respectively. After all the cleaning steps, reviews contained 12 words in “*sustainable*” and 6 terms in “*NONsustainable*” (see Table 1). These data frames are then used for the subsequent analysis.

Figure 1

20 most frequent words for each data set



Finally, tokenization is performed to convert reviews into a machine-readable format. This technique breaks the text into individual tokens, which in this case is a word level. It enables each word to be treated as a separate token or individual unit, which can be used for subsequent analysis.

5.2. Understanding reviews sentiment scores with Sentiment Analysis

The focus now shifts to applying Sentiment Analysis, which is an NLP technique that aims to extract the emotional intent of a writer, that is, provide the sentiment or emotion expressed in a text such as reviews in this study (Kwartler, 2017). The sentiment scores are calculated for each review and each sentence of the data, and they indicate the degree of positivity or negativity of the text. This process is also known as indicating the polarity of a text and it allows a deeper understanding of the reasons behind rating.

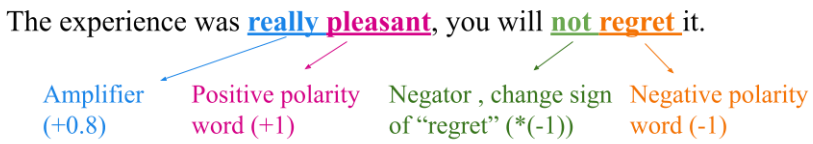
In this case, a subjectivity lexicon-based approach is taken. A subjectivity lexicon is a collection of words or phrases that have a polarity value associated, which indicates if it is positive, negative, or neutral. In this case "hash sentiment huliu" dictionary is used. Additionally, polarity functions not only account for words such as "bad", "hate", "love" or "amazing" but also for amplifiers, de-amplifiers and negators, also called *"valence shifters"*. Amplifiers are those words that increase the intensity of a positive or negative word such as "very" or "really". De-amplifiers are words that decrease this intensity such as "hardly" or "barely" and, finally, negators are terms that reverse the intent of a positive or negative word such as "not". This is a huge improvement over a simple dictionary without valence shifters since in the case of negators, for example, they can completely change the sentiment of the text (e.g.: *"I do (not) like its quality")*. Moreover, this is the reason stop words were kept.

The process of the polarity function starts with scanning positive and negative words within the subjectivity lexicon. Then, the function in our case is going to consider five preceding words and two posterior words to create a cluster of terms. In this cluster, neutral terms count as zero, positive words count as one and negative words count as negative ones. Then, the other words are considered valence shifters. Amplifiers sum 0.8 to the function, while de-amplifiers rest 0.8 and then, negators shift the sign of the sign of the polarity word. Finally, all these scores are summed and divided by the square

root of all words in the analysed text (Kwartler, 2017). This can be easily visualised in the next example:

Figure 2

Polarity function process for an example sentence



$$Polarity\ score = \frac{Total\ raw\ polarity}{\sqrt{Total\ number\ of\ words}} = \frac{0.8 + 1 + (-1) * (-1)}{\sqrt{10}} = \frac{2.8}{\sqrt{10}} = 0.88$$

Finally, to find what customers mention in positive reviews concerning negative reviews, the ratio of the word frequencies is calculated. First, word frequencies for positive and negative reviews are computed. Then, the ratio between them is calculated and the most frequent words used by customers are presented.

Sentiment Analysis enables us to uncover underlying emotions and sentiments expressed in cleaning product reviews, which provide insights into the reasons behind ratings. This help identifies areas of improvement to enhance customer satisfaction.

5.3. Visualising the data with Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS) is employed in this study as an initial approach to understand and interpret the data. This is a statistical technique that is used to visualize and understand similarities or dissimilarities between objects, like words in this study. It reduces complex data into a simpler form for easier interpretation. This is done by creating a map of a lower dimension where words that often appear together are closer on the map and those words that are different are farther apart. The potential derives from the fact that it preserves as much as possible the relative distances between words. By implementing this method, patterns and relationships in the data can be visualized, which makes it helpful for interpreting the information more intuitively.

This dimension reduction is achieved by first calculating word co-occurrences in a predefined context (e.g.: within a review, a sentence, or a window of five words…). The word co-occurrence matrix capture similarities which then need to be transformed to a dissimilarity matrix, the input for MDS, through normalisation. Then, this dissimilarity matrix is transformed into a distance matrix, that is, dissimilarity values are converted into distances. Given this distance matrix that represents distances between each pair of objects, the MDS technique will place each object into a lower dimensional space, where distances are conserved as possible.

In this case, non-metric MDS was performed, which means that the dissimilarity measure is based on non-metric or non-Euclidean distances such as co-occurrence measures. Since the main objective was to preserve the rank order of the dissimilarities or distances between data points, this was the best option. This means preserving relative differences such as "point A is more similar to point B than to point C".

For this reason, points that are more similar and, therefore, closer together in the original dissimilarity matrix, will be also closer in the lower-dimensional representation. Contrary, points that are more dissimilar or farther apart in the dissimilarity matrix will be also farther apart in the MDS map. In general, non-metric Multidimensional Scaling can capture the inherent structure and patterns in the dissimilarity data, even when the distances are not strictly adhering to a metric measurement.

Kruskal (1964) performs a monotone regression of the distances on the dissimilarities. The goal of non-metric MDS is to find the best-fitting monotonic relationship that most accurately reflects the original dissimilarities. This means that as the dissimilarity between two objects (words) increases, the distance between their corresponding points in the lower-dimensional map consistently increases as well. Therefore, to measure how well the distances in the lower-dimensional representation match these original dissimilarities, a measure called *stress* is used:

$$\sigma(X) = \sqrt{\frac{\sum_{i<j} w_{ij}(\hat{d}_{ij}^2 - d_{ij}^2(X))^2}{\sum_{i<j} w_{ij} d_{ij}^2(X)}} \quad \text{subject to } \sum_{i<j} w_{ij}\hat{d}_{ij}^2 = \frac{n(n-1)}{2} \tag{5.1}$$

In this function, X are the n points in the lower-dimensional space. Then, d_{ij} denotes the observed distance between the two objects (word i and word j) in the high-dimensional space and it is obtained from the dissimilarity matrix. Next, $\hat{d}_{ij}$ denotes the

fitted values based on the *configuration* X, that is, the predicted distances between all pairs of words (i and j) in the lower-dimensional space that are obtained from MDS results. Then w_{ij} represents the weights assigned to each pair of dissimilarities, which are optional. These are useful when some dissimilarities are missing, where the weight becomes 0. That allows the function to not consider that dissimilarity in the analysis. If it is not missing, w_{ij} is equal to 1 and therefore, it is considered. Then, the constraint ensures that the sum of the weighted squared fitted dissimilarities across all unique pairs of objects is equal to a constant value. The constant value represents the number of unique pairs of words in a dataset with n words.

The *stress* function measures the discrepancy between the observed dissimilarities in the high-dimensional space and the predicted dissimilarities in the lower-dimensional space (configuration). A lower stress value indicates a better fit since the configuration preserves the distances between objects or words more accurately.

Stress is a quantitative way of evaluating any configuration based on the minimization of the residual variance or residual sum of squares of the monotone regression. Additionally, MDS software performs a standardization step to avoid the stress value depending on the absolute magnitude of the dissimilarities. Finally, as aforementioned, the desired configuration is the one that minimizes *stress* since it is the configuration that is best fitting the data.

To achieve this minimization, we start with an arbitrary configuration and through an iterative procedure called SMACOF (Scaling by Majorizing a Complex Function). This method is going to update and gradually improve the configuration in the reduced-dimensional space until reaching the point of no improvement (Rabinowitz, 1975). Although this method is not using gradient descent, it shares some similarities with gradient-based optimization techniques.

5.4. Predicting customer satisfaction

Star ratings are a useful metric for overall satisfaction; however, they do not provide information on specific themes, topics, or aspects of the product that customers mention, which difficult the understanding customers' preferences or motives not to adopt sustainable cleaning products. Consequently, as aforementioned, to find which

factors lead to the low adoption rates of sustainable cleaning products, 20 PCA factors, 20 LDA topics, 10 extracted emotions, 50 most frequent words in the data after cleaning, 14 most frequent bi-grams or pairs of words, 30 word embeddings and the number of words in a review are going to be used as predictors. Therefore, these methods are going to be explained next.

5.4.1. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a mathematical technique that is quite useful for identifying relevant patterns in high-dimensional datasets and reducing their dimensionality while preserving as much information as possible. One of its main purposes is to identify features that contribute the most to the variability of the data.

PCA achieves this by finding possible groupings of similar variables, that is, by finding new variables, called Principal Components (PCs), which are linear combinations of the original variables. These new variables (PCs) are independent of one another and are ranked by how well they capture the variance in the data. Factors or Principal Components obtained from PCA are used in this study as variables to predict customer satisfaction.

To compute PCA, review data is transformed into a Document-Term matrix (DTM), where each row represents a document (review) and each column represents a unique term from the entire text. The cells of this matrix contain the frequency or count of occurrence of words in the text after cleaning. Then, PCA internally computes the covariance matrix of the data, which represents the relationships and variances between the different variables.

Then, to obtain the Principal Components and their corresponding eigenvalues, PCA performs Singular Value Decomposition (SVD) on the covariance matrix. In general, when a vector in space is transformed, it can change its direction and length. Nevertheless, there are special vectors called eigenvectors that, although being stretched or compressed, remain in the same direction after the transformation. Eigenvectors represent the “directions” along which the transformation takes place and the amount of stretching or compression of these eigenvectors is determined by their eigenvalues. In

PCA, as aforementioned, the purpose is to uncover patterns of variation in a data set with multiple features. The idea is that the first PC corresponds to the direction of maximum variance in the data, then, the second PC is orthogonal to the first one and provides the second largest source of variation and so on. Eigenvalues indicate the importance of each PC. Larger eigenvalues indicate a larger portion of the variance of the data that is captured (James et al., 2013).

Singular Value Decomposition is a mathematical technique that breaks down the covariance matrix into three components. First, the left singular vectors, which are the eigenvectors (PCs) of the covariance matrix and represent the directions where the data shows higher variation. Second, the singular values, which are related to the eigenvalues of the covariance matrix and quantify the amount of variation explained by the eigenvectors. The larger the singular values, the higher variation captured by that specific PC. Third, the right singular vectors, which represent another set of vectors that are orthogonal to the left singular vectors, describe how the PCs are combined to reconstruct the original data, that is, they provide the transformation matrix that maps the PCs (the reduced dimensional space) back to the original feature space. Each right singular vector corresponds to a specific PC and its elements represent the weight of that PC on each original variable, which helps describe the relationship between the PCs and the original variables (James et al., 2013). Finally, eigenvectors or singular vectors from SVD are ordered based on their eigenvalues from highest to lowest. Therefore, the one with the highest eigenvalue is going to be the first PC and so on since they are ordered based on the variance they explain.

Moreover, each PC is denoted by a set of loading vectors or coefficients, which indicate how and how much each variable contributes to the specific PC. The length and direction of the loading vectors describe the strength and direction of influence (positive or negative) that a variable has on a specific PC. Therefore, the longer the loading vector, the stronger the influence (positive or negative). Analysing the loading vectors for each PC helps in understanding which variables (in our case, words) contribute the most to each PC.

Additionally, if the data points (reviews) are projected onto the direction of the loading vectors, we obtain the Principal Component scores (James et al., 2013). This informs

about how each review (data point) aligns with the patterns represented by the PCs since higher scores indicate stronger alignment with the specific PC.

Finally, when the data is measured in different scales, it should be scaled when performing PCA. Nevertheless, to help interpretation of the resulting factors, rotation was performed. It might be the case that the resulting vectors are not aligned with the underlying structure of the data. In that case, rotation achieves a more concise representation of the data, while keeping the variance explained by the PCs. In this case, varimax rotation is implemented, which aims to create more perpendicular loading vectors by rotating and adjusting the angles and magnitude of the loading vectors. By doing this, a clearer separation of the features across the PCs and more distinct and concentrated loading vectors is achieved, which leads to simpler and more interpretable PCA factors. This is the reason scaling is not performed before but after rotation, to ensure that the resulting PC scores are on the same scale as the rotated loadings and maintain consistency.

