Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis
Economics and Business
Data Science and Marketing Analytics

An assessment of feature mining and sentiment analysis techniques for extracting customer needs from online product reviews

Author:	*Supervisor:*
Cristian Rojas	Prof. Dr. Patrick Groenen
Student number:	*Second assessor:*
561827	Prof. Dr. M.G. de Jong

Abstract

Driven by customer-centricity, product design methodologies identify the value of understanding customer needs (CNs) in the early steps of product development. Traditional sources for this information, such as surveys, interviews and focus groups, have been described as expensive, time-consuming, and unfit for responding to new fast-paced markets. Contrarily, digital data sources offer richer, cheaper, and faster customer knowledge. Information Retrieval (IR) techniques that leverage online customer reviews (OCRs) have gathered much interest in the last two decades, yet a low proportion of them have focused on mining CNs as an input for product design procedures. In our thesis, we divide the mining task in two: feature extraction and sentiment analysis. First, we assess Bag of Words, Part of Speech tagging and Dependency Parsing, and secondly a Sentence- and Aspect-Based Sentiment Analysis. From our experiment, we conclude that the framework that combines Dependency Parsing and Aspect Based Sentiment Analysis is the most suitable for extracting CNs from OCRs.

Table of contents

1. Introduction

Driven by technology advancements and increasingly competitive markets, Product Design methodologies have changed dramatically over the past century (Koomsap and Risdiyono 2013). For many years, this task was commonly approached from a manufacturer perspective, disregarding customer preferences in benefit of production efficiency. In the 1950s, as new information technology capabilities allowed for significantly better collection, storage and analysis of customer data, product design slowly shifted towards a customer-centric approach. However, it was only during the 1990s that this new perspective witnessed widespread adoption and became central for the success of most companies (Shah et al. 2006). Since then, customer data is widely accepted as one of the most valuable assets for a firm and failing to anticipate customer needs by launching a product that does not address them, can produce severe financial losses (M. M. Tseng and Du 1998).

Product design continued to evolve through the years and several methodologies now recognize the importance of identifying customer needs (CNs) early on. Axiomatic Design states that functional requirements, which are the minimal set of requirements that a design must satisfy, need to be translated from CNs that are traditionally obtained through methods such as surveys, interviews, focus groups or ethnographic studies (Kulak, Cebi, and Kahraman 2010). Similarly, the Kano model of customer satisfaction identifies CNs through a questionnaire where customers express sentiments (i.e., if they feel satisfied or unsatisfied) towards a hypothetical situation, which are then used to define the functional requirements (Tontini 2007). Quality Function Deployment methodology starts by listening to the customers' voices through personal interviews and focus groups where customer express their needs in their own words, to then be used in the definition and prioritization of product requirements (Akao 1990; Clausing 1993; Cohen 1995). Similarly, choice-based conjoint analysis is used to extract customer preferences over a set of product features, through surveys that force the respondents to choose between products with similar features but differently combined, mimicking a real-life decision process (Wang and Tseng 2011).

While these methods have proven useful for extracting CNs, they also pose several challenges for researchers and companies. In fact, the process of translating CNs into tangible design parameters (DP) can be expensive and time-consuming, as it mainly depends on the expertise of the design team (Ireland and Liu 2018). Moreover, Wang and Tseng (2008) recognized three main limitations in these systems as they rely heavily on (1) the ability of the design team to identify the right customer group to extract knowledge from; (2) on the customer's knowledge and experience with the product and its limitations; and (3) on their competence and willingness to express their needs and preferences through pre-defined formats or in terms that the design team understands. Hence, even when these constrains are effectively dealt with, the data extracted (CNs) may not be ideal because as researchers must elicit customer responses, their thoughts and opinions are not freely expressed and can differ substantially from what they really think (Ulrich 2003). Often, these challenges and miss-communications between customers and designers result in convoluted interpretations of CNs, triggering changes in product specifications along the design process increasing costs and time to market (Tseng, Jiao, and Merchant 1996; M. M. Tseng and Jiao 1998; Nellore 2001; Tseng, Kjellberg, and Lu 2003).

Consequently, as more retailers shift towards developing responsive systems to better address their CNs, and as new products are being introduced more frequently, it becomes crucial to develop faster and more effective methods for extracting more accurate CN and translating them into design features (Calantone et al. 2010; Wang, Mo, and Tseng 2018). With the ever-growing amount of online customer reviews (OCR) and opinions available since the early 2000s, a growing body of literature has examined new frameworks so that user-generated content may be leveraged as a tool for improving product design processes. For example, Hu and Liu (2004) in their seminal paper were some of the earliest to see the value of mining OCRs observing promising results with their "feature-based opinion summarization" algorithm and stated that it could already be implemented in practical settings. Some years later, Kang and Zhou (2013) outperformed state-of-the-art techniques when investigating the extraction of new, subjective, features improving upon previously researched methods. More recently, Chen et al. (2019) successfully extracted product features by analysing Kindle E-reader OCRs through Google Cloud Platform, making it easier for firms to include in their operations. The next chapter of this thesis examines in detail these and other proposed frameworks.

Despite this interest, there is a considerable amount of research focusing on the effects of OCRs from the customer perspective, for example, as deciding factors for purchase decision in undecided customers and their effects in overall sales for online retailers (Dellarocas 2003;

Chevalier and Mayzlin 2006). Still, in the context of product design, the number of studies attempting to connect OCRs to design tasks is significantly lower (Yang et al. 2019), and some of the proposed techniques are still unsuitable for industrial practice (Lutters et al. 2014). However, recent findings support the idea that within the next few years, examining OCRs for the latter reasons is likely to become an important component in research and firms alike. Additionally, in light of data analytics trends, Lee, Kao, and Yang (2014) described this field of research as highly lucrative given that can help designers create more competitive products. Similarly, Timoshenko and Hauser (2019) stated that OCRs are equally or more valuable as source of CNs than conventional methods, and that machine learning methods are more efficient (unique customer needs per unit of professional services cost) when identifying CNs from OCRs.

While these newly developed systems have proven useful for overcoming the pitfalls of classic methods, they are not free of drawbacks. Similar to the slow adoption of customer centric product design in the 1990s, the key problem with the collection, storage and analysis of vast amounts of OCR, regarded as a form of Big Data (Chen et al. 2019), is that relies heavily on processing/computational power. Moreover, unlike structured data (e.g., views, bookings, height, weight, etc.), OCRs are a type of unstructured data comprised of human expressions in text format (often including informal writing), which entails a fundamental challenge for any computational approach (Archak, Ghose, and Ipeirotis 2011; McAfee et al. 2012; Netzer et al. 2012). Fortunately, the advancements seen in the last decade in fields like Machine Learning, Artificial Intelligence, Cloud Computing, and the fast growth in the computational power of consumer-level CPUs and GPUs, allow for techniques that were seldom seen in business to become increasingly important for firms.

This thesis attempts to address two main gaps within this new field of research. First, over the last two decades a wide array of methodologies have been developed without a systematic review or unified approach still available for this task. Secondly, as Ireland and Liu (2018) identified, the majority of the empirical case studies disregard the value of analysing a bigger sample of competitors by focusing on products rather than product categories or sub-markets. Thus, the purpose of this research is to find the most suitable method for the extraction of CNs from OCRs extracted from several products within the same category in Amazon. Such approach will allow us to answer the following research questions:

Which (of the reviewed techniques) is the best method for extracting customer needs from online customer reviews of one Amazon product category? What insights (customer needs) does this method allow us to extract?

To answer these questions, first we search for relevant techniques from recent studies and narrow down the number of methods for empirical evaluation; secondly, the few selected methodologies will be implemented using an original dataset, which includes Amazon customer reviews from several best-selling products within the same category; and lastly, a final framework will be selected as the *best* by using relevant model evaluation metrics, ease of use, and interpretability of the information extracted. Accordingly, the following sections of this thesis are structured as follows: Section 2 is a literature review focused on product design trends, information extraction systems and appropriate Natural Language Processing (NLP) techniques. Section 3 will provide details on data collection process and manual annotation used as benchmark for model evaluation. Section 4 will describe the proposed methodologies for the case study, model evaluation and comparison, and the results will be shown in Section 5. Finally, in Section 6 we will discuss our results and offer conclusions.

2. Literature Review

This section is divided into three subsections. First, we introduce relevant product design trends and definitions. Then, we introduce some of the relevant work in the field of information extraction from online customer reviews. Finally, we introduce an important concept, recently defined in literature, that further supports the purpose of this thesis.

2.1. Product Design and Online Customer Reviews

The sustained increase in demand for manufacturing and consumer goods, mainly driven by (1) the rise in developing countries, (2) shortening in product life cycles and (3) the introduction of new products at a faster rate, have put pressure on mainstream product design and manufacturing processes (Schuh et al. 2014). Traditionally, these approaches are dependent on traditional tools intended for eliciting customer knowledge, such as surveys, interviews, focus groups, ethnographic studies, etc. (Kulak, Cebi, and Kahraman 2010). In addition to being vastly recognized as time consuming and costly, several more limitations have been identified in literature and were discussed in the previous section (Tseng, Jiao, and Merchant 1996; M. M. Tseng and Jiao 1998; Nellore 2001; Ulrich 2003; Tseng, Kjellberg, and Lu 2003; Wang and Tseng 2008; Ireland and Liu 2018).

In response, several authors have proposed that a migration from classic methodologies towards digital data sources is necessary as it can provide richer, cheaper, and faster information, as online reviewers – previously known as the respondents or participants in classical methods – give their feedback voluntarily and at no cost (Groves 2006; Chevalier and Mayzlin 2006; Qi et al. 2016). From a business perspective, OCRs have become the centre of attention of many researchers (Mudambi and Schuff 2010). For example, Ba and Pavlou (2002) and Pavlou and Gefen (2004) examined the positive feedback mechanisms that OCRs can have on buyer’s trust. Similarly, Clemons, Gao, and Hitt (2006) observed that strongly positive OCRs increased product sales and Chen, Dhanasobhon, and Smith (2008) concluded that reviews regarded as highly useful by other customer also influence sales positively. On the other hand, significantly fewer attempts have been made to investigate the unique value of OCRs for manufacturers and designers, particularly in terms of how to understand and process large amounts of OCRs through frameworks that can extract useful customer knowledge (Yang et al.

2019). In fact, such insights have been described as particularly important during the early stages of product development (Hu and Liu 2004; Hedegaard and Simonsen 2013; Jin et al. 2016; Jin, Ji, and Liu 2014; Jin, Ji, and Gu 2016; Jin, Ji, and Kwong 2016).

The fast growth of e-commerce and the lethargy with which current product design frameworks respond to fast-changing customer needs, makes more evident the need to close the knowledge gap and develop more agile and responsive systems (Dellarocas 2003; Yin, Bond, and Zhang 2014; Wang, Mo, and Tseng 2018, 2018; Ireland and Liu 2018; Calantone et al. 2010). As a result, different authors have established methodologies for extracting information from OCRs to aid product design processes, although there is yet no consensus on the scope of these systems, i.e., the type of output that designer's value most. For instance, Lee and Bradlow (2007) proposed a framework for extracting product features and their levels and use as the initial input for Conjoint Analysis. In the study of Ireland and Liu (2018) they proposed the extraction of word pairs (feature + sentiment) from OCRs, to offer insights that empower product designers to make better decisions (still requiring human analysis). Contrarily, Wang, Mo, and Tseng (2018) attempted further automation with a system that first extracts CNs and then maps them into Design Parameters (DP) using a deep learning model. In their study, they treated keywords from OCRs as CNs supported by the work of Timoshenko and Hauser (2019) who observed significant similarities between OCRs' contents and CNs extracted with traditional methods. Most recently, Chen et al. (2019) leveraged different Machine Learning and AI techniques through a cloud-based framework intended for small- and medium-sized companies, as it offers straightforward interpretation of results.

2.2. Relevant Work

One of the first studies that tried to leverage customers' feedback was motivated by the idea of automatically classifying customer reviews, and even though feature extraction was out of the scope of their research, Turney and Littman (2002) shed light over a thriving field of study. They designed a simple unsupervised algorithm that (1) identifies phrases that contain adjectives and adverbs, (2) estimates their semantic orientation (SO) and (3) classifies the review. The second step, regarded as the core of the method, used Point Wise Mutual Information (PMI) (Turney 2001) to assess the similarity between pairs of words or phrases (calculated by comparing its similarity with a positive reference word to its similarity with a negative one). Each review was then classified as ¨recommended¨ if its average SO was positive, or as ¨not recommended¨ if negative. Turney´s novel method for calculating SO observed a mixed performance, in part due to the nature of the products being analysed. In fact, in their experiment the accuracy obtained when analysing OCRs of banks and cars was much higher than for movies, where the semantic orientation of phrases such as ¨more evil¨ hindered the classification as they received a negative SO, yet do not mean that a film was not worth recommending. Hence, in the methodology section we address this limitation by including steps to identify and deal with nuanced words.