To sum up, PCA can abbreviate a set of variables into less uncorrelated components, while maximizing the variability of the original data. The resultant PCs are considered collinearity-free and this helps reduce overfitting (James et al., 2013).

5.4.2. Latent Dirichlet Allocation (LDA) for topic modeling

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that is used for topic modeling since it can identify themes or topics within a collection of M documents, which in this study are Amazon reviews. Topics obtained from LDA are used as well as variables to predict customers' satisfaction.

The main idea in LDA is that a document is a mixture of topics and that a topic is a mixture of words. However, this method is considered a soft-clustering method since a term can belong to multiple topics and a document can be about multiple topics (Blei et al., 2003).

Therefore, a probability distribution over words for each topic is given first (β_k). These probabilities (δ_k) indicate the likelihood of a word belonging to a specific topic k.

$$\beta_k \sim Dirichlet(\delta_1 \dots, \delta_k) \tag{5.2}$$

Additionally, LDA observes the word frequency distribution among all documents to define the number of k topics and then, every document (n) is given a probability for each topic (θ_n).

$$\theta_n \sim Dirichlet(\alpha_1 \dots, \alpha_k) \quad (5.3)$$

Both, the word distribution for topic k (β_k) and the topic proportions for document n (θ_n) follow a Dirichlet distribution, which is defined over a vector of positive parameters. Then, β_k and θ_n are defined by δ and α parameters, respectively, which influence the spread of the distributions (Formulas 5.2 and 5.3). Higher values of these parameters lead to a more uniform distribution, while lower values of them lead to sparser distributions. Dirichlet distribution is employed to model prior distributions, which means that it incorporates prior knowledge or assumptions about the expected topic proportions and word distributions. During the training process to find the posterior distributions of topic proportions θ_n and word distributions β_k, these prior distributions work as a regularization mechanism that influences and helps shape the final estimations, also called conditional distribution. This is considered a flexible method since these beliefs or expectations about the distribution of topics in a document and the distribution of words within each topic guide the posterior estimations and then they are later combined with the word occurrences in the document collection (reviews).

Once β_k and θ_n are generated, topic assignments for every word (i) in the documents (n) must be determined (z_{in}). This is done using a multinomial distribution with topic probabilities provided by θ_n. This means that each word (i) in the documents (n) is assigned a topic (z_{in}), which indicates the topic to which the word belongs (see Formula 5.4).

$$z_{in} \sim Multinomial(\theta_n) \quad (5.4)$$

Additionally, words are also assigned to topics (w_{in}), which is also done using a multinomial distribution. Probabilities, in this case, are given by $\beta_{z_{in}}$, which refers to the distribution of words within the assigned topic (z_{in}). In this case, multinomial distribution assigns a probability to each word in the vocabulary that indicates the likelihood of that word being generated from the topic (z_{in}) (see Formula 5.5).

$$w_{in} \sim Multinomial(\beta_{z_{in}}) \quad (5.5)$$

Once identified these parameters, the joint distribution of the topic mixture (θ), the topic assignments (z) and the observed words (w) can be explained. The joint distribution is defined by the following equation (Formula 5.6), and it represents the likelihood of observing a specific combination of these variables, given α and δ (Blei et al., 2003):

$$p(\theta, z, w|\alpha, \delta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \delta), \quad (5.6)$$

where N is the total of words, $p(\theta|\alpha)$ is the prior distribution over topic proportions in a document, $p(z_n|\theta)$ is the probability of assigning a topic to a particular word in a document and $p(w_n|z_n, \delta)$ is the probability of generating a specific word (w_n), given the topic assignment (z_n) and the word distribution (δ).

Overall, different techniques can be used to approximate the joint distribution, however, Gibbs sampling was the selected method since it is particularly useful when handling complex models with various parameters and variable dependencies. This is particularly important in LDA, which is a method that comprises multiple latent variables and their interactions. Additionally, it is more computationally efficient than other techniques (Kwartler, 2017).

Gibbs sampling is a Markov Chain Monte Carlo (MCMC) method that can be used to sample from complex probability distributions. Gibbs sampling starts with a random allocation of words to one of the k topics and, through an iterative procedure, it updates and gradually improves the topic assignments and the word distributions until the topics converge (no significant changes) or until a certain number of iterations (Kwartler, 2017). The idea is that if an iterative sample from the conditional distributions of the variables is performed while keeping the other variables fixed, at some point, the sampled values will come from the true joint distribution of all the variables. In this process, MCMC will ensure that all these samples are correlated with each other in a Markovian approach, that is, each sample will depend only on the previous one. By doing this, LDA can efficiently identify those hidden topics that best explain the observed word occurrences in the reviews.

For the last step, the number of topics (k) and α parameter need to be chosen. The number of topics is going to be selected based on perplexity and coherence measures since there is a trade-off between the model complexity and the model fit. Higher values

of k would lead to better fitting the data, however, the model can become too complex and computationally ineffective. Additionally, there is also a risk of overfitting the data and having a poor generalization with unseen data. On the other side, if the model is too simplistic (low k) underfitting happens, which means that the model cannot capture the patterns and main themes in the data (Kwartler, 2017). Therefore, a balance should be found to achieve good performance while remaining interpretable.

The first method used to achieve this purpose is called perplexity (see Formula 5.7), which measures how well the trained LDA model predicts out-of-sample data. This is done by measuring how surprised is the method on average when trying to predict each new word based on what the model learns during training. As defined by Blei et al., (2003), perplexity is monotonically decreasing in the likelihood of test data, therefore, lower perplexity scores indicate that the model is fitting the data well and, therefore, that it can predict the next word with low surprise (low perplexity).

$$perplexity\ (D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (5.7)$$

In the previous formula, D represents a collection of M documents, also called *corpus*, for a test set, N represents the total number of words in the document d and $p(w_d)$ indicates the estimated probability of observing the word w in document d. Therefore, this formula computes the average log-likelihood per word in the test dataset and then transforms it to perplexity by exponentiating the negative value of that average.

The second method used to define the number of topics is coherence from the textmineR package (W. Jones, 2021) which measures how well the most important words per topic correlate with each other. In other words, it measures the interpretability and coherence of the topics, as the name suggests. In this case, coherence is going to be measured by the average probability of word co-occurrence within a moving window. The intuition behind the coherence measure using in this study is the following: for each pair of words $\{a, b\}$ in the top N words from a topic, where $\{a\}$ is more likely than $\{b\}$ in that specific topic, probabilistic coherence will calculate $p(b|a) - p(b)$. Here, $p(b|a)$ measures how likely is $\{b\}$ in those documents containing $\{a\}$, and $p(b)$ measures how likely is $\{b\}$ in the corpus. The idea is that this should be positive since knowing that $\{a\}$ is in the document would make $\{b\}$ also more probable. In general, we would like to pick the number of topics when the coherence reaches its maximum since that would indicate higher semantic similarity.

Then, the α parameter determines the document-topic distribution, which then controls the topic prevalence. Topic prevalence captures the importance of topics in the revies which refers to the average weight each topic has across documents. Consequently, the higher α for a topic, the more likely that topic is to be observed in the data, thus, the higher topic prevalence. Contrary, if α is small, the resultant topic distribution will be very sparse.

5.4.3. GloVe word embeddings

Word embeddings are vector representations of text data that can be used for machine learning algorithms and that capture semantic relationships and the context of each word. By transforming words to their numerical representations, the machine learning methods applied can capture more subtle information and context from the reviews. In this study, 30 word embeddings are going to be used as independent variables for the predictive models.

GloVe (Global Vectors for Word Representation) is an unsupervised learning method that is used in this study to generate word embeddings. This method is able to predict how often a word co-occurs in the context of another word, therefore, it captures the meaning of words based on the semantic relationships between words derived from a co-occurrence matrix. A co-occurrence matrix X_{ik} is a squared matrix where each row and each column represent a unique term in the vocabulary. It represents word co-occurrences in a text by indicating how many times two words (i and k) appear together in a text within a defined window of words. In this case a window of 8 words before and 8 words after is selected. Therefore $P_{ik} = P(k|i) = \frac{X_{ik}}{X_i}$ is the probability of term k being in the context of the focal word i, where X_{ik} represents the co-occurrence count of word k with word i, and X_i represents the total co-occurrence count of word i with all other words in the corpus. In Formula 5.8 the objective function of GloVe can be found, which has to be minimised using a weighted least squares approach (Pennington et al., 2014):

$$J = \sum_{i,k=1}^{V} f(X_{ik})(w_i^T \tilde{w}_k + b_i + \tilde{b}_k - log(X_{ik}))^2 \qquad (5.8)$$

In this function, V represents the size of the vocabulary. Then, $w_i^T \widetilde{w}_k$ represents the product of the two-word vectors for i and k. This product reflects the similarity or the extent to which the meanings of both words align in the vector space. For this reason, if this product is high, that means that these embeddings point in the same direction in the vector space and there is a strong fit and a high co-occurrence probability. If the product $w_i^T \widetilde{w}_k$ is low, these words point in opposite directions and there is a low co-occurrence probability. Next, $b_i + \tilde{b}_k$ represents a bias for each word to account for the effect of selecting i or k as the focal word. Then, $log(X_{ik})$ represents the logarithm of the co-occurrence probability. The weighting function f is used to adjust the influence of the co-occurrence counts in the model. It is defined by $f(x) = \begin{cases} \left(\frac{x}{x_{max}}\right)^{\alpha} & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases}$.

This function gives small weights to the very low co-occurrence frequencies, which is defined when $x < x_{max}$ but also controls the weights to the very large values of X_{ik} by setting it to 1. That is, it stops giving weight. Additionally, α ranges from 0 to 1 and it determines how quickly the weights decrease for low co-occurrence counts. The default value of 0.75 is selected, however, larger values of α lead to a quicker reduction of the impact of infrequent words. Contrary, smaller values result in a slower reduction in the influence of low co-occurrence counts.

GloVe updates word embeddings and weights through an iterative procedure that optimizes the previous objective function. In this case, the Stochastic Gradient Descent (SGD) iterative procedure is used with 100 iterations.

Word embeddings are used also to measure similarities between words, which is done by computing *cosine similarity* for word embeddings (Pennington et al., 2014).

$$Cosine\ similarity = \frac{w_i^T w_k}{\|w_i\| \|w_k\|} \tag{5.9}$$

This measure computes the cosine of the angle between the two embeddings in the high-dimensional space. The intuition behind it is that words with similar meanings have similar locations in the high dimensional space or will point in the same direction. Therefore, if two words have similar positive (or negative) values on the same dimension, they will be more similar and will have a higher cosine similarity. This is a scaled version of the product $w_i^T w_k$ from the objective function (5.8).

5.4.4. Logistic Regression

In this study, a Logistic Regression is utilised as a statistical method to predict customer satisfaction. Logistic regression is a powerful and interpretable technique used for binary classification like in this case, where the outcome is *happy* or *not happy*.

Logistic regression is going to model the probability of Y, the "satisfaction" or "rating" response variable belonging to a particular category (1= happy) given some predictors (X). A Logistic Regression has the same right-hand side as a linear regression; however, the left-hand side of the equation represents the log odds or the logit, which is represented by $\frac{P(Y=1|X)}{1-P(Y=1|X)}$. This quantity takes any value between 0 and ∞. Values close to zero indicate a low probability of "happiness" or "satisfaction", while values close to infinity indicate a very high probability. This leads us to the Logistic Regression function, which is limited between 0 and 1:

$$log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... \ + \beta_j X_j \qquad (5.10)$$

In Formula 5.10, β is the coefficient of the feature X, and all of them represent the impact of that feature on the log-odds. Then, ε is the error term, which accounts for the effects that are not explained by the selected independent variables. Logistic Regression fits an S-shaped curve, called the Sigmoid function, to estimate the probability of the response belonging to a specific category, which can be $Y = 1$. These coefficients are estimated by using *Maximum Likelihood Estimation (MLE).* The intuition behind it is that we are going to try to find $\widehat{\beta}$ such that the final model generates predicted probabilities as close as possible to the observed data (James et al., 2013). Then, all cases with a predicted probability of more than 50% are classified as "happy", while those cases with less than 50% are classified as "not happy".

5.4.5. Penalised regression: Lasso Logistic Regression

Logistic Lasso Regression is an algorithm that combines Logistic Regression previously explained with a shrinkage method called LASSO. It is used with the same purpose of modeling the probability of a binary response (satisfaction vs. no satisfaction) based on

predictors. However, it is employed to enable the introduction of variable interactions as new predictors.

Lasso algorithm shrinks all the coefficients towards 0 by adding a penalty parameter λ to the maximum likelihood that converts unimportant variables to exactly 0 (see Formula 5.11). Then, relevant variables continue to be higher than 0. Additionally, Lasso also removes variables showing collinearity, which can lead to unreliable coefficient estimates. As the penalty parameter λ increases, the coefficients suffer a higher shrinkage towards 0 and more variables are converted to 0, which usually helps interpretation when several variables are included in the model. The traditional regression is equivalent to having $\lambda = 0$, that is, no shrinkage. However, λ needs to be tuned, which is done through cross-validation to achieve the minimum classification error plus one standard deviation. Therefore, a balance between model complexity and prediction accuracy is sought.

$$L(\beta_1, \dots, \beta_p) = \sum_{i=1}^{n} y_i x_{ij} \beta_j - log\left(1 + e^{x_{ij}\beta_j}\right) + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (5.11)$$

Tibshirani (1996) propose that although this shrinkage increases the bias, performing variable selection also decreases the variance of the predicted values, which leads to higher prediction accuracy. Additionally, Lasso can identify the most influential predictors which help interpretation. This is one of the major reasons why Lasso is utilised since interactions between all the predictors with the emotions are considered, which increases a lot the number of variables in the model.

In general, by employing this method, more variables can be included in the model, which can provide interesting insights, without creating overfitting and an interpretation problem.

5.4.6. Random Forest (RF) to predict customer satisfaction.

To predict customer satisfaction based on summary measures of the review text data, the Random Forest classification method is going to be performed. This method is highly powerful, it can handle high-dimensional data and large feature spaces and it can capture complex interactions between variables, which makes it unnecessary to add

interactions as new predictors. Additionally, Random Forest can cope with nonlinearity, which is quite common in text data and strives when talking about overfitting, since it performs random feature selection and bagging, which helps reduce it. Therefore, this model is going to be used as a comparison.

Random Forest is an ensemble method, which means that it tries to improve the accuracy of a model by combining multiple simple models or *weak learners* such as Decision Trees. Decision Trees are tree-like models that make decisions by recursively splitting the feature space based on conditions or questions that best separate the data into the two classes, which is a top-down greedy algorithm. This means that the algorithm starts to split the feature space from the top, trying to find the most informative splits for prediction, however, it is greedy because it makes the best split at each step of the process instead of looking at the split that could lead to the best final tree. Moreover, splitting decisions are guided by impurity-based criteria, which in this case is the *Gini Index*:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{5.12}$$

Gini Index indicates how likely a randomly selected element is to be incorrectly classified. It computes the sum of the squared probabilities of each class within the node and then subtracts it from 1. It ranges from 0 to 1, where 0 means that the nude is pure and all the observations in a node are from the same category (better class separation) and 1 means that all the observations are randomly distributed across categories (maximum impurity). The former is the most optimal scenario since it means that the classification error is going to be minimal (James et al., 2013).

Recursive binary splitting is going to be performed until reaching a stopping criterion and, if this is not specified, the tree is going to grow as much as possible, resulting in a complex tree that may overfit the data. Although this tree might have low bias it will have high variance and pruning techniques should be used.

In general, even though pruning helps reduce the complexity and risk of high variance, Decision Trees still suffer from high variance. This can be mitigated using Random Forest, which introduces two crucial techniques, *Bootstrap Aggregation* and *split-variable randomization*. Bootstrap Aggregation involves training each Decision Tree on

a random sample of the original data created through bootstrapping, which is random sampling with replacement. Random Forest then reduces variance and produces a more stable prediction by training multiple trees on different bootstrapped samples (B) and then aggregating their predictions through a majority vote (for classification purposes). This process can be easily visualised in Figure 1 in Appendix.

Nevertheless, this could lead to tree correlation since, even though the model-building process is independent, trees may have similar structures, especially on first splits. The reason is that some predictors might be more important or stronger than others. A solution is then provided by Random Forest, which additionally performs *split-variable randomization*. That is, instead of considering all p features at each split decision, Random Forest only contemplates a random sample of p features, designated by m.

Generally, two parameters provide a special and crucial improvement in predictive accuracy when being carefully selected. These are the total number of trees and the number of features m. The first should be chosen based on error stabilization and minimization. This leads to one of the major advantages of this method, which is the *Out-Of-Bag (OOB)* sample. The bootstrap method uses approximately two-thirds of the observations during training, which leaves out one-third of the observations, known as *Out-Of-Bag (OOB)* samples. These can be used to estimate the prediction error of the Random Forest since they have not been used during training and, therefore, act as a validation set. This is especially useful when working with large data since enough trees (B) allows the OOB error to be comparable to the error obtained through cross-validation, which can be computationally expensive. Then, the number of features is $m = \sqrt{p}$ for classification, as in this case, or $m = \frac{p}{3}$ for regression. However, it can be tuned by computing the OOB error for different m values and taking the one that gives the lowest error.

Finally, the importance of each variable is computed to compare the results with the other classification methods used. This presents the most important variables based on the Mean Decrease Gini coefficient. It measures the impact of each variable on the model by measuring how much the Gini Index decreases when that variable is included in the splitting decisions of the model. Therefore, the higher the Mean Decrease Gini score, the higher the importance of the variable since it means that including that

variable in the model's splitting decisions leads to a greater reduction in impurity and, consequently, a higher accuracy.

5.4.7. Performance metrics for models' comparison

Logistic Regression, Lasso Logistic Regression and Random Forest performance are compared based on Matthews Correlation Coefficient (MCC), first formulated by Matthews (1975). This is a measure of quality for binary classification models. This method is preferred over the accuracy measure especially when dealing with imbalanced data. It ranges from negative one to positive one. The first (-1) means that the predicted outcome is different to the ground truth and the second (+1) means that predictions are perfect, and everything is correctly classified. Then, MCC equal to zero means that there is a random prediction.

MCC formula can be derived from a confusion matrix, which is a table that informs about the actual classes in the data and the models' predicted classes and from which the number of true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) can be obtained. Then, MCC can be computed as follows:

$$MCC = \frac{TP * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{5.13}$$

Additionally, given the imbalanced data, sensitivity and specificity are also provided. Sensitivity refers to the true positive rate, that is, the fraction of positive cases that have been correctly classified out of all the positives. Specificity refers to the true negative rate, which is the fraction of negative cases that have been correctly identified out of all the negatives. Consequently, the higher the percentage, the better the prediction quality.

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{5.14}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{5.15}$$

6. Results

In this section, the results of the analysis are presented. First, the findings of Sentiment Analysis and Multidimensional Scaling are presented and discussed for both *"sustainable"* and *"NONsustainable"* data sets. This is followed by the models' results of Logistic Regression, Logistic Lasso Regression and Random Forest for the *"sustainable"* data set. Similarly, the *"NONsustainable"* data set model's results are presented, which is followed by a discussion and comparison of the findings between both data sets to identify similarities and differences.

6.1. Visualizations for "*sustainable*" and "*NONsustainable*"

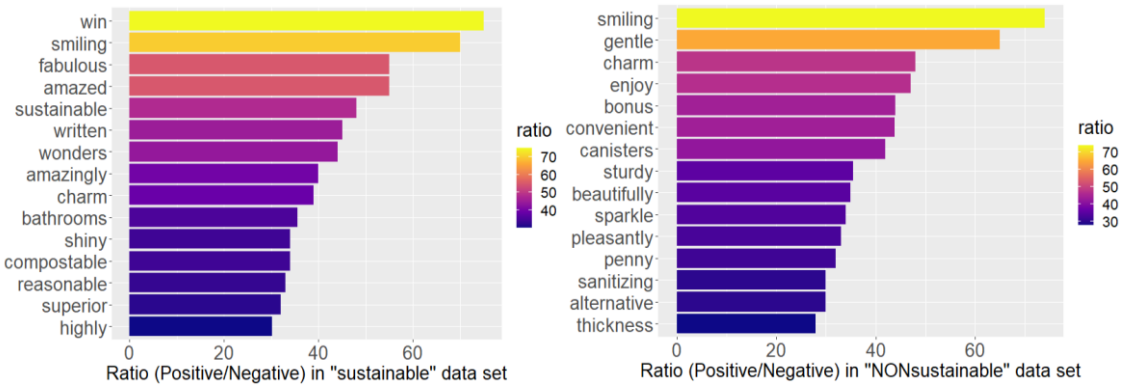
6.1.1. Sentiment Analysis

First, the most frequent words for both data sets are computed for negative and for positive reviews. However, most words were common for both positive and negative reviews. Therefore, the most informative option is to describe relative frequencies to distinguish which words are relatively more common in positive than negative reviews and which ones are relatively more frequent in negative reviews.

In Figures 3 and 4 the ratios are presented. A ratio of 70%, for example, for words "win" and "smiling" means that these words appeared seventy times more frequently in positive reviews than in negative reviews.

Figure 3

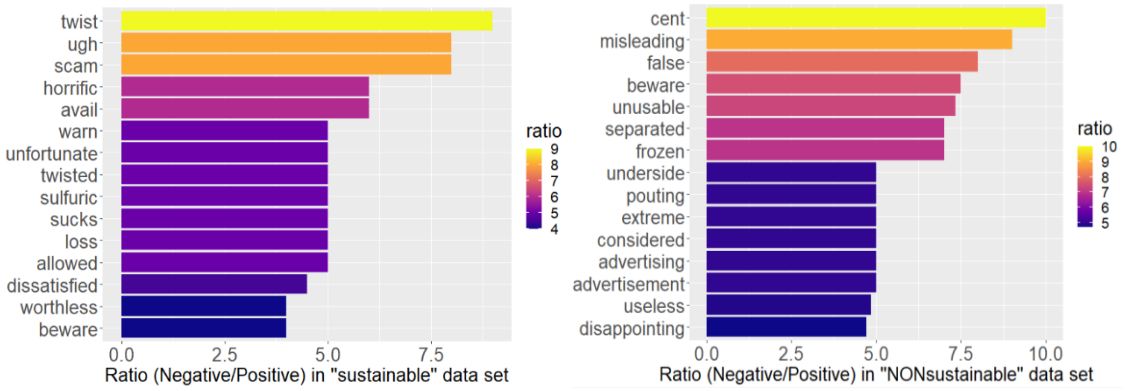
Relative frequencies: fifteen words more frequent in positive reviews



In Figure 3, words' relative frequencies for positive reviews in both data sets are presented. In this case, both present similarities such as "fabulous", "smiling" or "amazed" used in *"sustainable"* and "smiling", "charm" or "gentle" used in *"NONsustainable"*, which suggest positive emotional experiences regardless of sustainability. However, the *"sustainable"* data set contains specific words ("sustainable", "compostable", or "reasonable") that indicate that not only customers are paying attention to the sustainability aspect of the product but also, they express optimism about it.

Figure 4

Relative frequencies: fifteen words more frequent in negative reviews



In Figure 4, words' relative frequencies for negative reviews in both data sets are visualized. The *"sustainable"* data set contains words such as "twist", "scam" or "sulfuric", which are not present in *"NONsustainable"*. This indicates customers' concerns regarding the veracity of the sustainable product since "sulfuric" might suggest disappointment with the presence of chemicals. Similarly, the *"NONsustainable"* graph contains "misleading", "unusable", "frozen", "false" and "advertising". This could indicate that buyers are dissatisfied with the quality, functionality, or usability of the product, but also with false advertising or false claims.

In addition, Sentiment Analysis is performed by applying the polarity function on a review and sentence level for both data sets. To recall, the polarity function represents the average sentiment score given for each piece of text. Moreover, the indicated value in the polarity function is (+/-) 0.8 for valence shifters, however, different weights are explored (0, 1, 2, 5, 10) to observe the impact on sentiment analysis. By performing this

method on a review level, a comparison between sentiment scores and the actual star ratings is made and a scatterplot is created to visualize this relationship.