The foundations for extracting CNs from reviews were laid in the 80s and 90s, and by the beginning of the 2000s, most of the previous work on text summarization had mainly focused on (1) the identification and extraction of certain core entities and facts in a document (DeJong 1982; Tait 1982; Radev and McKeown 1998); or (2), on the development of text extraction frameworks (Paice 1990; Kupiec, Pedersen, and Chen 1995; Hovy, Lin, and others 1999), that identify some representative sentences to summarize a document. Although these approaches were interesting, they suffered from not being domain independent (Jones 1993a, 1993b), i.e., they required domain-specific knowledge, and more importantly, could fail to extract customer needs and preferences as their opinions are not always representative of a text (Hu and Liu 2004).

Noticing this gap and the potential value of extracting information from OCRs, Hu and Liu (2004) were among the first authors that identified and dealt with these shortcomings. Although from a customer perspective rather than designer, they coined a two-step approach, or feature-based opinion summarization system, based on data mining and natural language processing methods like Part of Speech (POS) tagging or grammatical tagging (labelling words as verbs, nouns, adjectives, and so on). Unlike traditional text summarization, their methodology did not summarize reviews by rewriting a subset of the original sentences to convey its main ideas. Instead, based on the assumption that people often use the same words when they express their thoughts, it (1) used association rule mining (Agarwal, Srikant, and others 1994) to extract product features (nouns) about which consumers had voiced opinions (adjectives); (2) ranked these opinion features (noun + adjective) according to how frequently they occurred in reviews, and (3) it determined the semantic orientation of the opinion sentences based on the dominant orientation of the opinion words within each sentence.

The experiment conducted to assess their framework included reviews from five electronic products (2 digital cameras, 1 DVD player, 1 mp3 player, and 1 cellular phone) from Amazon and CNet. To evaluate the extracted features and semantic orientation, a human tagger read all reviews and generated (1) a list of features per each product, both explicit (e.g., *pictures* in *"the pictures were absolutely amazing"*) and implicit (e.g., *"size"* in *"it fits in a pocket nicely"*), and (2) semantic orientation of opinion features. The best model performance was obtained when they included the identification of infrequent features in their system. This last step was further detailed in a subsequent publication (Hu and Liu 2006) and required authors to build a lexicon of words with a binary classification (positive or negative depending on its SO). The Bing Sentiment Lexicon (Hu and Liu 2004) was included in their publication and now serves as the basis for several Sentiment Analysis algorithms. The authors qualified their results as promising and concluded that these techniques were effective for dealing with this task.

Based on their conclusions and seeing that their work is often used as benchmark in subsequent publications, we regard it as particularly relevant for our research and therefore we include several techniques from their methodology in this thesis.

Similar solutions have been proposed for this task over the years. Namely, Ren (2007) and Popescu and Etzioni (2007) validated the approach built by Hu and Liu (2004), by arguing that an overall negative sentiment classification of a review does not necessarily imply that a customer completely dislikes a product (and vice versa). Thus, they built on the same intuition and heuristic behind the previous feature-based approach: (1) an opinion word (adjective) associated with a product feature (noun) will occur in its vicinity, and (2) feature/opinion pairs from reviews mostly appear as noun/adjective pairs in the same phrase. More particularly, Ren (2007) reasoned that using association rule mining was a complicated method for extracting frequent features, and so after investigating the use of PMI between two words (Turney and Littman 2002), they reached the conclusion that a simplified approach was also viable. On the other hand, Popescu and Etzioni (2007) performed a systematic review on the five datasets used in (Hu and Liu 2004) and, using the original results as a baseline, reported significant improvements in the subtask of feature extraction through their OPINE algorithm. Their novel unsupervised system, which was based on the KnowItAll information-extraction system (Etzioni et al. 2005), used PMI-IR to assess if feature candidates were actual product features before extracting them. Nonetheless, despite their favourable results these three approaches were only well suited for identifying and extracting explicit features but were not capable of detecting implicit ones. Likewise, they only offered a positive/negative classification which did not entail how strong those sentiments were (Kang and Zhou 2013).

In 2007 and from a designer perspective, Lee and Bradlow (2007) proposed a method for automatic exploration and extraction of product attributes and their levels, to serve as initial input for Conjoint Analysis studies. Exploiting the co-occurrences of words within customer generated pros and cons lists, allowed for a completely unsupervised bag-of-words approach (disregarding grammar and order) that avoided the complexities of NLP techniques. The graph-based methodology clusters pros and cons phrases into product attributes, then splits them into dimensions and finally, each dimension is further divided into levels. For example, by analysing thousands of entries like ¨*Long 6x optical zoom*¨, ¨*standard 3x optical zoom*¨ and ¨*nice digital zoom*¨, their system creates an output such as “*Attribute*: *zoom”*; *“Dimension: magnification”*; and *“Levels: 2x, 3x”*. Alas, the number of limitations their system posed reflected in low evaluation scores and in ambiguous outputs that required more interpretation.

Their results remain important for our work, as here we address these limitations by adopting some of the authors’ propositions regarding further implementation of Natural Language Processing (NLP) techniques to (1) extract bigger n-grams, instead of just unigrams (one word), and (2) deal better with synonyms, misspellings, slang, etc.

Two years later, Qiu et al. (2009) identified the weaknesses of lexicon-based sentiment methods, stating that as different words are used in different domains or contexts, it is almost impossible to collect and maintain one universal lexicon. To deal with this, they developed a state-of-the-art double propagation system that exploits the relationships between sentiment words (adjectives) and product features (nouns). By first using POS tagging, their method extracts a few sentiment words and product features using a seed sentiment lexicon, then searches for new sentiment words and features using the existing ones as a starting point. Then again, with the newly acquired sentiment words and features, more are extracted in the same way. The procedure is repeated until no more sentiment words are added. Every time a new sentiment word is extracted, it inherits the polarity of the word used to extract it (unless negations are present within a five-word window). It is important to note that this work differed from previous ones as it uses syntactic relations or dependency grammar to describe relationships between words, rather than a distance-based approach. To understand this, the author offered the following example: *“The newly released iPod is amazing”* in which *“newly”* depends on *“released”* which depends on *“iPod”* and *“iPod”* itself depends on *“is”*. This type of relationship, where one word directly depends on another (or both depend on a third word directly), was named Direct Relationship and was the only one used in their study by stating that more complex ones were not suited for the informal nature of the text found in OCRs. In their experiment, they used the original sentiment lexicon of 1752 words in increasing proportions (10%, 20%, 30%, and so on) as seeds and compared to other methods. The authors concluded that their model was powerful at generating large numbers of new sentiment words, with a good level of accuracy for assigning polarity scores. However, while these statements were true and the approach offered an improvement to contemporary methodologies, Kang and Zhou (2013) observed three main problems with it: (1) the model deals well with medium size corpora but shows low precision when dealing with large corpora and can miss important features when dealing with a smaller dataset; (2) sentiment words and features can have longer dependencies that are not captured by the five-word window; and (3) ignoring more complex dependencies also ignores objective features (statements about the product without sentiment or opinion).

For dealing with these shortcomings, Kang and Zhou (2013) designed a new set of methods for feature extraction. They argued that extracting objective features could be an improvement upon previous methods, because as they mainly focused on subjective features (statement about the product with sentiment or opinion), they failed to capture important features in the form of customers statements rather than customers opinions. In other words, while a subjective statement or opinion involves a sentiment or judgement (¨…these headphones are extremely comfortable…¨), an objective statement merely describes something about the product (¨…they come in white, grey, and black colours…¨). According to the authors, these descriptions made by customers, are still features and valuable information worth capturing. In their methodology, for subjective feature extraction they use double propagation (Qiu et al. 2009) and include comparison patters (words that finish in ‘er/est’ and a list of manually gathered words) to extract more features. On the other hand, for extracting objective features the system identifies the following lexico-syntactic patters: NP+Verb+NP, where NP is a noun or noun phrase that represents a product feature; and PRP/Ex+Verb+NP, where PRP/Ex is a pronoun that refers to a product. These two structures allow the processing of phrases such as ¨*…the headphones come with a case…*¨ and ¨*…they come with a case…*¨, respectively. Finally, as extracting more features might lead to a lower precision, they implemented a three-step pruning on candidate features by using (1) TF-IDF, (2) a distance-based word similarity approach to deal with lexically redundant features such as ¨noise-cancelling¨, ¨noise cancelling¨ and misspellings, and lastly (3) through Wordnet´s similarity score to compare candidates to a set of predefined features previously extracted from the results of (Hu and Liu 2004). However, after their experiment the authors observed that while this method outperformed the results of Hu and Liu (2004), it also exhibited lower precision than the double-propagation method (Qiu et al. 2009). Thus, the authors concluded that while their method was able to extract more features than the latter, it could benefit from further pruning strategies that get rid of irrelevant features, and that a dependency relation-based feature extraction worked better than using only term frequencies.

In our work, we are addressing some of the limitations observed by Qiu et al. (2009) and Kang and Zhou (2013), by including a dependency relationship-based system that leverages complex dependencies for the extraction of both subjective and objective features. However, since our purpose is to stablish a sentiment baseline and compare with manually annotated opinions, we do not attempt double-propagation and base our analysis on a sentiment lexicon instead.

With the purpose of aiding designers to make data-driven decision, Ireland and Liu (2018) developed a framework for processing large amounts of qualitative data (OCRs) into quantitative data. The authors took a six-step approach to extract Feature-Sentiment Pairs (FSP) from the reviews of one Amazon product (a camping chair) and compared the machine-generated results to a human analysis. Like several of the previous methods, this framework starts with Part of Speech (POS) tagging to then identify and split the reviews into sentences. However, differing from prior approaches, it uses the opinion words (adjectives, adverbs, and verbs) to train a Naïve Bayes model to classify each word and sentence. For this, it considers one- and two-star reviews as negative, three-stars as neutral, and four- and five-stars as positive. The fifth step, generation of FSPs, works by pairing each sentiment word (adjective) with their respective features (nouns) within a sentence. In other words, if there are two nouns in a sentence the sentiment words will be paired with both. Still, they argued that their model ensured that even when incorrect pairs occurred, the small size of sentences guaranteed that the most important FSPs would be identified. In the last step, the framework determined the importance of FSPs by bringing words to their lemma (inflected form of a word, e.g., “better” and “best” become “good”), measuring WordNet´s similarity between words and assuming that those within two ¨nodes¨ of distance (in the algorithm´s tree-like structure) are similar. The algorithm then pairs similar nouns with similar sentiment words to create FSPs, and finally (4) counts frequencies of each FSP.

After conducting their experiment, they were able to extract valuable insights for designers, yet encountered several limitations through the process. For example, they observed that nearly one fifth of the sentences were disregarded for the FSP generation process, as the algorithm only identified the sentiment word but not to which feature it belonged. Thus, for improving the method and dealing with this issue they proposed to employ a distance-based approach similar to the one proposed by Hu and Liu (2004). Similarly, they observed several sentiment misclassifications as well as discrepancies among machine- and human-analysis, most likely due to not including negations into that step. Consequently, in this work we account for these negations when assessing the semantic orientation of sentences and product features. Furthermore, in this thesis we analyse ten products from the same category, based on the authors’ conclusion that a multi-product analysis might create more insightful information than a one-product approach.

Two of the most recent studies employ advanced techniques to extract insights from OCRs through very distinct approaches. While the contribution of Wang, Mo, and Tseng (2018) is their novel and fully automated system, Chen et al. (2019) offers a simplified and ready-to-deploy framework by leveraging a popular cloud-based system.

The methodology proposed by Wang, Mo, and Tseng (2018) attempts to automatically map CNs into DPs by training three consecutive classifiers. Based on whether a sentence contained information about product features and not, for example, regarding shipping or customer service, they manually annotated review phrases as relevant or irrelevant to train a first model that determined the importance of new sentences. Then, through statistical analysis they identified n-grams (sequence of n number of words) as keywords or CNs and used the most frequent ones as class labels to train a second classifier that mapped relevant sentences onto CNs (e.g., CN = “comfortable to hold”). Lastly, using products specifications from manufacturers' websites as labels (e.g., DP = Android system, weight = 178 g, CPU = Qilin 970, storage = 128 GB, RAM = 6 GB, …), they stored each combination of product description and CN to build the third classification model. From their experiment, they obtained accuracies of 86.02% in the sentence classification task, between 93% and 99% for the keyword extraction task but significantly lower (as low as 50% for DP = RAM) when automatically mapping CNs onto DPs. To the best of our knowledge, this is the first attempt at automating the process of translating CNs into DPs.