Additionally, Pearson's correlation and a confusion matrix are used. Pearson's correlation measures the relationship between these two variables (polarity score and rating) by quantifying the strength and direction of the association. It ranges from -1 to 1, where -1 means perfect negative correlation, 1 indicates perfect positive correlation and 0 means no correlation. A positive correlation, therefore, means that as sentiment score increase, the rating also increases, that is, if people express more positive sentiment in the reviews, they give higher ratings. A negative correlation, the contrary, indicates that as sentiment scores increase, the rating decreases. Furthermore, a confusion matrix is a metric that, in this case, summarizes the performance of the sentiment classification. It allows us to compare the sentiment categories created from the polarity function and the original sentiment categories that the customers provide. This technique helps to understand the accuracy of the sentiment analysis classification.

Ratings are categorized as positive (4 and 5 stars), neutral (3 stars) and negative (1 or 2 stars) and the derived sentiments are positive for scores above 0, neutral for scores equal to 0 and negative for scores below 0. Results for the *"sustainable"* data set are interpreted first and complete results can be observed in Table 2.

The confusion matrix for ratings and derived sentiment scores using the default weight of 0.8 gives an accuracy of 76.40%, which means that the sentiment analysis model correctly classified the sentiment as positive, negative, or neutral in 76.40% of cases. Additionally, the correlation is 0.46, which indicates a moderate positive relationship between customers' star ratings and derived sentiment scores. It can be observed in Table 2 how giving too much weight to valence shifters such as "very", "really" or "barely" decreases the accuracy and decreases the correlation. The same happens in *"NONsustainable"*, where the accuracy and correlation using the default value 0.8 are slightly lower, 73.10% and 0.42 respectively. In that data set, given a weight of 1 provides a small increase in accuracy of 0.02%.

This effect can be observed in Figure 2 and Figure 3 in Appendix as well, which present the scatter plots representing the relationship between review ratings and sentiment scores. These plots represent on the x-axis the actual ratings and on the y-axis the

review stars. Each point represents one review, and the results suggest that the higher the valence shifters' weights, the more extreme values are generated or the higher the variance in the distribution of reviews (y-axis). Therefore, for weights set to 10, reviews with many valence shifters will get very high or very low scores, whereas reviews with not many valence shifters will get sentiment scores closer to 0, that is, to a neutral opinion.

Table 2

Sentiment Analysis results for both data sets

	"sustainable"		***"NONsustainable"***	
Weights	Accuracy (%)	P. Correlation	Accuracy (%)	P. Correlation
0	76.07	0.46	72.72	0.43
0.8	**76.40**	**0.46**	**73.10**	**0.42**
1	76.29	0.45	73.12	0.42
2	76.15	0.42	73.08	0.39
5	75.74	0.33	72.83	0.31
10	75.49	0.26	72.66	0.24

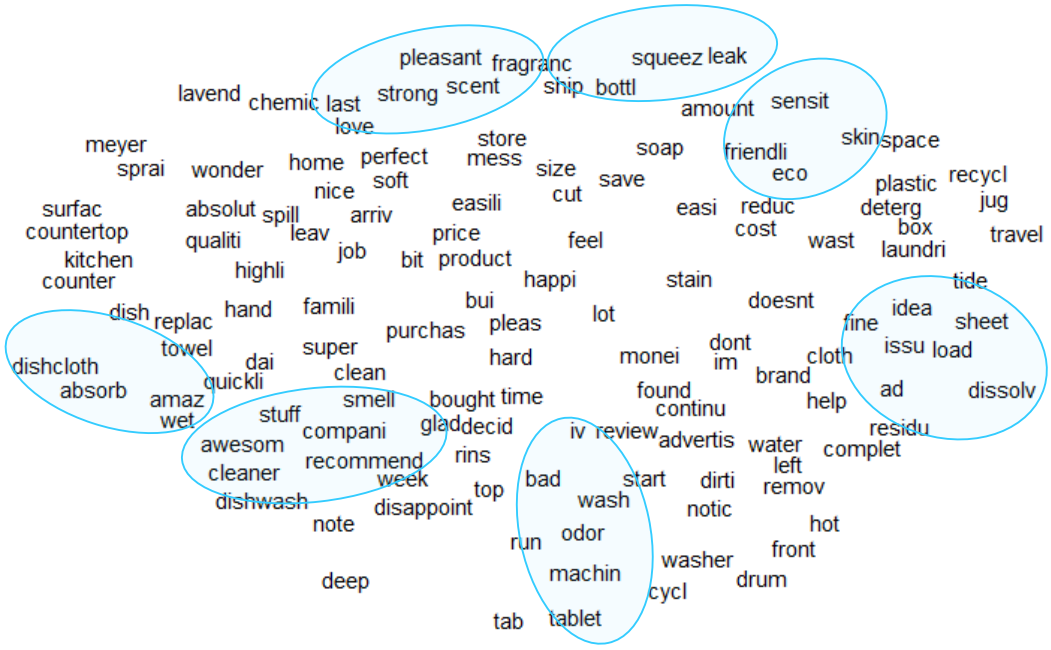
Additionally, sentence-based Sentiment Analysis is also performed and the distribution of the sentiment score is presented in histograms in Figure 4 for *"sustainable"* and in Figure 5 for *"NONsustainable"* in Appendix, which are similar. Like with review-based, it can be seen that with the weight set to 10, the distribution is wider than with the weight set to 0.8. In *"sustainable"* the range is (-21.7, 22.8) and in *"NONsustainable"* the range is (-21, 23.4) for weights set to 10, whereas with the default value (0.8) the range is (-2.60, 3.20) in *"sustainable"* and (-2.68, 2.63) in *"NONsustainable"*.

6.1.2. Multidimensional Scaling

To recall, MDS allows to represent the data in a lower dimensional space. The feature co-occurrence matrix utilized in this case considered words co-occurring in the same review. This approach was selected given the low average number of words per review after stemming and cleaning. In Figures 5 and 6, the resultant MDS maps for both data sets can be observed.

Figure 5

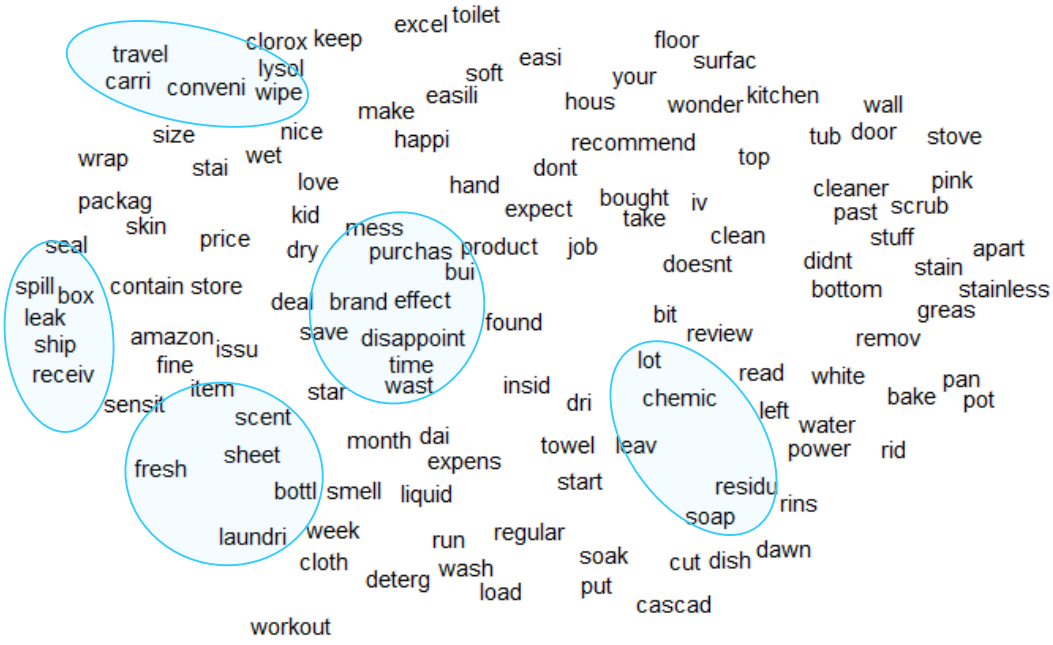
MDS map for "sustainable" data set



Regarding the sustainable options (see Figure 5), one potential positive theme could also be the scent of the product (pleasant, strong, scent, fragrance, last, love). Another common theme is the leakage problem (squeez, leak, ship, bottle). Then, customers mention sensitive skin (sensit, skin, friendly, eco), which could be positive. However, in the case of being negative in the reviews, that would suggest that customers might prefer appropriate products for sensitive skin. People also mention the problem with residues (sheet, load, dissolv, residu, issu). Another problem relates to the bad smell of the machine (odor, machin, wash, tablet, bad). Additionally, customers seem satisfied with sustainable dishcloths because they absorb well (dishcloth, absorb, amaz, wet) and with dishwash cleaner (dishwash, cleaner, recommend, company, awesome).

Figure 6

MDS map for "NONsustainable" data set



Some of the themes found for the non-sustainable products data set (see Figure 6) are travel wipes (travel, carri, conveni, wipe), which denotes satisfaction, or product leakage (spill, leak, box, ship, receiv), which could reflect discontentment with the shipping method. Another theme could be the satisfaction related to the scent of the laundry products (scent, sheet, fresh, laundri, bottle). Finally, there seems to be dissatisfaction with the purchase of one of the non-sustainable brands (brand, purchas, bui, effect, disappoint, time, wast). Finally, people mention chemic and residues about soap (lot, chemic, leav, residu, soap), which suggests a complaint regarding product remains.

Regarding the measure of goodness of fit, *stress* is used, which measures how well the MDS configuration reflects the original dissimilarities. A rule of thumb indicates that stress should be lower than 20% and 0% would indicate that there is a perfect representation of the original dissimilarities in the MDS map (Kruskal, 1964). However, Mair et al. (2016) already exposed some of the problems of this rule. One of them is that, as the number of words considered increases, stress also increases and there is a higher probability of finding discrepancies between the observed dissimilarities and the fitted dissimilarities. This is a major constraint, given that nowadays, the number of

words is quite large. Additionally, the number of dimensions also affects the stress value since a higher number of dimensions allows for a more flexible representation of the data, which will result in a lower stress value and a better fit (Mair et al., 2016). However, the primary goal of MDS in this study is to obtain a visual representation and gain insights into themes discussed by customers. Therefore, two dimensions are selected to be able to interpret the map more accurately.

Different approaches were considered regarding the number of words and 200 words are finally considered. The reason is that the results with fewer words were similar but less informative and 200 words provide a reasonable number of terms for the analysis without overwhelming the map. For this representation, the stress value is 34% and 35% for *"sustainable"* and *"NONsustainable"* respectively, which means that there is some degree of discrepancy between the original dissimilarities and the distances in the MDS map. Nevertheless, although the distances might not be perfectly preserved, the overall patterns and relationships between the words still provide valuable information as an exploratory and suggestive interpretation.

6.2. Response and predictors of the prediction models

To recall, a binary classification is performed, where the possible outcomes are *"happy"* for a rating of 4 or 5 stars or *"not happy"* for a rating equal to or lower than 3 stars. Then, 145 predictors are used to identify those factors leading to low adoption rates, which are the 50 most frequent words in the data after cleaning, 14 most frequent bigrams, 10 extracted emotions, 20 PCA factors, 20 LDA topics, 30 word embeddings and the predictor indicating the total number of words in a review after cleaning. For these predictions the data is partitioned in 70% as train data and 30% as test data.

Unigrams or single words have been selected based on occurrence, therefore, the 50 most common words used in the reviews are utilised as predictors as well. The reason is that they are representative terms that contain valuable information about topics people mention in the reviews. At the same time, capturing important information is balanced with interpretability and computational efficiency.

Bigrams are a set of two adjacent words or tokens in a text. As an example, in the sentence *"The quality is good"* bigrams are *"the quality"*, *"quality is"*, and *"is good"*.