On the other hand, Chen et al. (2019) integrated a variety of AI and Machine Learning processes through Google Cloud Platform to offer a practical framework for non-data professionals. Some of the techniques employed in the system are Entity Analysis (a noun taking the form of a person, an organization, a product feature, etc.), Sentiment Analysis, Image Analysis, and so on. Combined, these methods summarize the information contained in OCRs, producing sentiment scores over product features as outcome to be leveraged by designers. In their experiment, the authors analysed eight services (customer service, shipping, warranty, etc.) from 5.000 Kindle reviews from Amazon. Although they do not present metrics for evaluating the performance of the feature extraction task, the visualization offered by the cloud-based service simplifies the interpretation of customers opinions. The framework presents average sentiment scores of each extracted feature which, in line with the authors' main goal, can prove useful for the adoption of these techniques by designer teams.

2.3. Crowdsourcing Design

Despite the lack of a unified vision over (1) what techniques to include in a framework and, from a product design perspective (2) what type of output is most useful, the purpose of this thesis is not to completely automate a product design process but to offer insights distilled from OCRs through NLP techniques that can inspire subsequent steps. This idea falls into the concept of crowdsourcing design, defined by Liu and Lu (2016) as *"…the process of accomplishing design by soliciting contributions from massive users upon the online community"*. This concept exploits the inherent value of customer knowledge within OCRs without pretending to replace (automate) the role of experienced product designers, even though it has been described that end users can sometimes outperform professionals in some design tasks (Evans et al. 2015).

Building on the work presented in this section and following the ideas behind crowdsourcing design, the techniques employed in this thesis for the analysis of OCRs are described in detail in a subsequent section. Similarly, even though we have already mentioned OCRs on several occasions, they are defined at length in the next section.

3. Data

This section is subdivided in four categories, (a) selecting OCRs as a data source, (b) the data collection process, (b) manual annotation of reviews, and (4) the pre-processing steps required before the analysis.

3.1. Selecting OCRs as Data Source

As a source for OCRs, we have selected Amazon.com for three main reasons. First, is one of the largest and more popular e-commerce platforms, offering a wide range of products and millions of OCRs. Secondly, is one of the most preferred data sources among publications in the text analytics field. Third and most importantly, in addition to the main text and a product rating (from 1 to 5) reviews from this website contain a helpfulness score (voted by other customers) and information regarding whether a review comes from a verified purchase. This additional information is imperative for this thesis, as (1) is our method for avoiding fake reviews purposefully published by manufacturers to mislead customers (Chen et al. 2019), and (2) based on the perceived helpfulness of OCRs as an important factor in purchasing decisions (Qi et al. 2016), we assume that these reviews contain valuable information for designers.

Our basic unit of analysis or OCRs are defined as peer-generated product evaluations posted on company or third-party web sites and are usually composed of a numerical rating (e.g., from 1 to 5) and an open-ended commentary about the product (Mudambi and Schuff 2010). They contain the consumer's perceived quality of the product and several authors have observed that they are a much more reliable, and a richer source of information, than traditional sources (Qi et al. 2016). In addition, OCRs are regarded as big data as they meet the standard definition of the three V's: high volume (i.e., a vast number of OCRs are available on the Internet), high variety (i.e., they differ greatly from each other and come from different sources) and high velocity (i.e., new CRs are continuously published every day). Additionally, Chen et al. (2019) described them as also possessing high veracity as they are unedited and crowdsourced from ordinary customers.

3.2. Data Collection

Since our purpose is to extract CNs from the OCRs of several products within the same category, two choices or data selections need to be made. First, we select articles within one product category to then select a fraction of their OCRs. First, we propose the category "External Headphones" from which we take the 9 most sold products between February 27 and March 4 of 2021. The status of top-selling product is offered by the e-commerce as the only measure of sales volume which often changes within a few weeks and offers no further information about its calculation. In other words, this list of products is very likely to change by the time this article is read. In Table 1 we show the 9 products selected for the analysis where we also include their position in the sales ranking and their product code.

Table 1. Top 9 Products in the "External Headphones" category

Brand	Model	Ranking	Product Code
Anker	Soundcore Life Q20	1	B07NM3RSRQ
Cowin	E7	2	B019U00D7K
iJoy	Matte	3	B01HNMTCE2
Audio Technica	ATH-M20X	4	B00HVLUR18
Zihnic	WH-816	5	B07K5214NZ
Sony	WH-1000XM4	6	B0863TXGM3
Sony	MDR-7506	7	B000AJIF4E
Mpow	059	8	B082TWSSTM
Bose	700	9	B07Q9MJKBV

On the other hand, the criteria for selecting OCRs from these 9 articles is based on Lead User Theory (Liu and Lu, 2016) which allows us to focus on the most promising reviews instead of mining every single one of them. To this end, we leverage the website's capabilities, where (1) OCRs are automatically sorted by Top Reviews (voted by peers as most helpful) and (2) "Verified Purchases" are automatically selected. Thus, we extracted 110 OCRs per product using the R programming language (R Core Team 2020) on RStudio version 4.0.2 (RStudio Team 2020). These were also used for all subsequent analyses.

Namely, a total of 990 product reviews were automatically extracted (or web-scraped) from Amazon.com using the R Library rvest (Wickham 2021) and the Google Chrome extension SelectorGadget, to select specific HTML code lines from each product site (Cantino 2013). However, as the OCRs ranged from April 1st, 2008, until March 4th, 2021, we included only OCRs published from 2017, leaving a total of 967 to be analysed. In the table below (Table 2) we included two sample reviews, but we have left out the publication date as it will not be used in this study.

Table 2. Examples of the sampled reviews

Title	Review
Sony I bow to You!	*I should say I'm blown away by this amazing headphone. I have owned many headphones, and I'm ready to say I have arrived! This headphone nails it in every category I care about: sound quality, battery life, bass, comfort, look, etc. The quality of music is already good in wireless mode, but if you want to see what this headphone can do pair it wired with a dedicated Daq (I'm using arcam-rpac). Great work Sony!*
Noise cancelling creates distant traffic sound	*Just got these and not sure I am going to keep them. One big feature I got these was for Noise Cancelling. But compared to my Bose QC35s these things are loud when canceling noise. It sounds like distant traffic. In comparison the QC35s are silent. Since I bought these to create a dead zone when I work or sleep, these may not work for me. They are definitely nicer fitting and look better than the QC35s. but are failing at a very key requirement for me.*

The OCRs in our dataset are quite short, almost always below 1500 terms, with an average length of only 224 words. The negative skewed shape of this distribution and the unimodal pattern of rating scores can be seen in the figures below.

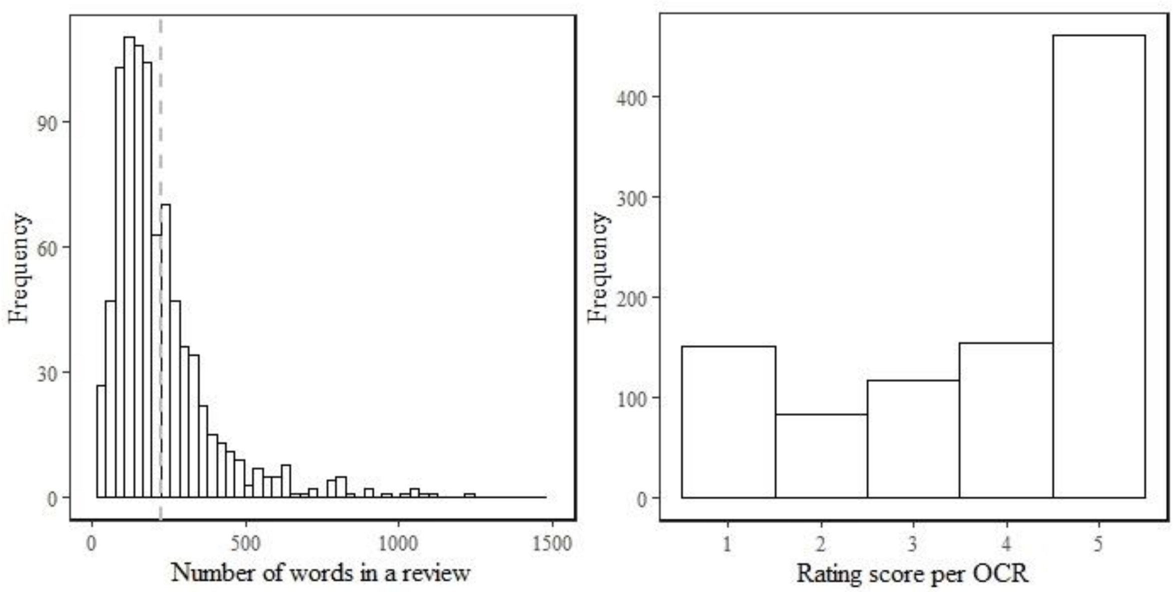

Figure 2: Frequency histograms of review length in number of words (left) and of rating scores per OCR (right).

3.3. Pre-processing

For the implementation of most Natural Language Processing (NLP) techniques, the text data requires transformation to a more structured format that permits analysis. However, as the pre-processing techniques vary depending on the task, choosing the appropriate one(s) is crucial. Thus, with the purpose of testing different approaches for the extraction of CNs (Section 4), we applied a distinct set of pre-processing steps for each case. However, they share some commonalities as in all cases we (1) brought the text to lower cases and (2) removed a customized list of terms containing brand names and models, as our focus is on product category.

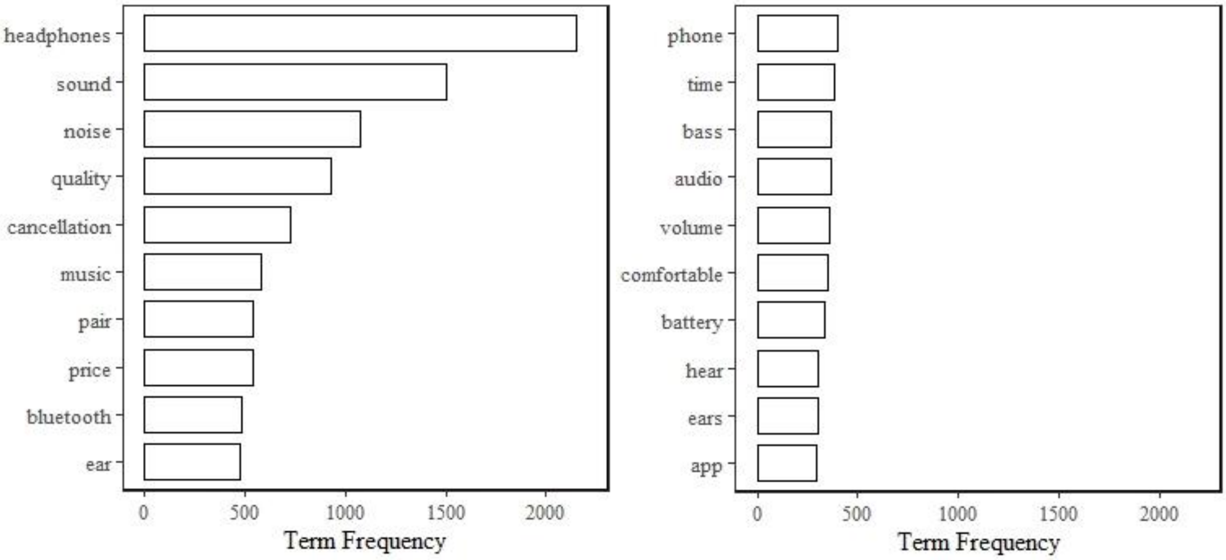

Figure 3: Frequencies of the 20 most occurring terms, from 1st to 10th (left) and 11th to 20th (right).

Moreover, after conducting a preliminary analysis of most frequent terms (Figure 3), we decided to (3) match often misspelled terms with the correct word: *cancelation* with *cancellation*, *canceling* with *cancelling*, and *noice* became *noise*.

Finally, as the title of an OCR can contain useful information (see examples in Table 2) we attached it as the first sentence of that review, increasing the total number of sentences of each OCRs by one (this is already accounted for in the figures in this section). The remainder pre-processing steps are detailed in the methodology chapter (section 4.1) as they are task dependent.

3.4. Manual Annotation

In this work, we define a CN in two different ways based on current research: (1) as extracted keywords (or Product Features), and (2) as a combination of a Product Feature (or Product Aspect) plus a Sentiment or Semantic Orientation towards that aspect. Thus, in order to assess the performance of the two tasks (Feature mining and Sentiment Analysis), we compare the systems outputs with manually extracted Customer Needs (CNs) from a smaller sample. This smaller dataset corresponds to the OCRs from the first page of each product, i.e., the most helpful. Namely, three annotations were made: (1) a list of features was created from each review, plus a (2) positive, negative, or neutral comment. Lastly, (3) the annotator was asked to group the features into more general product aspects, e.g., "*tactile controls*", "*dedicated buttons*" were grouped as "*controls*". The manual annotation was performed on 83 reviews, by a third party without prior domain knowledge of the product, and the output list contains her perceived Product Features and Sentiments from reading the reviews. Table 23 in the Appendix Section shows an example of the manual annotation process. As the smaller sample for annotation was not randomly selected, in the histograms included below, we see that the distribution of the length of reviews is less skewed, with a higher average (264 words) than the whole sample, and the distribution of the rating scores now shows a bimodal pattern.

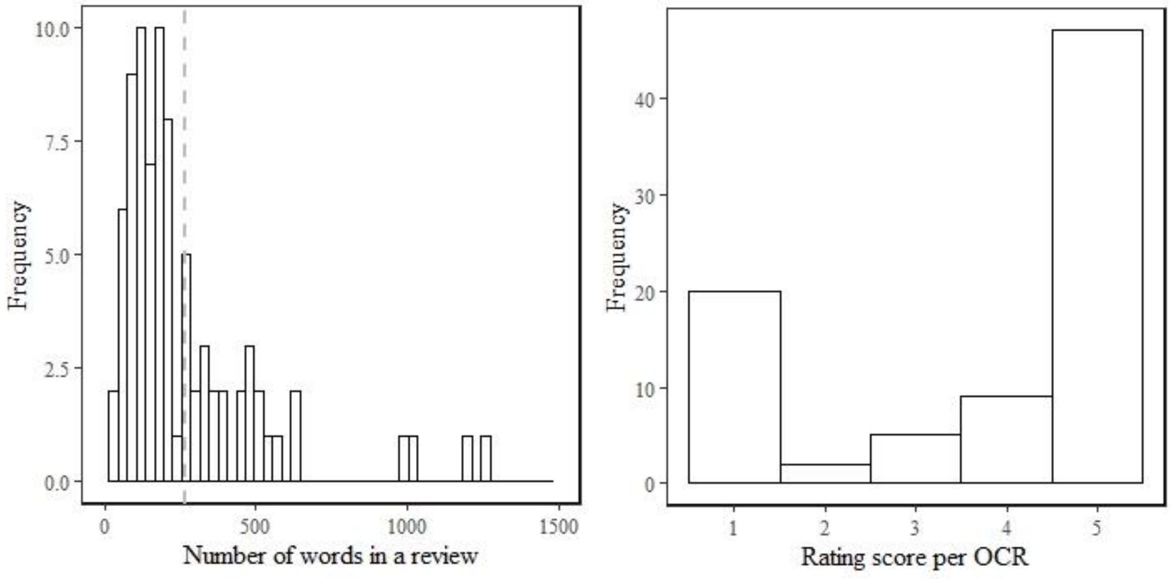

Figure 4: From the sampled data (*n* = 83), frequency histograms of review length in number of words (left) and of rating scores per OCR (right).

4. Methodology

As shown in our literature review, a wide array of systems have been proposed that encompass different techniques depending on the authors' assumptions and definition of customer need. Nevertheless, most of them divide the task into two main steps: the extraction or mining of product features, and an assessment of the semantic orientation or opinion concerning the extracted features. Thus, given that in this thesis we aim to find the most adequate system for extracting customer needs (CNs), we assess four different approaches for extracting product features or aspects, and two distinct tactics for sentiment analysis.

4.1. Product Feature Extraction

These three separate methods - or language models - are based on either (1) n-grams in a Bag of Words, (2) on Part of Speech (POS) tagging, or (3) on grammatical dependencies, although all of them follow a probability distribution over words or sequences of words,

$$P(w_1, w_2, w_3, \dots, w_n) \tag{1}$$

where the probability P of each term w_n is conditional to the word preceding it, as shown in the following equation,

$$P(w_1, w_2, w_3, \dots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2|w_1) \dots P(w_n|w_{n-1}, w_{n-2}, \dots, w_1) \tag{2}$$

however, computing these parameters is computationally expensive due to the high number of conditional probabilities $P(w_n|w_{n-1}, w_{n-2}, \dots, w_1)$ that need to be estimated. Hence, for each of the three approaches below we introduce intuitive ways to estimate them.

4.1.1. Bag of words (BoW)

This method treats every word - or combination of *n* number of words (n-grams) - as a unique feature in a text, disregarding the importance of word order (syntax), grammatical word types and dependencies. Nevertheless, this method is vastly used in text mining as is computationally inexpensive, offers a quick route for analysis, it is easier to interpret and has been characterized as effective for this particular task. Furthermore, with this methodology we follow the findings of Timoshenko and Hauser (2019) and the application of Wang, Mo, and Tseng (2018), where they defined CNs as keywords within the reviews. As a result, we characterize keywords as statistically relevant unigrams and bigrams.

Although there are several shapes in which text can be stored. For example, it can be stored as a (1) corpus which contains raw character vectors (i.e., strings), as a (2) Document Term Matrix (DTM) or, simply as a (3) table. In our work we choose the latter shape, a table with one-token-per-row, as it eases the subsequent implementation of sentiment analysis, and we rely on the R package tidytext (Silge and Robinson 2016). The same authors define token as *"a meaningful unit of text, most often a word, that we are interested in using for further analysis"* and tokenization is *"the process of splitting text into tokens"*.

In addition to the pre-processing steps mentioned in the previous section, the following steps are specifically applied here to allow proper implementation and analysis: removal of all punctuations and stop words from the SMART lexicon (Lewis et al. 2004), terms with digits are discarded as well as those with a letter appearing 3 or more times (suggesting slang and misspellings), and finally we apply lemmatization and exclude those tokens with a frequency lower than 5. The latter involves a lexicon of lemmas (Rinker 2018) which is used to convert inflected words to their root form, e.g., "cancelation" becomes "cancel" and "ears" become "ear".

The n-gram model is based on the assumption that the probability of a word depends only on the previous word, so after the data is cleaned and tokenized, we use Maximum Likelihood Estimation (MLE) to estimate the probability of a word w_i occurring in the text, which in the case of unigrams (no preceding word) is calculated as follows,

$$P(w_i) = \left(\frac{C(w_i)}{\sum_{k=1}^{n} w_k}\right) \tag{1}$$

where the estimated probability $P(w_i)$ is the ratio between the count C of term w_i and the sum of all terms in the document $\sum_{k=1}^{n} w_k$. Additionally, based on the work of (Hu and Liu 2004) we use the following threshold: for unigram w_i to be extracted it must have an estimated probability $P(w_i) \geq 0.01$. Similarly, for extracting useful bigrams or collocations we estimate the maximum likelihood of the probability of a bigram as follows,

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i)}{\sum_{k=1}^{n}(w_{i-1}, w_k)} \tag{2}$$

where the estimated probability of bigram $P(w_2|w_1)$ is equal to the ratio between the number of occurrences of a bigram, or $c(w_{i-1}, w_i)$, and the sum of all bigrams that start with term w_{i-1} or $\sum_{k=1}^{n}(w_i - 1,\ w_k)$. However, given that $\sum_{k=1}^{n}(w_i - 1,\ w_k)$ must be equal to the count of w_{i-1}, or $c(w_{i-1})$, the equation can be simplified as,

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})} \tag{3}$$

Additionally, as we need to extract only those bigrams that are potentially meaningful, in other words are not cooccurring by chance, we formulate the following hypotheses:

$$H_0\text{: } P(w_2|w_1) = P(w_2|\neg w_1) \tag{4}$$

$$H_1\text{: } P(w_2|w_1) \neq P(w_2|\neg w_1) \tag{5}$$

where H_0 states that the probability of w_2 occurring when w_1 precedes it is equal to the probability of w_2 occurring when w_1 does not precede it; and H_1 states that these probabilities are the same. Subsequently, to test these hypotheses we calculate the log likelihood ratio test as follows:

$$\log(\lambda) = \log\left(\frac{\mathrm{L}(H_0)}{\mathrm{L}(H_1)}\right) \tag{6}$$

where $\log(\lambda)$ is the ratio between the likelihood (L) of H_0 and the likelihood (L) of H_1. Lastly, to test association between words, a chi-squared test is applied to the results with a significance level of 0.05 and 1 degree of freedom. Thus, we extract only those bigrams with a p-value lower than the threshold, i.e., collocations.

4.1.2. Grammatical or Part of Speech tagging

For this approach, fewer pre-processing steps are applied to our data, as the order of words within a sentence is relevant for the analysis. In fact, in the first step a tagging software is implemented, in this case the R package UDPipe (Straka, Hajic, and Straková 2016), that reads the text and assigns parts of speech to each word (noun, verb, adjective, adverb, etc.), splits the reviews into sentences, and applies lemmatization. The program leverages Universal Dependencies (Nivre et al. 2020), which is a structure for reliable annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different languages.

This application also returns the data in a one-token-per-row format. In Table 3 we illustrate the most common parts of speech as described in the Universal Dependencies framework.

Table 3. Parts of speech and examples

Tag	Part of speech	Definition	Examples
ADJ	Adjective	Words that typically modify nouns and specify their properties or attributes	*Big, old, first, second, American*
ADV	Adverb	Words that typically modify verbs for such categories as time, place, direction, or manner. They may also modify adjectives and other adverbs, as in very briefly or arguably wrong.	*Very, well, exactly, up, down, never*
INTJ	Interjection	Word that is used most often as an exclamation or part of an exclamation.	*Pssst, ouch, hello, bravo*
NOUN	Noun	Words that denote a person, place, thing, animal, or idea	*Girl, cat, tree, air*
PROPN	Proper noun	a noun that is the name (or part of the name) of a specific individual, place, or object.	*Mary, John, NATO, HBO, Rotterdam*
VERB	Verb	Word that signals events and actions.	*Run, runs, running*

After employing POS tagging, the keyword mining process starts with the identification of one-word terms by applying the same calculations shown in equation (1) and selecting those with $P(w_i) \geq 0.01$, although in this scenario we only extract nouns ("*NOUN*") as product features. Next, for mining meaningful terms built of two or more words, we adhere to the trend observed in literature by moving beyond frequency-based approaches (such as word cooccurrences) in favour of the Pointwise Mutual Information (PMI) criterion for bi-grams, and the Rapid Automatic Keyword Extraction (RAKE) algorithm for longer n-grams (Rose et al. 2010). Both tests attempt to find words that together create a new concept and therefore might be relevant for a designer (e.g., the words "battery" and "life" versus "battery life").

The first of these methods is a statistical test that quantifies the likelihood of cooccurrence of two terms. In statistical terms, is a measure of independence or ratio between the (log) probability of two terms cooccurring $\mathrm{P}(w_1, w_2)$ and the product of the marginal probabilities for each term $\mathrm{P}(w_1)\mathrm{P}(w_2)$:

$$\mathrm{PMI}(w_1, w_2) = \log\left(\frac{\mathrm{P}(w_1,w_2)}{\mathrm{P}(w_1)\mathrm{P}(w_2)}\right) \quad (7)$$

Given that the result of this equation is a logarithm of the ratio, the interpretation is that when $\mathrm{PMI} = 0$ the two terms cooccur by chance, i.e., they carry no more meaning than when used independently. Conversely, when a result is positive, we are in presence of a statistically meaningful collocation. Thus, in this system we select only the bigrams with a value of $\mathrm{PMI} > 0$.

On the other hand, the RAKE algorithm was proposed to avoid the use of conditional probabilities over an arbitrary window of words and is based on the authors observation that keywords (words or phrases) rarely include stop words and punctuations (Rose et al. 2010). This method is included in our work for two reasons: first, it has been characterized as useful for extracting highly specific – i.e., domain specific – terminology; and secondly, by extracting keywords composed of combinations of nouns and adjectives (i.e., noun-noun is also a possible output), we can obtain both subjective and objective features.