Bigrams or pairs of words are selected after cleaning and stemming, which already removes infrequent and stop words. In this study, 14 bigrams in both data sets are selected as predictors based on the criterion of occurring more than 200 times and more than 130 times in *"sustainable"* and *"NONsustainable"* respectively. The reason is that infrequent bigrams may not provide enough meaningful information and they might introduce noise. Additionally, by selecting the most frequent ones, representativeness can be ensured since they are more likely to reflect patterns in the data and less computational complexity is achieved.

The selected emotions are based on the classification made by psychologist Plutchik (1982). He believed that there are eight emotions, which have been the basis for human and animal survival: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. A description model of Plutchik's wheel of emotions can be seen in Figure 6 in Appendix. Additionally, two extra emotions are included which are positive and negative.

The selection of the number of PCA factors, 20, is made based on the trade-off between complexity and information loss. In this case, using too few factors could result in information loss since the first PCs did not represent a huge variation of the data. However, using too many could lead to overfitting and model complexity. The selection of 20 factors represents a balance between the complexity and interpretability of the model and capturing sufficient information from the original dataset since it is possible to extract meaningful information and interpret the relationship between these factors and customer satisfaction. In *"sustainable"*, the variance explained by the first 20 PCs is 53.48% and for *"NONsustainable"* it is 47.46%.

Regarding the 20 LDA topics, the selection is made based on perplexity and coherence. For both data sets perplexity suggests 10 topics since it is when the measure is lowest and coherence 20 topics, which refers to the point where the coherence is maximized. Therefore, since interpretability and coherence are crucial for this analysis, 20 topics are selected. Moreover, the selection of α in LDA is done manually and set to default, resulting in $\alpha = 0.1$, however, α optimization was enabled. This means that even though the α was set to 0.1, the model will try to optimize it during the fitting process every 10 Gibbs iterations while considering the initial suggestion. The reason to do this is because of the nature of the data. Review data is expected to cover a range of topics related to the selected products and customers are expected to discuss their preferences

and experiences in many ways. Therefore, selecting this smaller value can encourage a more diverse distribution of topics within all the reviews as well as capture this variability. The final topics can be inspected in Figures 7 and 8 in the Appendix.

Finally, 30 word embeddings are used as predictors as well and this is how they are created. First, 10 dimensions are considered to create the word embeddings. Then, the minimum, maximum and mean values for each of the 10 dimensions across all the words present in reviews are calculated. Therefore, the result is 30 columns, which are the mean values of the word embeddings across the 10 dimensions, the minimum values of the word embeddings across the 10 dimensions, and the maximum values of the word embeddings across the 10 dimensions.

6.3. Models' results for "*sustainable*" and "NON*sustainable*"

In this section the different models are compared based on the Matthews Correlation Coefficient (MCC). Sensitivity and specificity are presented as well to inform about the classification quality of both outcomes.

Many models are created for each method. Every method is performed with all the predictors, only with LDA topics, only with PCA factors, only with words and bigrams, only with word embeddings and only with emotions. Additionally, Logistic Regression and Lasso Logistic Regression are performed with all the predictors and interactions with emotions. However, only the best models are selected, which are two Logistic Regressions, one with all predictors and the other with all topics and emotion interactions; two Lasso Logistic Regressions, one with all predictors and the other with all predictors and emotion interactions; and a Random Forest model with all variables. A summary of these models' performance can be found in Table 3.

Both for Logistic Regression and Lasso Logistic Regression a threshold of 0.7 is selected, instead of 0.5. This is done to address the class imbalance problem since class 1 (happy) contains 85.7% (77.64%) of the data and class 2 (not happy) represents 14.3% (22.36%) of the data for *"sustainable"* (*"NONsustainable"*). By doing this, the model classifies more instances as class 2 when the predicted probability is higher than 0.7, which helps capture more true negatives and increase specificity. That is, more instances

from the minority class will be correctly classified as such. The goal is to balance the sensitivity and specificity to achieve good prediction precision in both classes.

Table 3

Summary of the models' performance for both data sets

	"sustainable"			***"NONsustainable"***		
	MCC	Sensitivity	Specificity	MCC	Sensitivity	Specificity
Logistic R.: All predictors	0.129	61.29%	56.63%	0.391	83.61%	57.12%
Logistic R.: All topics + interactions	0.339	92.29%	39.10%	0.321	85.21%	46.55%
Lasso Log. R.: All predictors	0.324	95.30%	30.11%	0.386	86.76%	51.40%
Lasso Log. R.: All predictors + interactions	0.336	95.34%	31.24%	0.388	86.98%	51.25%
Random Forest: All predictors	0.310	87.20%	46.07%	0.428	86.03%	57.56%

6.3.1. Sustainable data set

In this data set, the best model based on the MCC is the Logistic Regression with all topics and interactions with emotions (0.339), with a sensitivity of 92.29% and a specificity of 39.10%. Including all parameters or all parameters with interactions led to overfitting the data. Therefore, since interactions between the emotions and the whole set of possible variables could provide interesting insights, a Lasso Logistic Regression is performed, which yields an MCC of 0.336, a sensitivity percentage of 95.34% and a specificity of 31.24%. Then, the same method is performed without interactions and a slightly lower MCC is achieved (0.324).

Finally, the Random Forest model is created since it can also investigate these close relationships between variables, given the split structure. Based on tuning $m = 12$, which is equal to the rule of thumb $m = \sqrt{p} = 12$. Additionally, the number of trees is 67, which yields the lowest OOB error estimate. Then, since there is a class imbalance,

an equal sample size for each class is specified, which allows for a balanced class distribution in the training data. This helps prevent the model from being biased towards the majority class, which is class 1 as “happy”. The results for this model are an MCC of 0.310, a sensitivity of 87.20% and a specificity of 46.07%.

Based on these results, Lasso Logistic Regression with emotion interactions is the best option since it accounts for collinearity and includes interactions with emotions for every predictor, which is considered to provide additional interesting insights. For this reason, it is selected as the preferred method.

In the selected Lasso model, the optimal value of λ is achieved through 10-fold cross-validation, which is a technique that splits the data into multiple subsets and iteratively fits the model using different values of λ. Then it evaluates the performance on the validation sets. The value of λ that is selected is the one that lies within one standard error of the optimal value, which is 0.0097. Moreover, from the 1542 coefficients, 95 have not become zero.

6.3.2. Non-sustainable data set

The best model in this data set is the Random Forest with a default $m = \sqrt{p} = 12$, equal sample size for both classes and 103 trees, which achieved an MCC of 0.428, a sensitivity of 86.03% and a specificity of 57.56%. In second place, the Logistic Regression with all predictors achieved an MCC of 0.391 and, in third place, the Lasso Logistic Regression with interactions achieved a coefficient of 0.388. Then, Lasso Logistic Regression without interactions (0.386) and Logistic Regression with all topics and all interactions (0.321) were the worst performers.

In this case, Lasso Logistic Regression with all predictors and interactions is selected as well. One of the main reasons is interpretability, the goal of these predictions, and simplicity. A simpler model such as Lasso Logistic Regression is preferred over a more complex one, such as Random Forest because of the clear and straightforward interpretation of relationships between predictors and the target variable. Contrary, the Random Forest model is more complex due to their ensemble nature, and although it may provide higher accuracy, it compromises interpretability, which is fundamental in this study. Lasso Logistic Regression demonstrates that its performance is good enough,

compared to Random Forest, to use it for interpretation. A second reason is that the regularisation in the Logistic Regression allows to get rid of irrelevant variables and reduce collinearity, which helps prevent overfitting and at the same time, it considers all emotion interactions, which can provide valuable information.

The selected Lasso model's λ value is 0.0130 and a total of 92 predictors out of 1542 are non-zero and are considered relevant to the model.

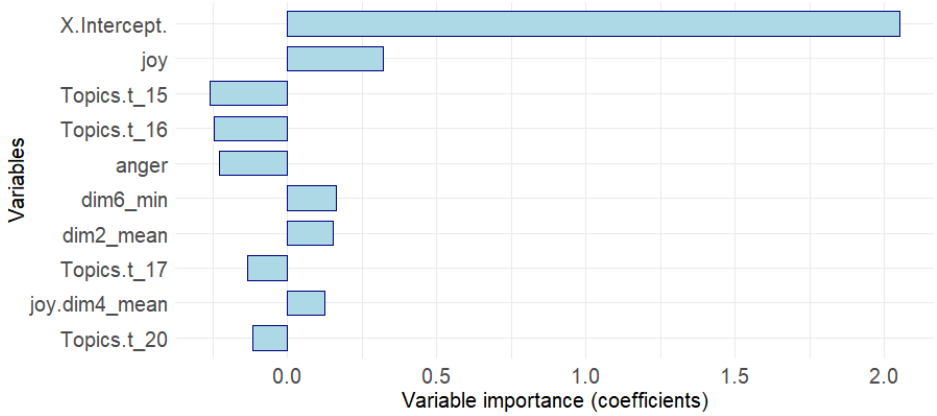
6.3.3. Comparison of both data sets' results

To shed light on which factors lead to the low adoption rates of sustainable cleaning products, the results of the Lasso Logistic Regression with emotion interactions for both data sets are explained in this subsection. It is important to mention that variables have been standardized before training the model to ensure that all variables are on the same scale and have a similar magnitude.

First, the variable importance is presented in Figures 7 and 8 for both data sets. It is measured by the impact of the coefficient on the response in absolute value, and the first 10 predictors are discussed. Note that variables were standardized before performing the Logistic Lasso Regression since predictors are not measured in the same scale.

Figure 7

Top 10 most impactful variables for Lasso Logistic Regression with interactions in "sustainable" data



Regarding the *"sustainable"* data set (see Figure 7), the first variable is the intercept, which is positive and means that when all other variables are held at a value of zero, the

log-odds of a review being “happy” is approximately 2.05. This suggests that even in the absence of other variables, the log-odds of a customer being satisfied or a customer giving a positive review is positive.

Among the highest coefficients in absolute value, emotion “joy” suggests that an increase in the standardized “joy” emotion variable leads to an increase in the log-odds of customer satisfaction, ceteris paribus, which means that, when customers express more joy in their reviews, they are more likely to be satisfied, if everything else remains constant. Then, an increase in the presence of LDA topics 15 or 16 is associated with a decrease in the log-odds of having a positive review, ceteris paribus. Topic 15 is mostly related to product leakage, bottle quality, and disappointment when receiving the package and it also mentions expectations regarding advertisements of that company. This indicates that customers that express these themes or issues are more likely to be dissatisfied. On the other side, topic 16 might seem positive since it mainly talks about smell, lavender, lemon or clean. However, it also mentions strong and chemic, which can lead to a negative experience if the scents are too strong. Moreover, an increase in the standardized emotion variable “anger” is associated with a decrease in the log-odds of having a positive review, keeping everything else constant.

Additionally, an increase in the values of the “dim6_min” or “dim2_mean” word embeddings variable is associated with an increase in the log-odds of customer satisfaction, ceteris paribus. First, “dim6_min” represents the minimum value encountered for word embeddings in the sixth dimension. The coefficient indicates that higher values of “dim6_min” are associated with a higher likelihood of customers being happy, ceteris paribus. The specific semantic aspects represented by this dimension are features or characteristics of the products. Relevant word vectors contributing to this dimension are “paper”, “towel”, “save”, “mirror” and “hardwood”. If we now check the cosine similarity for “mirror” (see Table 4), some of the words that are “captured” by this direction are “furniture”, “stovetop”, “wood”, “tile” and “glass”.

Table 4

Cosine similarity for direction “mirror”

Mirror	Furniture	Stovetop	Wood	Tile	Glass
1.000	0.9567	0.9371	0.9284	0.9181	0.8993

To recall, cosine similarity computes the cosine of the angle between the two embeddings in the high dimensional space. Therefore, words with similar meanings have similar locations in the high dimensional space or will point in the same direction, which means that they will have similar values on the same dimension. These results might indicate satisfaction with eco-friendly or money-saving products as well as products suitable for different types of surfaces.