The algorithm first uses a list of stop words (such as "the" and "of") and punctuations as delimiters for detecting and extracting relevant candidate keywords. Then, it creates a cooccurrence matrix and computes three scores for each word within a keyword: word frequency, word degree (cooccurrence with other words), and the degree to frequency ratio. Finally, the RAKE score of a keyword is the sum of these three values for each word within that keyword. Our implementation for this algorithm we select only keywords with a RAKE score of 2 or more.

4.1.3. Dependency Parsing

This approach is grounded on the work of Qiu et al. (2009) and Kang and Zhou (2013), where their dependency-based systems outperformed previous ones built on term frequencies. Dependency parsing is also enabled by the R library UDPipe, which in addition to grammatical labelling, tags each word within a sentence according to its syntactic relation with other words. The pre-trained model uses the 37 universal syntactic relation available in Universal Dependencies v2 (Nivre et al. 2020) that are constantly being updated since they were first introduced as the Universal Dependencies Treebank (McDonald et al. 2013). In Table 4 we include a few of these syntactic relations.

Table 4. Examples of dependencies and definitions

Tag	Dependency	Definition
nsubj	Nominal subject	*the syntactic subject and the proto-agent of a clause.*
obj	Object	*the second most core argument of a verb after the subject, denotes the entity acted upon.*
amod	Adjectival modifier	*adjectival phrase that serves to modify the noun (or pronoun)*
nmod	Nominal modifier	*relation used for nominal dependents of another noun or noun phrase and functionally corresponds to an attribute, or genitive complement.*
compound	Compound	*multiword expressions such as noun compounds (e.g., phone book).*
cop	Copula	*in English the verb to be is the only word that can appear with the cop relation*

For example, Figure 5 shows the output of analysing the phrase "*The overall build quality of these headphones is outstanding",* where "*build quality*" is recognized by the algorithm as the syntactic subject or the one performing the action, which in this case, is being outstanding. Moreover, it is also visible that the syntactic relation is captured regardless of the number of words (distance) between the two terms.

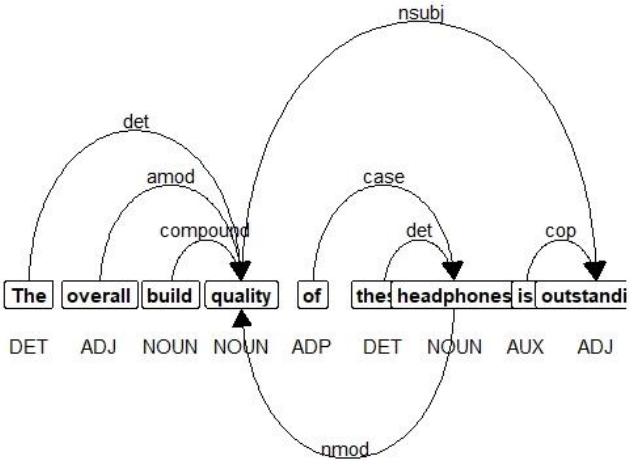

Figure 5: Part of Speech (POS) tagging and dependency parsing on a sentence extracted from the sampled data.

Therefore, the focus in this thesis is only on one of these syntactic relations: the subject or nominal agent of a clause (i.e., the "do-er"). We specifically look for the syntactic relations between subjects ("*nsubj*") and adjectives ("*ADJ*"), so that the output takes the shape of word pairs formed by a product aspect and an opinion. In other words, through this approach we extract feature-sentiment pairs (FSP), characterized by Ireland and Liu (2018), to enable a subsequent aspect-based sentiment analysis.

4.1.1. Evaluation

Precision and Recall, which are frequently employed in Information Retrieval and document classification studies, were chosen as assessment measures. In this work, if a keyword appears in the annotated list of features it is considered as a True Positive, and likewise, if a keyword extracted through a system does not appear in the list of true features, it is considered a False Positive. Next, we aggregate the mined keywords into product aspects following the annotated data as a reference, and if no feature falls into an aspect it is considered a False Negative.

		Reference	
		Event	Not Event
Predicted	Event	True Positives	False Positive
	Not Event	False Negative	True Negative

Figure 6. Confusion Matrix

We intuitively define Precision as the proportion of extracted features that are relevant, and calculate it as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{8}$$

Similarly, we characterize Recall (or true positive rate) as the fraction of aspects covered by the model to all relevant aspects, and compute it as follows:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{9}$$

In other words, Precision gives us a sense of the *correctness* of the extraction process, and Recall speaks of the *completeness* of the information extracted.

4.2. Sentiment Analysis

Opinion mining or Sentiment Analysis (SA) refers to a family of techniques that sit at the crossroads of statistics, natural language processing, and computational linguistics (Pang and Lee 2008). The aim of this discipline is to extract opinions from text written in natural language, or in other words, classify texts based on their semantic orientation by recognizing positive and negative expressions (Misuraca et al. 2020). This is referred to as the polarity or valence of a text and is assessed by assigning a polarity score of -1, 0, and +1 for negative, neutral, and positive terms, respectively (Liu, Hu, and Cheng 2005). Moreover, in one of the most common approaches for SA, the scores assigned depend on a lexicon of polarized terms, which can be created manually (e.g., the Bing sentiment lexicon) or created in a partial or fully automated way from the text (e.g., through double propagation). These polarized or subjective lexicons are lists words coupled with a specific emotion, for example terms such as “amazing” and “great” are labelled as positive, yet “awful” and “disgusting” are identified as negative.

Furthermore, as text is the basic unit of analysis it can be approached in three different levels: documents, sentences, and aspects. In other words, a document-level approach reflects the overall classification of a document as positive or negative. However, when a document is segmented into sentences, the amount of information extracted increases as we can observe the polarity of each sentence, as well as the overall orientation of the document (Tan et al. 2011). Likewise, aspect-based SA enables an even more detailed understanding, as it mines opinions regarding specific aspects of a written document. For example, customer reviews often consist of different opinions regarding each product or service feature: a guest at a hotel might leave a review with positive remarks about the cleaning service yet at the same time refer negatively about the food, personnel, location, etc. (Mohammad 2017).

With the purpose of capturing previously ignored nuances in natural language, many systems now incorporate a second lexicon that includes valence shifters or words that switch, intensify, or neutralise an opinion. For example, from the R library sentometrics (Ardia et al. 2020), a negator is a word that switches the valence of a term e.g., the word “*not*” completely changes the sentiment in the phrase “*I do like pie*” versus “*I do not like pie*”. An amplifier intensifies the original valence of another e.g., in “*I seriously do not like pie*” the word *seriously* increases an already negative valence, while a de-amplifier minimizes a positive sentiment in “*I barely like* *pie*". Additionally, the authors include a list of adversative conjunctions which trump the previous clause, e.g., "He's a nice guy but not too smart".

In this thesis, we try two different approaches for SA: on a sentence level, and on an aspect level. In both cases, we rely on the lexicon created by Hu and Liu (2004), which has been updated to contain a total of 6874 terms (Table 5). The labels included in this dictionary are manually annotated sentiment values of 1, 0, -1, -1.05 (for conditional phrases like "*could have*", "*should have*") and -2, where the latter indicates specific phrasings that, according to the authors, are always negative (e.g., "*too much fun*" and "*too much evil*" both denote negative even when the following word is positive and negative respectively). In addition, the lexicon includes frequently observed misspellings such as *"diappointed"*.

Table 5. Example terms from Bing sentiment lexicon

Term	Score
achievement, glory, inclusive, kudos, smart, trendy	1
abuse, despicable, disappointed, diappointed, lack, scratchy	-1
should have, could have, would have, would be	-1.05
too good, too many, too often, too much	-2

For the sentiment analyses, the lexicon was customized by removing the word "*cancellation*" (valence of -1), by setting the polarity of "*warm*", "*warmer*", "*warmly*", "*warmth*" and "*hot*" to -1, and by adding a list of new words that were deemed relevant for gaining a clearer understanding of customers opinions. These terms were selected from the reviews through the methodologies displayed in part 4.1. of this section and are shown in Table 6. Finally, our custom lexicon contains a total of 6891 terms and their polarity scores.

Table 6. Customized terms

Term	Score	Term	Score
tight	1	*low*	-1
mellow	1	*unimpressive*	-1
insulative	1	*non-existent*	-1
alright	1	*little*	-1
passable	1	*gone*	-1
ok	1	*heavy*	-1
high	1	warm	-1
potent	1	warmer	-1
long	1	warmly	-1
light	1	warmth	-1
customizable	1	hot	-1

Moreover, to address the limitations posed by Ireland and Liu (2018), we include a list of valence shifter in our analysis, taken from the R library sentometrics (Ardia et al. 2020). As mentioned above, the authors divide these terms into four classes: (1) negators, (2) amplifiers, (3) de-amplifiers and (4) adversative conjunctions; and include correct and incorrect spellings, much as the Bing lexicon. A few examples of these terms are visible in Table 7.

Table 7. Example valence shifters from sentometrics

Term	Score	Class
ain't, couldnt, haven't, mustnt, not, neither, never, no, shouldn't, wont	-1	1
absolutely, certain, deeply, enormously, heavy, massively, sure, very	1.8	2
almost, least, kinda, seldom, sorta, sparsely	0.2	3
although, but, however, whereas	-1	4

4.2.1. Implementation

In our first approach to create a summary of customers opinions toward product features, we use the features mined through every extraction system (from the previous subsection) as keywords to match sentences within our dataset. This approach resembles the implementation done by Hu and Liu (2004) where they counted the number positive and negative reviews for each product feature, yet in our work we will use the number of sentences instead. Therefore, the mined features are matched to the sentences where they are mentioned and then we obtain the polarity score of each phrase as a summary of the total number of positive, neutral and negative sentences per aspect. This approach receives the name of Sentence Based Sentiment Analysis (SBSA). Secondly, we leverage the Feature Sentiment Pairs mined through dependency parsing to perform an Aspect Based Sentiment Analysis (ABSA). To this end, we first define the sentiment of each FSP (i.e., the polarity of an adjective used to describe a nominal subject) and create aspect-level summary using the same aspects created in the manual annotation process.

The assessment of these two techniques is done qualitatively by interpreting summary visualizations created with the R package ggplot2 (Wickham 2016), and through statistical testing (Chi-squared and Fisher´s test) to compare the distribution of manually annotated sentiments (gold standard) to ABSA and SBSA, separately. To correct for multiple testing, we apply Bonferroni correction (Haynes 2013), so our significant cut-off is set as α/n, where n is the total number of tests (in our experiment, $n = 10$) and the initial $\alpha = 0.05$.

5. Results

In this section we report the results of our analyses in two parts. In Section 5.1, we display the list of features manually annotated to use as gold standard. Next, we present the extracted features using each of the methods proposed in Section 4 and assess their outputs qualitatively and by providing evaluation metrics. In Section 5.3 we assess the results of our sentiment analyses, and lastly, we present the results obtained when analysing the whole dataset.

5.1. Feature Extraction

5.1.1. Manually Generated Features

From the total of 83 reviews sampled for manual annotation, two contained only general comments about the product and therefore no features could be extracted from them. Table 8 shows the raw output of the manual annotation process.

Table 8. All manually annotated features

unigrams		n-grams	
volume	usability	noise cancellation	charge and listen
case	comfort	sound quality	charge speed
earcups	style	battery life	charging cable
mic	wireless	build quality	usb c
buttons	pairing	call quality	usb charger
feel	fit	bluetooth	micro usb
treble	bass	connect app	carry bag
aesthetics	storage	touch controls	ear pads
soundstage	padding	customer service	memory foam
cable	design	media controls	noise cancellation levels
radio	weight	voice assistance	normal mode
durability	build	voice prompts	audio quality
equalizer	packaging	bass boost	auxiliary cable
hinges	sound	ear cushions	travel bag
mids	price	audio quality	charge and listen
highs		auxiliary cable	charge speed
anc		charging cable	bass boost
range		micro usb	ear cushions

As customers can use several terms for referring to a product feature, e.g., *"ear pads"* and *"ear cushions",* the annotator was asked to group these keywords into more general product aspects. For example, “*vocals*”, “*mids*”, “*clarity*”, “*soundstage*” and “*bass*” were aggregated into “*sound features*”, and similarly, “*volume controls*”, “*media controls*” and “*touch controls*” became “*controls*”. As a result, the 69 different product features were grouped into 28 product aspects that we also leveraged for assess our methods. These broader categories are displayed in Table 9.