Then, "dim2_mean" represents the average value of the word embeddings in the second dimension for all the words. Some of the most relevant word vectors are "family", "super", "absorb", "towel" and "durable". This suggests that customers might like suitable products for families, which absorb well and the superior performance. This is also supported by the cosine similarity for "towel" (see Table 5), where some of the words that are "captured" by this direction are "replace", "paper", "absorb" and "reusable".

Table 5

Cosine similarity for direction "towel"

Towel	Replace	Paper	Absorb	Reusable
1.000	0.9412	0.9079	0.8443	0.8020

The next important predictor is topic 17. An increase in this topic's presence in the reviews is associated with a decrease in the log-odds of customer satisfaction, ceteris paribus. This topic mainly discusses that the product is not worth the price and that the brand is expensive, which generates disappointment.

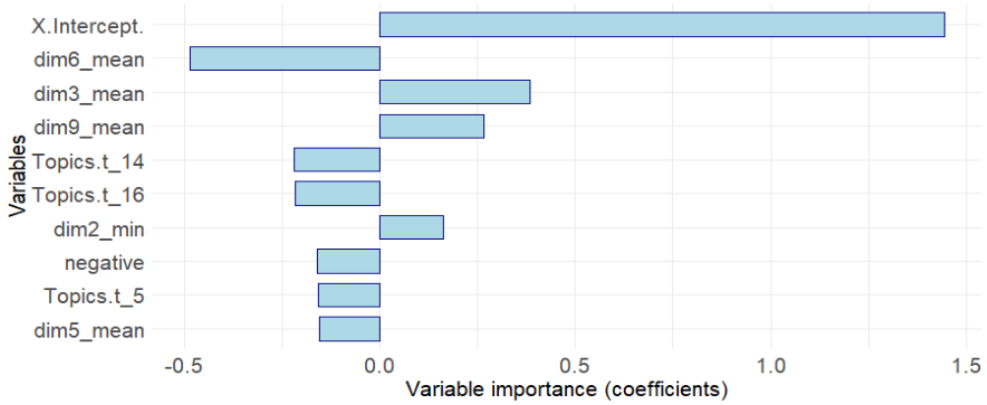
Regarding the interaction "joy.dim4_mean", the positive coefficient suggests that an increase in this interaction between the emotion of joy and the average value of the word embeddings in the fourth dimension is associated with an increase in the log-odds of customer satisfaction, ceteris paribus. Some relevant word vectors in this dimension are "pet", "safe" or "sensitive" and the cosine similarity for "safe" provides a high value also for "kid", "worried", "ingredients" and "toxic". This suggests that customers care about how safe these products for a house with pets and kids are. Therefore, the interaction with joy suggests that safe products can potentially lead to an increase in customer satisfaction.

Finally, an increase in the presence of LDA topic 20 is associated with a decrease in the log-odds of customer satisfaction. This topic is related to smell and odour as well; however, it seems to be more related to the odour of machines, which also generates disappointment given the expectations generated from other reviews.

On the other hand, *"NONsustainable"* Lasso Logistic Regression provides as important variables the ones shown in Figure 8, from highest to lowest impact in absolute value.

Figure 8

Top 10 most impactful variables for Lasso Logistic Regression with interactions in "NONsustainable" data



The positive intercept means that when all other variables are zero, the log-odds of customer satisfaction are approximately 1.44. The variable "dim6_mean", which represents the average value of the word embeddings in the sixth dimension has one of the strongest negative coefficients, which means that an increase in "dim6_mean" is associated with a decrease in the log-odds of a review being positive, all else being equal. Some words that contribute most significantly to this dimension are "leak", "box", "crack", "bottle", "waste", "arrive", "money" or "broken". The cosine similarity for "liquid" shows that words that are very similar to that word vector are "issue", "leak", "damage", and "spill". These results suggest that customers are dissatisfied when the product arrives damaged or with a leakage.

By contrast, the positive coefficient of "dim3_mean" indicates that when there is an increase in the mean value of the word embeddings in the third dimension, the log-odds of a customer being satisfied also increases, ceteris paribus. Relevant word vectors in this dimension are "scent", "fresh", "lemon", "laundry", "sensitive" or "pleasant" and

the cosine similarity for the word “lemon” indicates that “refresh”, “fresh”, “favourite” and “prefer” are close on the high-dimensional space. Therefore, this dimension captures words related mostly to scents that are fresh and respectful with sensitive skin and the coefficient suggests that customers who identify these features are more likely to be satisfied with the product.

Similarly, “dim9_mean” has a positive impact. An increase in the average value of word embeddings in dimension nine is related to an increase in the log-odds of customer satisfaction, ceteris paribus. Dimension 9 is mostly related to disinfection since some of the most relevant word vectors are “covid”, “disinfect” or “wipe”. Additionally, some brand names are mentioned, which based on cosine similarity, are also quite close in the high-dimensional space to “disinfect”.

Topics 14 and 16 have a negative coefficient, which means that their individual presence in the reviews leads to lower log-odds of a review being positive and, therefore, customer satisfaction, ceteris paribus. Topic 14 is related to the hype, hope, and expectations that customers have when reading reviews and it is mostly related to the disappointment with product quality. On the other side, topic 16 focuses on the leakage problem as well, and the product damages discovered once the product arrives.

Next, “dim2_min”, the minimum value of all word embeddings for dimension two, is expected to generate an increase in the log-odds of a customer being satisfied when it increases and everything else remains constant. Relevant word vectors are “kitchen”, “sink”, “bathroom” or “surface”. Moreover, the cosine similarity for “kitchen” is high for words such as “counter”, “easy”, “white” and the ones previously mentioned. This suggests that customers might like multi-surface products and quality cleaning wipes.

Then, when there are negative emotions expressed in the reviews or when topic 5 is present, the log-odds of customer satisfaction are lower, ceteris paribus. Topic 5 refers mostly to customers being upset by the money spent. It seems that customers also had hope and expectations based on reviews, which act as word-of-mouth (Chevalier & Mayzlin, 2006; Zhu & Zhang, 2010; Kostyra et al., 2016; Netzer et al., 2012).

Finally, “dim5_mean” has also a negative coefficient, therefore, an increase in the average value of word embeddings in the fifth dimension, leads to a decrease in the log-odds of having a positive review. The highest word vector coefficients for this

dimension are negative and the words are “recommend”, “oven”, “love”, “easy” or “grease”. The cosine similarity for “grease” indicates that similar words are “oven”, “glass”, “stove”, “pan” or “gunk”. This might indicate that some products have difficulties when cleaning greasy surfaces or utensils.

7. General Discussion & Conclusion

As stated by Kostyra et al. (2016), there is an urge for companies to understand the underlying motivations behind positive and negative reviews. Consequently, various theories and studies serve as a framework for understanding consumer behaviour and satisfaction towards eco-friendly cleaning products.

The Theory of Planned Behaviour and the Value-Belief-Norm Theory offer the basis for understanding the importance of values, beliefs, subjective norms, or social pressure in shifting consumer buying decisions towards sustainable options. Moreover, based on green marketing and growth marketing, marketers can tailor their strategies to promote sustainable consumption in this sector by applying the recommendations that are proposed in this study. Consequently, the purpose of this study is to answer the proposed research question and sub-questions:

What are the key determinants that positively influence customers' decision-making regarding the selection of sustainable cleaning products? Conversely, what are the factors that contribute to the relatively low adoption rates of these alternative options?

- Enhancing brand and product recognition: *How could these companies improve brand reputation and differentiate their products to gain a competitive advantage over non-sustainable ones?*
- Actionable insights for refining and optimizing marketing efforts: *What are the factors that can be derived from this analysis to refine and optimize marketing efforts for companies producing sustainable cleaning products? Furthermore, how can these factors be addressed and resolved to enhance their marketing strategies effectively?*

A summary of the findings can be found in the next sub-section, followed by concrete actionable recommendations and the limitations of this research.

7.1. Research Question

Based on this analysis, a potential reason influencing the low adoption rates of sustainable cleaning options is customers' pre-purchase expectations, which perfectly relates to the expectation disconfirmation theory (Oliver, 1977, 1980, 2014). Additionally, many authors have highlighted the importance of product quality on customers' satisfaction, market share, and long-term brand success (Jacobson & Aaker, 1987; Tellis & Johnson, 2007; Tirunillai & Tellis, 2012). The fact that customers feel disappointed when the product proves to be of poor quality or does not fulfil customers' expectations can lead to customer dissatisfaction. This problem is reflected in Sentiment Analysis results, where some of the most relatively frequent words in negative reviews for *"sustainable"* are "beware", "scam", "horrific", "warn", "sulfuric" or "dissatisfied".

Moreover, from the predictions of customer satisfaction, some of the strongest negative coefficients are related to expectations based on advertisements, emotions of anger and disappointment regarding the price and machine odours. This is intrinsically linked to concerns about product quality, such as the previously mentioned, or product leakage and strong scents. Finally, from Multidimensional Scaling, another defect that can discourage people to buy sustainable alternatives could be the residues of the products. This agrees with Wijekoon & Sabri (2021) findings, which highlight high costs and lack of trust, among others, as motives for why people do not adopt sustainable products.

Some of the sustainable products' strengths uncovered in this analysis are the fact that, in general, they are eco-friendly, compostable, safe for kids, pets and sensitive skin in terms of ingredients, versatile for different surfaces and, regarding towels, they are reusable and absorb quite well. However, it is important to note that some reviewers noticed possible chemicals on the products with which they are not satisfied (sulfuric).

Then, it is beneficial to also identify the competitor. From the *"NONsustainable"* data set, customers liked fresh and soft scents like lemon, multipurpose cleaners such as for the bathroom, kitchen, and tiles and especially, disinfectant properties on the products. Some of their weaknesses are also expectations, misleading advertisements, leakage problems in shipping, product residues, feeling that the product is not worth the price and difficulties to clean greasy surfaces or utensils.

7.2. Managerial implication

Although there has been considerable progress in green consumerism, sustainable companies producing cleaning products can be benefited by the following recommendations.

There are two common weaknesses for both markets. First, based on relationship marketing theory, sustainable companies should avoid false claims and advertisements about their products, since this fosters negative "word-of-mouth" on reviews and prevents the creation of a relationship of trust with the customer. Additionally, developing new processes to evaluate ingredients and materials is fundamental to ensure that there are no harmful chemicals in the products. Therefore, in the short term, advertisements can attract customers. However, in the long term, if the product quality and functionality do not meet or exceed expectations, individuals could feel betrayed, upset or angry, which impacts loyalty and adoption rates (Mazurek, 2019).

This leads to the second mistake: sustainable companies should address the problems of product quality. Especially, the discontentment with the leakage, machine odours, product residues and too-strong scents. Consequently, sustainable companies should ensure secure packaging and robust shipping methods. Moreover, investing in technologies that minimize the residue formation on products and that control washing machine odours is advisable based on customer reviews. Then, customers expressed their preference for fresh and soft scents, therefore, companies should consider focusing on fragrance quality and variety that aligns with this information.

Sustainable companies can also benefit from another competitor's strength, which is the development of disinfectant products since this had a strong positive impact on *"NONsustainable"*. Investigating sustainable disinfectant ingredients and building communication between industry segments can potentially impact customers' attitudes. Finally, it is important to keep working on strengths such as money-saving, reusable options, and quality and safe products such as absorbent towels.

These recommendations can potentially provide a more positive experience for customers and allow sustainable companies to distinguish themselves from non-sustainable competitors as well as gain market share. By addressing weaknesses, enhancing strengths, and learning from competitors' SWOT, sustainable companies can

create a competitive advantage. Additionally, by focusing on product and brand quality and recognition they can build trust and loyalty as well as a positive brand image that resonates with individuals. Developing alternatives that are functionally and environmentally sound and clearly exposing their benefits can attract environmentally conscious consumers who might be willing to pay a slightly higher price for quality sustainable alternatives.

Finally, once the product quality is ensured, sustainable companies can benefit from incentivized referral programs or marketing techniques such as partnerships with influencers, which can increase social sharing and brand recognition.

7.3. Limitations of my study

First, one of the limitations of this study is that it is based solely on Amazon reviews, which neglects other potential sources of customer feedback and sentiment. Additionally, although reviews were collected only for customers with verified purchases, the study may not be representative of the entire population. The reason is that it cannot capture the sentiment of those customers who did not leave reviews, which might create self-selection bias. Moreover, this study may lack contextual information about the customers such as purchase history or demographics, which could provide a deeper understanding of consumer preferences and motives. This could also help to target even more marketing messages and to deliver personalized offers and content for customers.

Another limitation relates to Sentiment Analysis. This technique has inherent drawbacks, such as the difficulty in accurately capturing sentiments when customers use sarcasm or irony. Additionally, sentiment analysis might have problems dealing with out-of-vocabulary words or domain-specific terms that might not be included in the lexicon.

Potential areas of future research could address some of these limitations by considering the incorporation of in-depth interviews, social media, forums, and online discussion data into the analysis. Additionally, future research could include contextual data such as demographics to provide a more fine-grained analysis. However, a major improvement could be to include purchase history, which could allow us to differentiate

between those customers who bought the sustainable cleaning product once and those who became loyal and kept buying it. Then, Sentiment Analysis could be benefited from using contextual embeddings such as Bidirectional Encoder Representations from Transformers (BERT) or Generative Pre-trained Transformer (GPT) since these methods can better comprehend the nuances of language such as irony or sarcasm by analysing the context words and phrases.

Finally, an interdisciplinary future research study could be benefited from the expertise and skills of marketing analysts but also behavioural professionals. Collaborating can potentially enhance the models' accuracy and interpretability, which will result in a deeper understanding of complex consumer decision-making processes regarding sustainable cleaning products.

8. Bibliography

Aaker, D. A. (1996). Measuring Brand Equity Across Products and Markets. *California Management Review*, *38*(3), 102–120.

Ajzen, I. (1985). From Intentions to Actions: A Theory of Planned Behavior. In J. Kuhl & J. Beckmann (Eds.), *Action Control: From Cognition to Behavior* (pp. 11–39). Springer.

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, *3*, 42.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science*, *57*(8), 1485–1509.

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*(4), 557–590.

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, *84*, 1–25.

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

Boegershausen, J., Datta, H., Borah, A., & Stephen, A. T. (2022). Fields of Gold: Scraping Web Data for Marketing Insights. *Journal of Marketing*, *86*(5), 1–20.

Carrington, M. J., Neville, B. A., & Whitwell, G. J. (2010). Why Ethical Consumers Don't Walk Their Talk: Towards a Framework for Understanding the Gap Between the Ethical Purchase Intentions and Actual Buying Behaviour of Ethically Minded Consumers. *Journal of Business Ethics*, *97*(1), 139–158.

Chen, Y., & Lee, S. (2023). User-Generated Physician Ratings and Their Effects on Patients' Physician Choices: Evidence from Yelp. *Journal of Marketing*.

Chevalier, J. A., & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, *43*(3), 345–354.

DataReportal, D. (2023). *Digital Around the World—Global Digital Insights*. https://datareportal.com/global-digital-overview

Ellis, S., & Brown, M. (2017). *Hacking growth: How today's fastest-growing companies drive breakout success*. New York : Crown Business, 2017.

Fielding, K. S., Terry, D. J., Masser, B. M., & Hogg, M. A. (2008). Integrating social identity theory and the theory of planned behaviour to explain decisions to engage in sustainable agricultural practices. *British Journal of Social Psychology*, *47*(1), 23–48.

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science*, *31*(3), 493–520.

Gonzalez-Arcos, C., Joubert, A. M., Scaraboto, D., Guesalaga, R., & Sandberg, J. (2021). "How Do I Carry All This Now?" Understanding Consumer Resistance to Sustainability Interventions. *Journal of Marketing*, *85*(3), 44–61.

Gordon, M. E., McKeage, K., & Fox, M. A. (1998). Relationship marketing effectiveness: The role of involvement. *Psychology & Marketing*, *15*(5), 443–459.

Jacobson, R., & Aaker, D. A. (1987). The Strategic Role of Product Quality. *Journal of Marketing*, *51*(4), 31–44.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York.

Kostyra, D. S., Reiner, J., Natter, M., & Klapper, D. (2016). Decomposing the effects of online customer reviews on brand, price, and product attributes. *International Journal of Research in Marketing*, *33*(1), 11–26.

Krawczyk, M., & Xiang, Z. (2016). Perceptual mapping of hotel brands using online reviews: A text analytics approach. *Information Technology & Tourism*, *16*(1), 23–43.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Kurz, T., Gardner, B., Verplanken, B., & Abraham, C. (2015). Habitual behaviours or patterns of practice? Explaining and changing repetitive climate-relevant actions. *WIREs Climate Change*, *6*(1), 113–128.

Kwartler, T. (2017). *Text Mining in Practice with R*. Wiley.

Laheri, V. K., Dangi, H., & Vohra, A. (2014). Green Marketing: Development of Construct and Its Evolution. *Asia-Pacific Journal of Management Research and Innovation*, *10*(2), 147–155.

Lee, T. Y., & Bradlow, E. T. (2011). Automated Marketing Research Using Online Customer Reviews. *Journal of Marketing Research*, *48*(5), 881–894.

Liu, L., Dzyabura, D., & Mizik, N. (2020). Visual Listening In: Extracting Brand Image Portrayed on Social Media. *Marketing Science*, *39*(4), 669–686.

Long, D. C. (2018). Greening of Consumer Cleaning Products. In W. Zhang & B. W. Cue (Eds.), *Green Techniques for Organic Synthesis and Medicinal Chemistry* (pp. 91–115). John Wiley & Sons, Ltd.

Mair, P., Borg, I., & Rusch, T. (2016). Goodness-of-fit assessment in multidimensional scaling and unfolding. *Multivariate Behavioral Research*, *51*(6), 772–789.

Markham, S. K., Kowolenko, M., & Michaelis, T. L. (2015). Unstructured Text Analytics to Support New Product Development Decisions. *Research-Technology Management*, *58*(2), 30–39.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442–451.

Mazurek, M. (2019). Brand Reputation and its Influence on Consumers' Behavior. In S. Grima, E. Özen, H. Boz, J. Spiteri, & E. Thalassinos (Eds.), *Contemporary Issues in Behavioral Finance* (Vol. 101, pp. 45–52). Emerald Publishing Limited.

Mishra, M. (2022). Customer Experience: Extracting Topics From Tweets. *International Journal of Market Research*, *64*(3), 334–353.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science*, *31*(3), 521–543.

Oliver, R. L. (1977). Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of Applied Psychology*, *62*(4), 480–486.

Oliver, R. L. (1980). A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research*, *17*(4), 460–469.

Oliver, R. L. (2014). *Satisfaction: A Behavioral Perspective on the Consumer*. Routledge.

Olsen, M. C., Slotegraaf, R. J., & Chandukala, S. R. (2014). Green claims and message frames: How green new products change brand attitude. *Journal of Marketing*, *78*(5), 119–137.

Oppong, A., & Caesar, L. D. (2023). A contingency analysis of brand reputation and loyalty in the banking sector. *SN Business & Economics*, *3*(7), 113.

Peattie, K., & Charter, M. (2002). Green marketing. In *The Marketing Book*. Routledge.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, *211*(2828), 42–45.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Plutchik, R. (1982). A psychoevolutionary theory of emotions. *Social Science Information*, *21*(4–5), 529–553.

Porter, M. F. (2006). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, *40*(3), 211–218.

Rabinowitz, G. B. (1975). An Introduction to Nonmetric Multidimensional Scaling. *American Journal of Political Science*, *19*(2), 343–390.

Radojevic, T., Stanisic, N., & Stanic, N. (2017). Inside the Rating Scores: A Multilevel Analysis of the Factors Influencing Customer Satisfaction in the Hotel Industry. *Cornell Hospitality Quarterly*, *58*(2), 134–164.

Rust, R. T., Rand, W., Huang, M.-H., Stephen, A. T., Brooks, G., & Chabuk, T. (2021). Real-Time Brand Reputation Tracking Using Social Media. *Journal of Marketing*, *85*(4), 21–43.

Sridhar, S., & Srinivasan, R. (2012). Social Influence Effects in Online Product Ratings. *Journal of Marketing*, *76*(5), 70–88.

Stern, P. C., Dietz, T., Abel, T., Guagnano, G. A., & Kalof, L. (1999). A Value-Belief-Norm Theory of Support for Social Movements: The Case of Environmentalism. *Human Ecology Review*, *6*(2).

Tellis, G. J., & Johnson, J. (2007). The Value of Quality. *Marketing Science*, *26*(6), 758–773.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tirunillai, S., & Tellis, G. J. (2012). Does Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance. *Marketing Science*, *31*(2), 198–215.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463–479.

W. Jones, T. (2021). *Topic modeling*. https://cran.r-project.org/web/packages/textmineR/vignettes/c_topic_modeling.html

White, K., Habib, R., & Hardisty, D. J. (2019). How to SHIFT Consumer Behaviors to be More Sustainable: A Literature Review and Guiding Framework. *Journal of Marketing*, *83*(3), 22–49.

Wijekoon, R., & Sabri, M. F. (2021). Determinants That Influence Green Product Purchase Intention and Behavior: A Literature Review and Guiding Framework. *Sustainability*, *13*(11).

Zhu, F., & Zhang, X. (Michael). (2010). Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. *Journal of Marketing*, *74*(2), 133–148.

9. Appendix

Table 1

Extract of sustainable cleaning products data set

Rating	Title	Review	Location	Date
5	It makes a lots of suds and clean my dishes.	It makes a lots of suds with a small amount of liquid.	United States	2023-05-14
5	My favorite sponge cloth	Superb quality!! They are just the right amount of thickness and absorb so well. PERFECT PERFECT PERFECT for wiping counters!! They are somewhere between a sponge and a rag, like an amazing hybrid, kind of hard to describe (see video). Absolute perfect size for dishes as well as wiping countertops. Comes in plastic free packaging AND from a small business, win win. Environmentalist and minimalist approved ⭐⭐⭐⭐⭐Super happy tysm!!	United States	2023-05-13
3	these are not any better than any others i have used	these are not any better than any others i have used . they do the job !	United States	2023-05-13
2	not for me	this is very very stiff when it's dry. it does absorb super well but doesn't ring dry and starts to smell bad very quickly.	Canada	2022-07-30
5	It Works!	Not quite what I imagined them to be, but still does the job!	United Kingdom	2021-05-14