Table 9. Manually annotated aspects

accessories	durability	range
battery life	ear pads	sound
build quality	equalization	sound features
call quality	fit	storage
charge	modes	style
comfort	noise cancellation	usability
connectivity	notifications	voice assistance
controls	packaging	weight
customer service	phone app	
design	price	

5.1.2. Automatic Feature Extraction

The table below shows the outcome of our first approach based on a Bag of Words model. In total, 9 unigrams and 14 collocations were extracted with this methodology.

Table 10. Bag of words output

unigrams		Collocations	
sound	music	audio quality	memory foam
noise	battery	battery life	noise cancel
quality		build quality	phone call
pair		cancel feature	sound quality
cancel		connect app	touch control
ear		ear cup	voice assistant
time		highly recommend	volume control

Before evaluating each system through classification metrics, we assess their outputs qualitatively, in terms of ease of interpretation.

Namely, when we compare this first output to the results of Table 10, it is possible to observe that the system successfully captures features such as “*battery life*”, “*build quality*”, “*connect* *app*”, “*touch controls*”, “*sound quality*”, “*voice assistance*” and “*volume controls*”. Contrarily, some features are not mined in a literal way, e.g., the manually extracted “*noise cancellation*” appears as “*noise*”, “*cancel*” (lemma of “*cancellation*”) and “*noise cancel*”, similarly, the “*pairing*” feature is extracted as “*pair*” (lemma of “*pairing*”); and “*ear cushions*” is mined as “*ear cup*” (lemma of “*cups*”). Moreover, some of these keywords may require some degree of interpretation from a designer, for example, when our system obtains “*phone call*” it might be that a customer was referring to “*call quality*”, as observed in the following excerpt from an annotated review: *“...Clarity of phone calls - both input and output (i.e., the mic works well) - is excellent…”*. However, it might mean something different as this extract shows “*…After six support phone calls and a live chat session, Sony did not offer to replace my unit…*”. Likewise, although more straightforward, “*cancel feature*” could be interpreted as the noise cancelling feature. Nevertheless, certain features, e.g., “*bass boost*” and “*customer service*”, were not captured by this automatic approach.

As mentioned in Section 4.1.2 for the second framework we leveraged POS tagging. Here we first extracted nouns as single-word features, and then implemented two different techniques for mining relevant collocations in the shape of noun + adjective. The outcomes of these two tactics (PMI and RAKE) are listed in Table 11.

Table 11. Part of Speech tagging (POS) output

Nouns	Collocations (PMI)		Collocations (RAKE)	
sound	memory foam	bass boost	customer support	battery life
quality	dedicated buttons	touch control	noise cancellation	ear cup
noise	usb c	long flight	ambient noise	volume adjustment
cancellation	customer service	battery life	long period	custom button
time	pro con	minute charge	customer service	big deal
pair	aux cable	ear cup	major flaw	rap songs
battery	default setting	second device	voice assistant	return window
music	customer support	high end	default setting	few hour
phone	voice assistant	last year	huge deal	audio quality
bass	voice prompt	price point	poor quality	build quality
	year old	noise cancellation	ambient sound	overall sound
			minor issue	bass boost
			long flight	

Correspondingly, through grammatical tagging the system automatically mined 12 single-word features, 22 collocations through Pointwise Mutual Information (PMI) and 25 when we implemented the Rapid Automatic Keyword Extraction (RAKE) algorithm. When we compare these results with those from the previous approach, we observe that several previously ignored features are now captured, e.g., "*bass*", "*bass boost*", "*customer service*", "*voice prompts*", "*price point*", among others. Furthermore, both PMI and RAKE were able to mine the following features: "*customer support*", "*customer service*", "*battery life*" and "*noise cancelling*". However, their similarities end there as RAKE failed to detect any keywords related to price, voice prompts, and cables/cords, and PMI did not mine any sound or audio-related aspects. Nonetheless, the latter system extracted the implicit feature "long flight" which, based on the source reviews, can be understood as a need for either enduring comfort or effective noise cancellation over a long flight: *"…The cushions are roomy and comfortable, and I expect to be able to wear them for a longer period of time (like on a long flight)…"* and *"…this is great, as I am looking to cancel out midrange and bass sounds like an air conditioner, nearby dogs, and the wearing rumble of airplane engines during long flights...".* Lastly, some manually annotated features, such as "*storage*" or "*travel bag*" and "*connect app*" were not caught by either of these methods.

In our last approach, when extracting Feature Sentiment Pair (FSP) through dependency parsing, we noticed that many FSPs appear only once in the whole data. This is due to customers using different words when referring to a product feature (e.g., a customer might use "*anc*" instead of active noise cancelation), and often describe them through an array of adjectives (e.g., "*good*", "*great*", "*amazing*", etc.). Therefore, in Table 11 we display part of the aggregated results. The complete list can be found in Table 19 in the Appendix.

Table 11. Dependency parsing output

Feature (noun)	Sentiment or Opinion (adjective)
app	*solid, customizable, buggy*
bass	*tight, good, neat, solid, potent, apparent, strong*
battery	*amazing, great, unprecedented, same, low, dead, awesome, great, good, respectable*
bluetooth, connection, connectivity	*easy, good, gone, problematic, stable, great*
anc, cancellation	*alright, decent, good, greatest, great, excellent, outstanding*
cable, cord	*great, little*
color, design, frame	*beautiful, sleek, attractive*
body, build, construction	*durable, better, solid*
sound, soundstage	*smooth, great, good, impressive, awesome, unimpressive, wonderful, comparable*
fit, fits	*perfect, ok, fine*
cushion, earpad, pad, pads	*comfy, soft, comfortable, roomy, wide*
case	*slimmer, better*

Dependency parsing captures features that remained hidden to previous attempts, for example, the system extracted FSPs related to "*case*" which refers to the annotated features "*storage*" and "*travel bag*" (e.g., "*…The case is slimmer than Sony's, adequate for the airplane seat pocket...")*; *"fit"* which can be used to described both head fit and size (e.g., *"…They fit comfortably over the ears (they adjust and bend easily for folding) and the sound is awesome!..."* and *"…The case fits in my work bag just fine…"*), and "design" (e.g., *"…The design is sleek, the sound is great, and the noise canceling is excellent…"*). Furthermore, the FSPs displayed above require less interpretation as, for example, *"comfy earpad"* and *"smooth sound"* are easier to understand. Additionally, this strategy was also able to capture implicit features, for example, while an FSP such as "*easy bluetooth*" mentions bluetooth capabilities, the word easy implies that customers value the feature's simplicity to use: *"…Bluetooth is easy to connect and has a far distance...".*

5.1.3. Method Evaluation

First, we assess the precision of our systems for identifying relevant keywords or product features by using the manually generated feature list as our gold standard. The results shown in Table 12, evidence that the approaches based on Bag of Words (BoW) and Dependency Parsing are the best and worst performing methods, respectively. Specifically, the former mined a total of 23 keywords yet only 20 were true features, and the latter extracted 152 features but 49 were irrelevant.

Table 12. Assessment of feature extraction systems (Feature level)

	True Positives	False Positives	Precision
BoW	**20**	**3**	**0.87**
PMI	25	7	0.78
RAKE	26	9	0.74
DP	103	49	0.68

However, as mentioned above, the terminology used by customers to describe products features often varies and several of these True Positive features can pertain to the same product aspect. For example, if we look once again to the results of our BoW approach (Table 10), four of the 23 keywords/features referred to the same aspect ("*noise*", "*cancel*", "*noise cancel*" and "*cancel feature*" refer to the *noise cancellation* aspect). As a result, a model with high precision is not necessarily the most insightful for a designer team and a subsequent analysis is needed to better assess the models.

Consequently, to obtain the Recall of our four approaches we leveraged the 28 (manually annotated) product aspects and aggregated the automatically mined features accordingly. Thus, each keyword extracted through our methods was manually assigned to one or more of these aspects. To use the example from above, the keywords "*noise*", "*cancel*", "*noise cancel*" and "*cancel feature*" were grouped as the *noise cancellation* aspect. In Table 13, we outlined which product aspects were effectively captured by each of the four systems. Additionally, in Table 14 we summarized the total number of aspects covered by the features mined through each approach and display their Recall.

Table 13. Aspect captured by each system

Manual Aspect	BoW	PMI	RAKE	DP	Manual Aspect	BoW	PMI	RAKE	DP
modes	0	1	0	0	accessories	0	1	0	1
noise cancellation	1	1	1	1	battery life	1	1	1	1
notifications	0	1	0	0	build quality	1	0	1	1
packaging	0	0	0	1	call quality	1	1	1	0
phone app	1	1	1	1	charge	0	1	0	1
price	0	1	0	0	comfort	1	1	1	1
range	0	0	0	1	connectivity	1	1	1	1
sound	1	1	1	1	controls	1	1	1	1
sound features	1	1	1	1	customer service	0	1	1	0
storage	0	0	0	1	design	0	0	0	1
style	0	0	0	1	durability	0	0	0	1
usability	0	1	1	1	ear pads	1	1	1	1
voice assistance	1	1	1	1	equalization	0	0	0	0
weight	0	0	0	1	fit	0	0	0	1

From the 28 aspects, 10 were covered by all systems and only one was completely ignored. Dependency Parsing was the only system that captured features related to the product *packaging*, *weight*, *style*, *range* (of bluetooth) and *fit*, and only the PMI-based approach obtained keywords associated to *notifications* and *modes*. Moreover, when comparing Table 13 to the results shown below (Table 14) we observe that while Bag of Words (BoW) obtained the highest Precision (0.87), it only mined enough features to cover 12 product aspects and thus obtaining a Recall of 0.43 (i.e., most of the mined keywords were true features, yet several were redundant). On the other hand, while many of the features captured through Dependency Parsing (DP) were false positives (Precision = 0.68), the true features contained information concerning 22 of the 28 product aspects (Recall = 0.79).

Table 14. Assessment of Feature extraction system (Aspect level)

	True Positives	False Negatives	Recall
BoW	12	16	0.43
PMI	18	10	0.64
RAKE	14	14	0.50
DP	**22**	**6**	**0.79**

5.2. Sentiment Analysis

With the purpose of mining customers opinions regarding specific aspects of the products, we performed Sentiment Analysis (SA) on the 10 aspects that were captured by all the feature extraction systems. However, as Dependency Parsing mined more features than any other method, we used these FSPs in this part of the study. These keywords, and the aspects that contain them, are displayed in Table 15.

Table 15. Aspects covered in the Sentiment Analysis.

Aspect	Dependency Parsing Features (Nominal Subjects)
Battery life	*"battery", "life", "data", "transfer"*
Comfort	*"comfortable", "cushion", "cushions", "soft", "comfy", "feel", "cushioning", "roomy", "wide"*
Connectivity	*"bluetooth", "pairing", "wireless", "pair", "connectivity", "connection"*
Controls	*"command", "switch", "buttons", "media", "controls", "touch", "tactile", "volume"*
Earpads	*"ear", "pads", "cups", "pad", "cushion", "memory", "foam"*
Noise Cancellation	*"ambiance", "ambient", "amplification", "anc", "noise", "cancellation"*
Phone App	*"connect", "app"*
Sound	*"audio", "sound", "sibilance"*
Sound Features	*"bass", "buffer", "clarity", "highs", "lows", "latency", "mids", "soundstage", "sub-bass", "treble", "vocals"*
Voice Assistance	*"voice", "assistance"*

The features above are leveraged in two different ways: first, we used them to match the sentences where they are mentioned, enabling us to perform a Sentence Based SA; and secondly, we use them as Feature Sentiment Pairs, obtain the polarity of each adjective used and finally group them by aspect to obtain a summary for each one. In Figures 7, 8 and 9, we display the distribution of polarity scores for each aspect, as obtained from the manual annotation process and from the two Sentiment Analyses (the summary data of these distributions, used to build these visualizations is listed on Tables 20, 21 and 22 in the Appendix section).

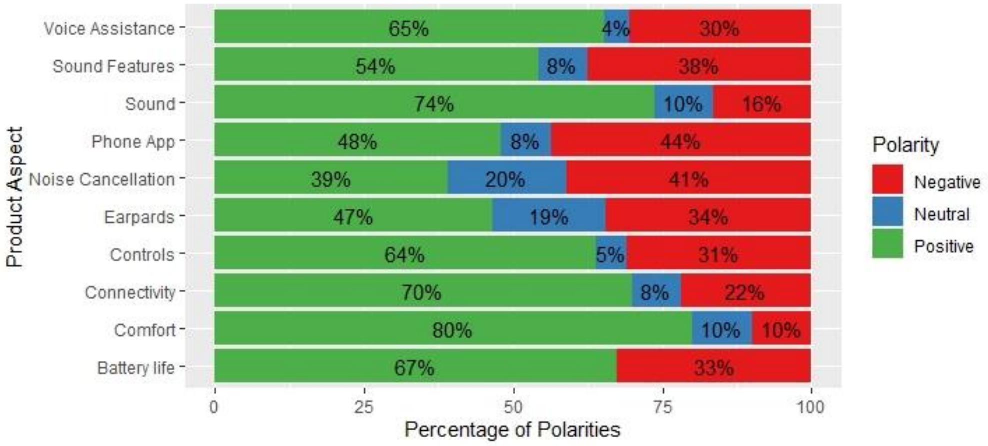

Figure 7. Distribution of Sentence-based polarities per product aspect.

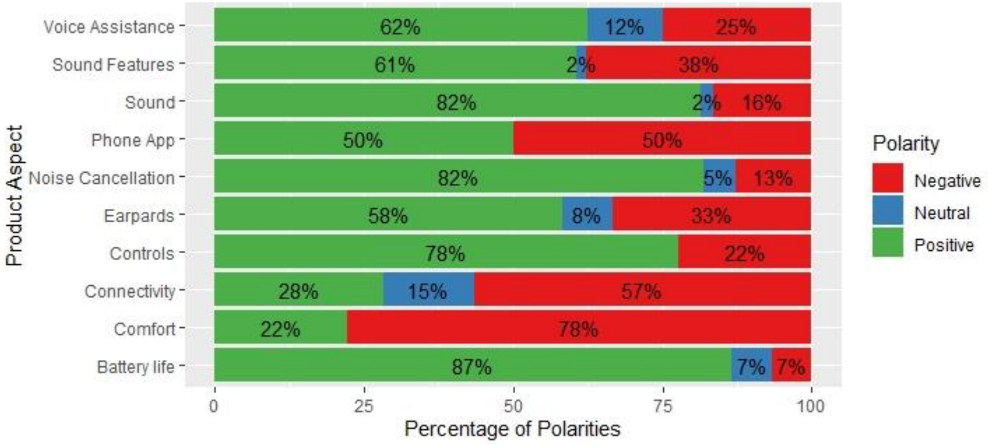

Figure 8. Distribution of Manually annotated polarities per product aspect.

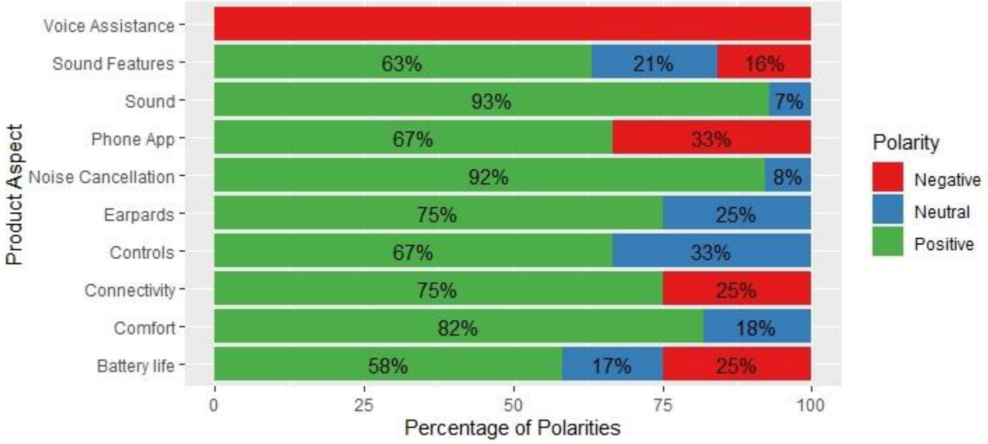

Figure 9. Distribution of Aspect-based polarities.

5.2.1. Sentiment Analysis Evaluation

In broad terms, though the figures above we can assess that the results from the Sentence-Based SA (SBSA) mainly differ in three aspects, *Noise Cancellation*, *Connectivity* and *Comfort*, versus our gold standard. Similarly, some discrepancies appeared among the gold standard and the output from Aspect-Based SA (ABSA), which seems to overestimate the rate of Positive polarities in two of these aspects. Nonetheless, to test if these differences are significant, we performed statistical tests on the proportions of Positive polarities (from the distributions displayed above) in each aspect *vs* the gold standard. These results are displayed below.

Table 16. Chi-squared test for Positive proportions from SBSA vs Manual Polarities

Aspect	Chi-squared	DF	p-value	Estimated Prop. 1	Estimated Prop. 2	99.5% Conf. Interval (Prop. 1 - Prop. 2)
Battery life	2.644	1	0.104	0.674	0.867	(-0.48, 0.09)
Comfort	**35.068**	**1**	**0.000**	**0.800**	**0.222**	**(0.340, 0.816)**
Connectivity	**16.152**	**1**	**0.000**	**0.642**	**0.283**	**(0.123, 0.596)**
Controls	1.740	1	0.187	0.638	0.778	(-0.408, 0.128)
Earpads	0.182	1	0.670	0.466	0.583	(-0.608, 0.372)
Noise Cancellation	**24.132**	**1**	**0.000**	**0.389**	**0.818**	**(-0.646, -0.211)**
Phone App	0.000	1	1.000	0.467	0.500	(-0.467, 0.426)
Sound	1.806	1	0.179	0.736	0.816	(-0.229, 0.072)
Sound Features	0.235	1	0.627	0.541	0.606	(-0.351, 0.221)
Voice Assistance	0.000	1	1.000	0.652	0.623	(-0.555, 0.609)

The results of the Chi-squared test confirm that the few significant differences (p-value < 0.005) among proportions of positive sentiments, between SBSA and gold standard, appear in *Comfort*, *Connectivity* and *Noise Cancellation* aspects. In other words, the ratio of positiveness concerning other aspects did not differ significantly from the annotated opinions.

Table 17. Fisher's test for Positive proportions from ABSA vs Manual Polarities

Aspect	Estimated Prop. 1	Estimated Prop. 2	p-value	99.5% Conf. Interval (Prop. 1 - Prop. 2)
Battery life	0.583	0.866	0.115	(-0.154, 0.718)
Comfort	**0.769**	**0.222**	**0.001**	**(-0.833, -0.076)**
Connectivity	0.750	0.283	0.032	(-0.794, 0.106)
Controls	0.714	0.777	1.000	(-0.331, 0.639)
Earpads	0.875	0.583	0.374	(-0.761, 0.390)
Noise Cancellation	0.923	0.818	0.652	(-0.331, 0.325)
Phone App	0.667	0.500	1.000	(-0.769, 0.661)
Sound	0.928	0.815	0.527	(-0.285, 0.286)
Sound Features	0.611	0.606	1.000	(-0.354, 0.387)
Voice Assistance	0.000	0.625	0.889	(-0.716, 0.962)

Similarly, we assessed the ABSA output *vs* the manually annotated polarities. However, as the number of Feature Sentiment Pairs per aspect was lower than the number of sentences in SBSA, we used the Fisher's exact test (Fay and Fay 2020). In the table above, we observe that the only significantly different proportion of positive scores (p-value < 0.005) appeared on the *Comfort* aspect. In other words, ABSA's polarity estimations for the other aspects did not differ greatly from the gold standard.

The assessment of these sentiments analyses suggests that they both represent viable methods for opinion mining on this data. Surprisingly, the statistical tests found no significant differences in the ratio of positive scores obtained through SBSA in 7 out of 10 aspects, and in 9 out of 10 for ABSA, when compared to the gold standard. Therefore, based on the better performance of ABSA and the higher number of features captured through Dependency Parsing (as Feature Sentiment Pairs), in the following subsection we show the outcome of applying both techniques on the rest of our complete dataset (n = 884).

5.3. Implementation in complete dataset

Table 18 shows the results from the implementation of Dependency parsing-based feature mining. These outputs highlight that by increasing the number of reviews, the 1130 unique FSPs extracted now covered 27 of the 28 manually generated aspects. For example, the previously ignored "price" aspect now contains 20 unique FSPs, such as: "good price", "right price", "low price", "fantastic price", etc. However, while we consider more important aspects those that are more frequently extracted (i.e., more frequently mentioned)*,* there are still 11 that show a less than 10 unique Feature Sentiment Pairs. For example, *style* contains "*good finish", "nice finish", "beautiful finish", "slick appearance", "great appearance", "cute color",* and *"silver color".* Lastly, in the Figure 10 we show the final distribution of polarities (positive, neutral, or negative) as the final output for a designer team.

Table 18. Dependency Parsing output (FSPs)

Aspect	Mined Features (nouns)	Unique FSP
Sound Features	*"bass", "volume", "highs", "mids", "base", "midrange", "boost", "frequencies", "latency", "treble", "vocals", "booster"*	78
Sound	*"sound”, "audio", "music", "sounds", "issue", "one", "popping", "speaker", "speakers"*	76
Noise Cancellation	*"cancellation”, "anc", "noise", "isolation", "cancelling", "nc", "cancel"*	64
Earpads	*"cups", "pads", "ear", "padding", "earpieces", "cup", "ears", "earpads", "earcups", "foam", "muffs", "seal"*	51
Usability	*"instructions", "update", "codec", "function", "functions”, “feature", "features", "manual", "settings", "customization", "functionality", "setting", "setup"*	51
Connectivity	*"connection", "pair", "bluetooth", "connectivity", "pairing", "connections", "signal", "antenna", "connect"*	46
Controls	*"controls", "buttons", "control", "button"*	35
Battery Life	*"battery", "batteries"*	28
Call quality	*"microphone", "calls", "mic", "voice", "speak", "microphones"*	28
Comfort	*"cushions", "comfort", "cushion", "headband", "headbands", "cushioning"*	22
Price	*"price", "cost"*	20
Accessories	*"cord”, "cable", "wire", "cables"*	19
Phone app	*"app", "interface"*	17
Range	*"range”, "distance"*	15
Modes	*"mode", "level", "levels"*	14
Build quality	*"construction", "force", "material", "build"*	13
Design	*"design", "dimensions", "frame", "mounts"*	11
Customer service	*"service", "support"*	8
fit	*"fit"*	7
Storage	*"case", "pouch"*	7
Style	*"appearance", "finish", "color"*	7
Equalization	*"equalizer"*	4
Packaging	*"packaging", "wrapping"*	4
Charge	*"charging"*	3
Durability	*"plastic", "months"*	3
Weight	*"weight"*	2
Voice Assistance	*"assistance"*	1