Figure 1

Random Forest final prediction decision by majority vote

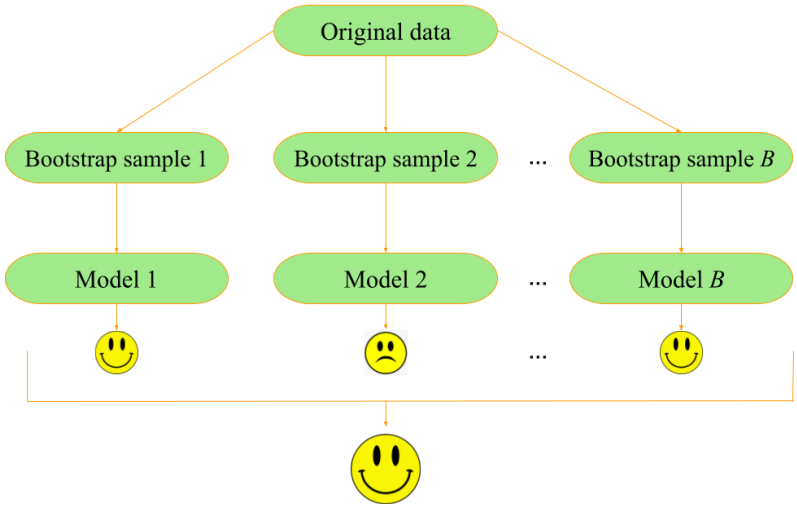


Figure 2

Scatter plots for "sustainable" sentiment score and star ratings

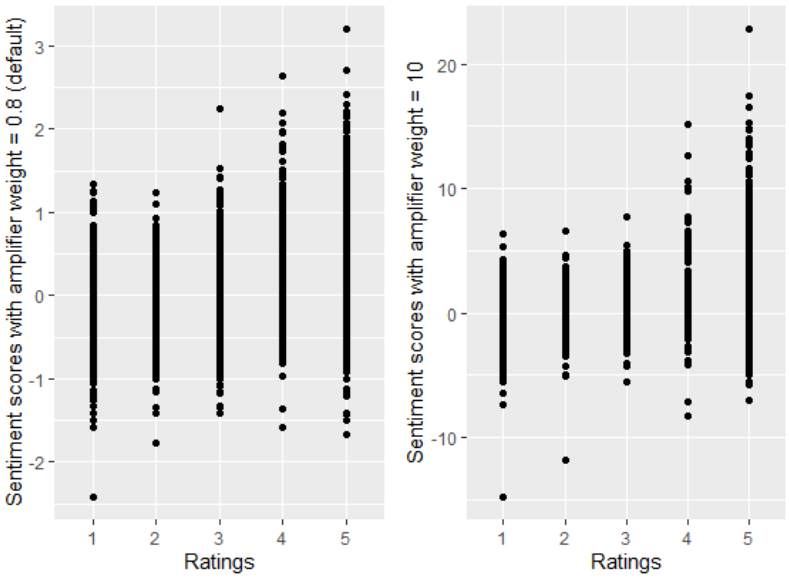


Figure 3

Scatter plots for "NONsustainable" sentiment score and star ratings

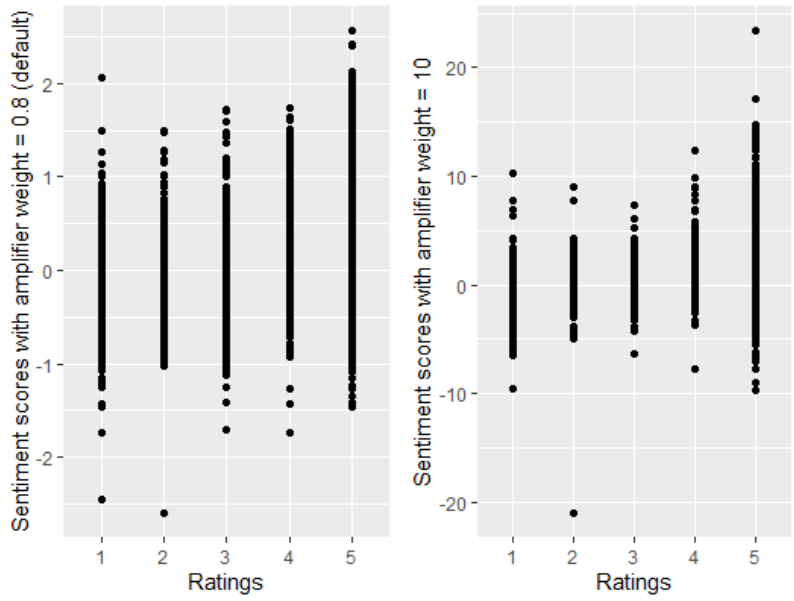


Figure 4

Distribution of sentiment scores for sentences in "sustainable"

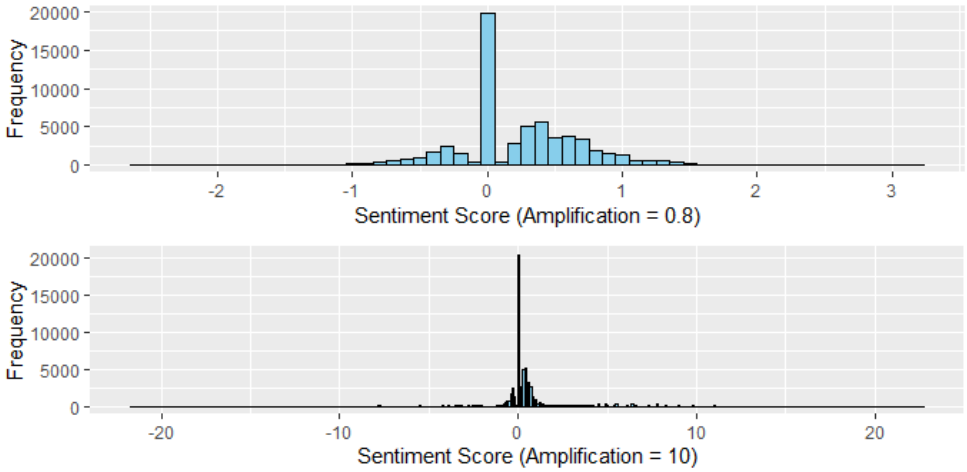


Figure 5

Distribution of sentiment scores for sentences in "NONsustainable"

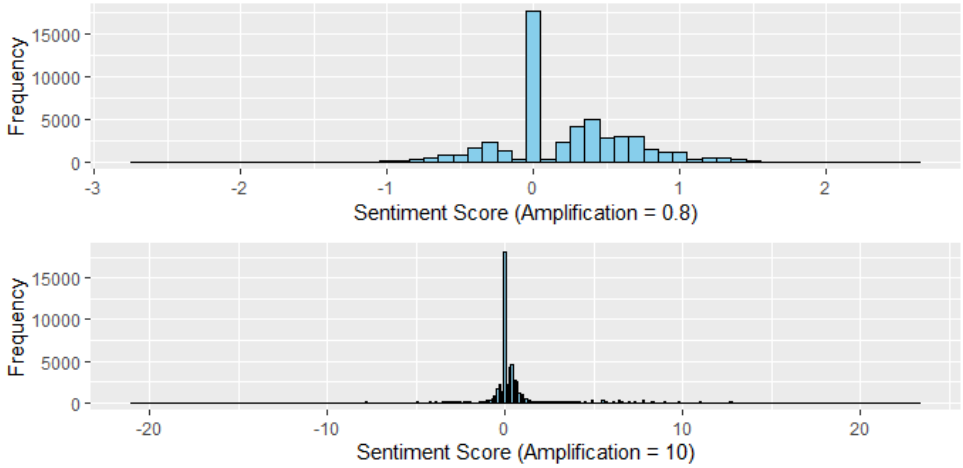


Figure 6

Plutchik's Wheel of Emotions

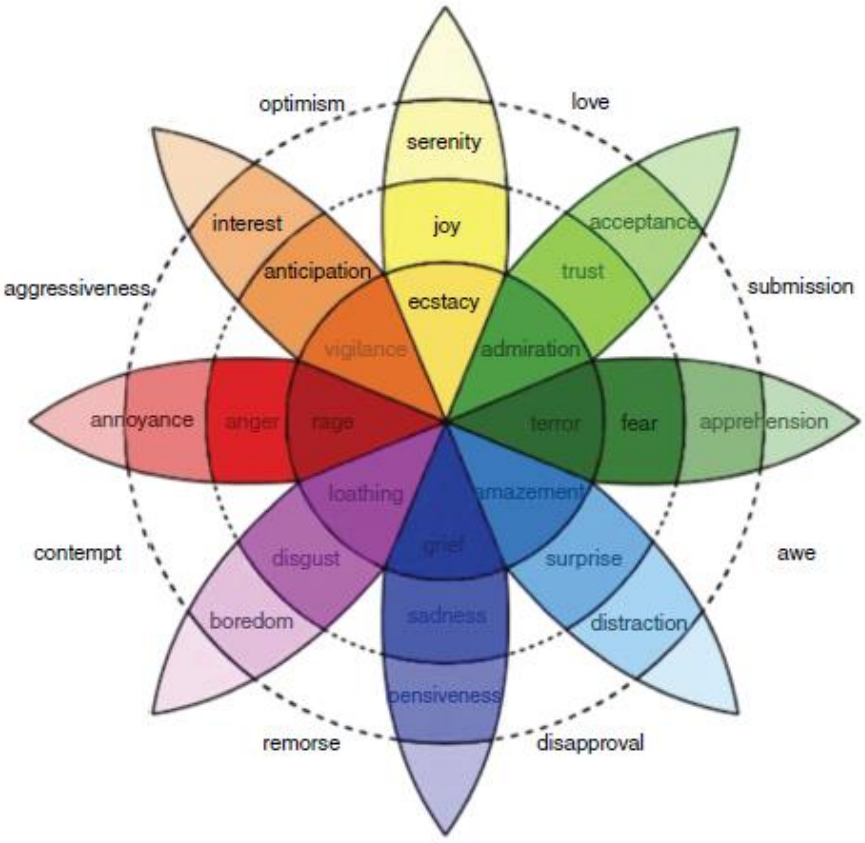


Figure 7

"sustainable" LDA topics description

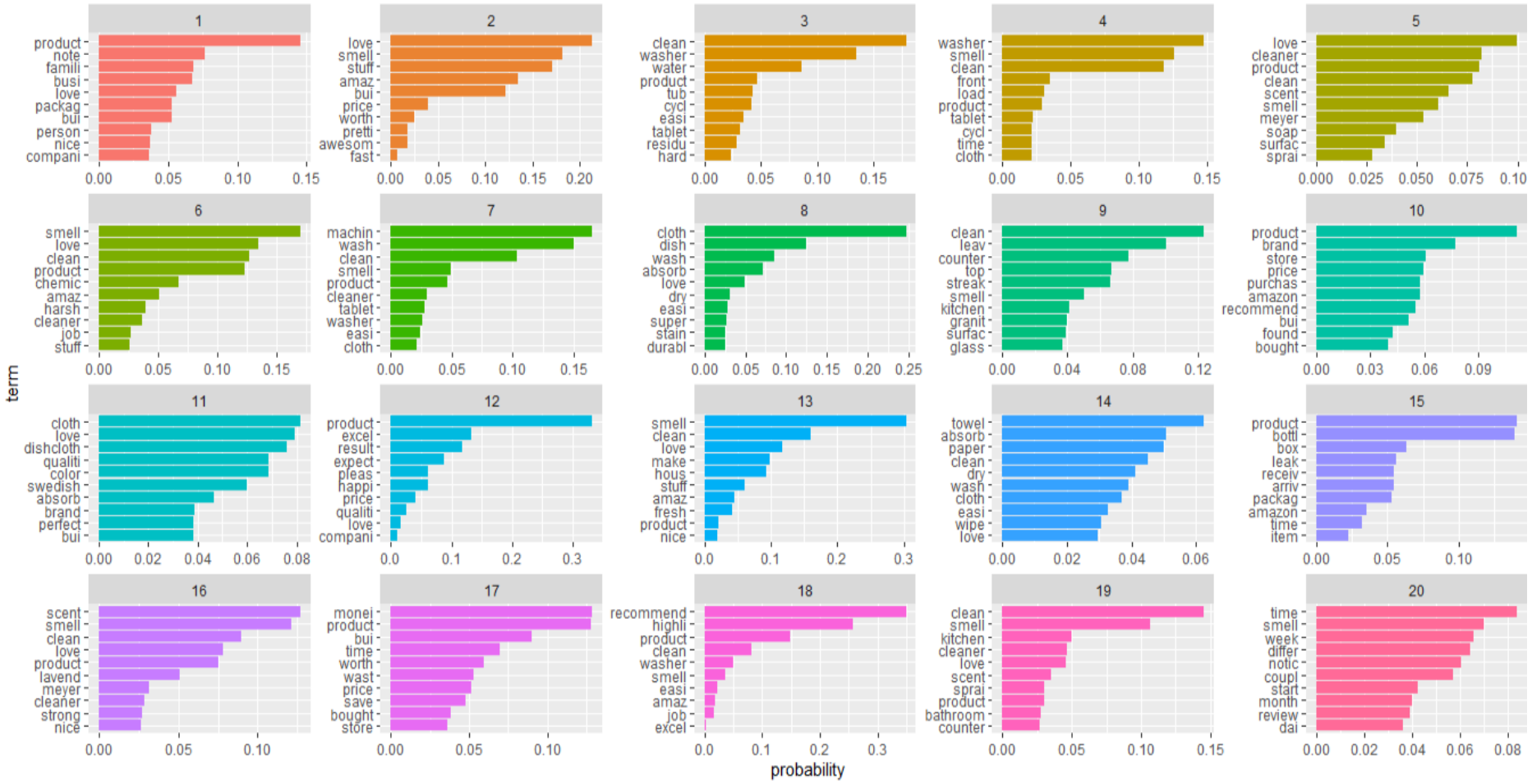


Figure 8

"NONsustainable" LDA topics description