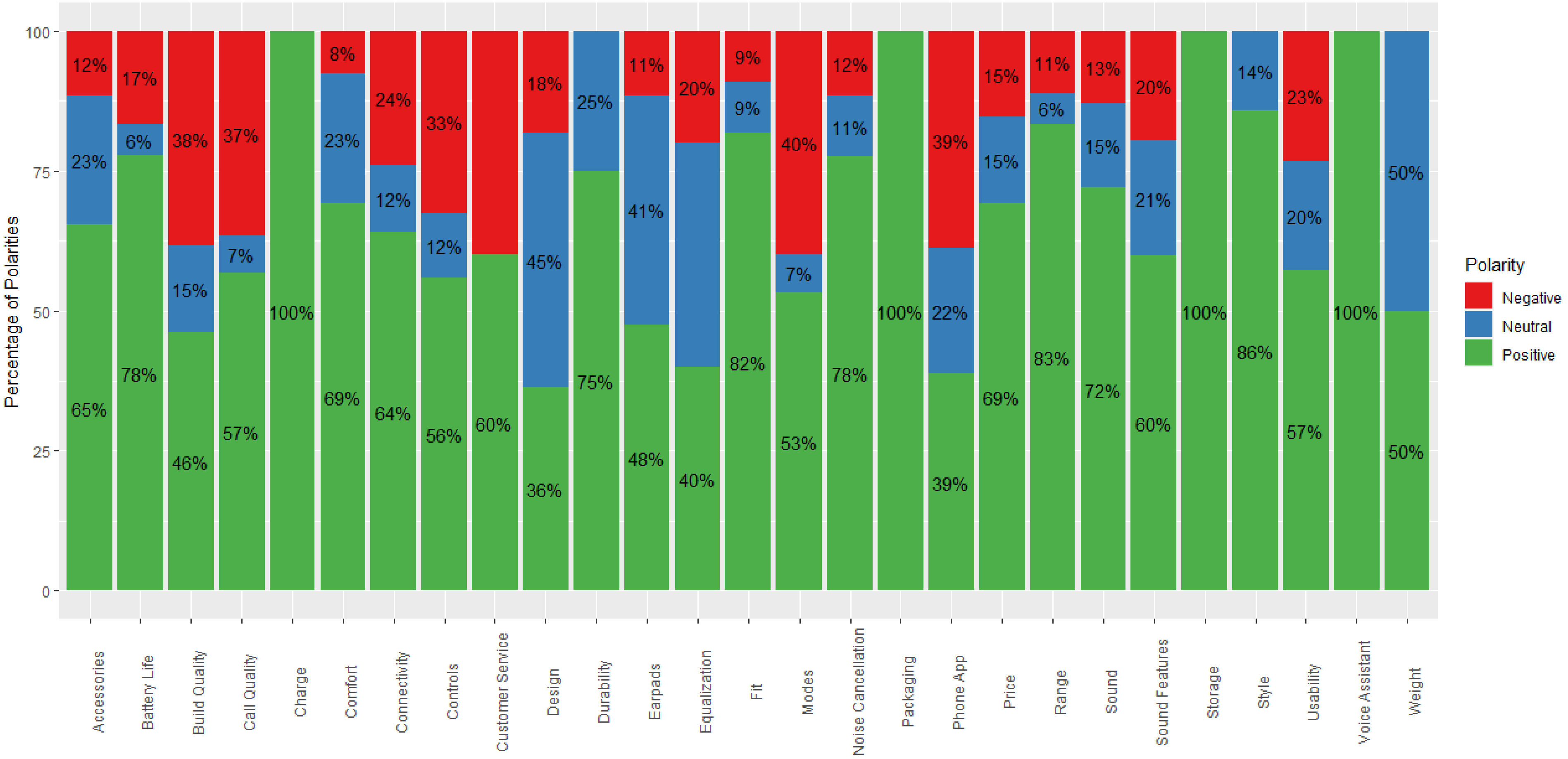

Figure 10. Aspect Base Sentiment Analysis in the complete dataset

6. Discussion and Conclusions

Following our research question, in our work we tested several approaches to create a summary of customer needs in the shape of product features and opinions. Based on our results, the system that leveraged dependency parsing (DP) was assessed as the most effective method for our particular application (both qualitatively and in statistically terms) which is consistent with the results observed by Kang and Zhou (2013). Furthermore, with the purpose of capturing customers opinions on the mined features, we tested two types of Sentiment Analyses (SA) with a customized lexicon and were surprised to see that both showed little disparity versus the human annotated data. Nonetheless, in line with the results reported by Zhou and Song (2015), our Aspect Level SA performed slightly better than the Sentence-based.

In this work, we addressed several of the limitations posed in previous research, e.g., by including valence shifters, leveraging grammatical dependencies, customizing our sentiment lexicon to our specific domain, and most importantly, by performing our analysis on a product category instead of a single product with the corresponding manual labelling of features and aspects. However, a few limitations still remain, for example (1) pruning irrelevant keywords/FSPs and (2) including more complex syntactic relations in the feature mining process. First, a better pruning methodology, such as the one proposed by Kang and Zhou (2013) where they leverage textual and semantic similarity, could allow us to improve the quality of the insights created by our system by automatically filtering uninteresting, irrelevant, or redundant keywords (which should increase the Precision of our final system). Secondly, even though our results allow for a straightforward interpretation (i.e., a feature plus an adjective, such as "*good battery*"), Feature Sentiment Pairs (FSPs) are comprised of only two words which offers a very narrow view of CNs without much detail (Ireland and Liu 2018). In other words, through our framework a designer could easily capture that the *phone app* aspect is mostly associated with negative opinions but would need more research to elucidate causality (i.e., how to improve said product aspect).

Furthermore, we are aware that our final approach has limitations of its own and we identify two main areas for improvement: (1) human annotation and (2) automation. First, several issues have been observed in literature regarding the methodology we followed for manual annotation (Hu and Liu 2004). For example, a person might not process all the data or forget to tabulate information when organizing by frequencies or carry personal biases by actively searching for features and sentiments that they believe more important while disregarding the less important ones (Ireland and Liu 2018). From our work, we argue that some product aspects could have been clustered differently, for example, "*style*" and "*design*" or "*build quality*" and "*durability*" might have been combined by a different annotator due to their semantic similarity. Similarly, given that none of the systems covered the *equalization* aspect, we argue that its labelling might need reviewing. Nonetheless, a way to deal with these issues (although not completely) is to have several annotators separately label the data and, when an issue arises, discuss and deliberate on it to then display their level of agreement through measures like Inter-annotator agreement (IAA) (Artstein 2017). On the other hand, while in our systems automated the processes of feature extraction and sentiment analysis, the middle step of our Aspect Based Sentiment Analysis, i.e., the aggregation of mined features into broader product aspects, was done manually. Even though this was a conscious decision to keep the focus on testing the former steps, it also means that before an implementation in a practical setting (e.g., a firm) the feature aggregation process should be automated with, for example, topic modelling through Latent Dirichlet Allocation (Blei et al., 2003). We argue that this unsupervised machine learning methodology, which stems from traditional clustering methods (Reisenbichler and Reutterer 2019), represents a natural next step from this research.

Another important point to highlight, is that although Lead User Theory allowed us to use a relatively small dataset, other authors claim that by leveraging a bigger sample of reviews the insightfulness of the information extracted increases (Ireland and Liu 2018). Although, to the best of our knowledge there is no agreement on the number of reviews needed to produce quality insights, in fact, this number ranges from 50 to tens of thousands in similar studies. Given that the number of aspects covered increased from 22 to 27 when we increased the number of reviews analysed from 83 to 884, we argue that this improvement in the *completeness* of the insights produced, might be due to the size of the data. Regardless, we also argue that even if using every single review of a product, the data would not be representative of all customers (as not all of them buy through Amazon and/or leave a review). Thus, when concluding that, e.g., the most negatively perceived aspects were "*Customer Service"* and "*Phone app"*, we cannot generalize to the whole market or argue that these opinions represent all customers. Therefore, given the interpretability of our results and in accordance with relevant literature, we conclude that our insights can be useful for a team of designers, although a controlled experiment is necessary to assess the real impact of using such a framework in a design process (Liu and Lu 2016).

Finally, in this thesis we tested several frameworks for analysing online customer reviews as an alternative to classic methods, and to enable a data-driven product design process. By leveraging different text models, we successfully distilled customer needs (explicit and implicit) in the form of feature sentiment pairs (FSPs) and keywords, and assessed them versus a gold standard. Namely, our best performing framework produced easily interpretable CNs that could prove useful during the early steps of a product design process. Furthermore, we propose that the extension of the techniques leveraged in this thesis might allow designers and firms to quantify the sentiment of product aspects, monitor competitor products, evaluate the performance of new product features on current markets, and even predict where new design opportunities reside. Moreover, performing an Aspect Based Sentiment Analysis regularly could allow firms to measure the effect of design improvements, by tracking how sentiments vary over time as new features are added. The author believes this potential is particularly important for small and medium size businesses, as they can seldom afford traditional market research for gathering vital customer knowledge.

7. References

Agarwal, Rakesh, Ramakrishnan Srikant, and others. 1994. “Fast Algorithms for Mining Association Rules.” In Proc. Of the 20th Vldb Conference, 487:499.

Akao, Yoji. 1990. “QFD: Integrating Customer Requirements into Product Design.” *Cambridge, MA*.

Archak, Nikolay, Anindya Ghose, and Panagiotis G Ipeirotis. 2011. “Deriving the Pricing Power of Product Features by Mining Consumer Reviews.” *Management Science* 57 (8): 1485– 1509.

Artstein, Ron. 2017. “Inter-Annotator Agreement.” In Handbook of Linguistic Annotation, 297–313. Springer.

Ardia, David, Keven Bluteau, Samuel Borms, and Kris Boudt. 2020. “The R Package Sentometrics to Compute, Aggregate and Predict with Textual Sentiment.” Forthcoming in Journal of Statistical Software. https://doi.org/10.2139/ssrn.3067734.

Ba, Sulin, and Paul A Pavlou. 2002. “Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior.” MIS Quarterly, 243–68.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022.

Calantone, Roger J, Sengun Yeniyurt, Janell D Townsend, and Jeffrey B Schmidt. 2010. “The Effects of Competition in Short Product Life-Cycle Markets: The Case of Motion Pictures.” *Journal of Product Innovation Management* 27 (3): 349–61.

Cantino, Andrew. 2013. “InspectorGadget.” GitHub Repository. https://github.com/cantino/selectorgadget; GitHub.

Chen, Diandi, Dawen Zhang, Fei Tao, and Ang Liu. 2019. “Analysis of Customer Reviews for Product Service System Design Based on Cloud Computing.” *Procedia CIRP* 83: 522–27.

Chen, Pei-Yu, Samita Dhanasobhon, and Michael D Smith. 2008. “All Reviews Are Not Created Equal: The Disaggregate Impact of Reviews and Reviewers at Amazon. Com.”

Chevalier, Judith A, and Dina Mayzlin. 2006. “The Effect of Word of Mouth on Sales: Online Book Reviews.” *Journal of Marketing Research* 43 (3): 345–54.

Clausing, Don P. 1993. “World-Class Concurrent Engineering.” In *Concurrent Engineering: Tools and Technologies for Mechanical System Design*, 3–40. Springer.

Clemons, Eric K, Guodong Gordon Gao, and Lorin M Hitt. 2006. “When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry.” Journal of Management Information Systems 23 (2): 149–71.

Cohen, Lou. 1995. *Quality Function Deployment: How to Make Qfd Work for You*. Prentice Hall.

DeJong, Gerald. 1982. “An Overview of the Frump System.” Strategies for Natural Language Processing 113: 149–76.

Dellarocas, Chrysanthos. 2003. “The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms.” *Management Science* 49 (10): 1407–24.

Etzioni, Oren, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. “Unsupervised Named-Entity Extraction from the Web: An Experimental Study.” Artificial Intelligence 165 (1): 91–134.

Evans, Richard D, James X Gao, Sara Mahdikhah, Mourad Messaadia, and David Baudry. 2015. “A Review of Crowdsourcing Literature Related to the Manufacturing Industry.” Journal of Advanced Management Science 4 (3): 224–321.

Fay, Michael P, and Maintainer Michael P Fay. 2020. “Package ‘Exact2x2’.”

Groves, Robert M. 2006. “Nonresponse Rates and Nonresponse Bias in Household Surveys.” Public Opinion Quarterly 70 (5): 646–75.

Haynes, Winston. 2013. “Bonferroni Correction.” In Encyclopedia of Systems Biology, edited by Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, 154–54. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-9863-7_1213.

Hedegaard, Steffen, and Jakob Grue Simonsen. 2013. “Extracting Usability and User Experience Information from Online User Reviews.” In Proceedings of the Sigchi Conference on Human Factors in Computing Systems, 2089–98.

Hovy, Eduard, Chin-Yew Lin, and others. 1999. “Automated Text Summarization in Summarist.” Advances in Automatic Text Summarization 14: 81–94.

Hu, Minqing, and Bing Liu. 2004. “Mining Opinion Features in Customer Reviews.” In *AAAI*, 4:755–60. 4.

Hu, Minqing, and Bing Liu. 2006. “Opinion Extraction and Summarization on the Web.” In *AAAI*, 7:1621–4.

Ireland, Robert, and Ang Liu. 2018. “Application of Data Analytics for Product Design: Sentiment Analysis of Online Product Reviews.” CIRP Journal of Manufacturing Science and Technology 23: 128–44.

Jin, Jian, Ping Ji, and Rui Gu. 2016. “Identifying Comparative Customer Requirements from Product Online Reviews for Competitor Analysis.” Engineering Applications of Artificial Intelligence 49: 61–73.

Jin, Jian, Ping Ji, and Chun Kit Kwong. 2016. “What Makes Consumers Unsatisfied with Your Products: Review Analysis at a Fine-Grained Level.” Engineering Applications of Artificial Intelligence 47: 38–48.

Jin, Jian, Ping Ji, and Ying Liu. 2014. “Prioritising Engineering Characteristics Based on Customer Online Reviews for Quality Function Deployment.” Journal of Engineering Design 25 (7-9): 303–24.

Jin, Jian, Ying Liu, Ping Ji, and Hongguang Liu. 2016. “Understanding Big Consumer Opinion Data for Market-Driven Product Design.” International Journal of Production Research 54 (10): 3019–41.

Jones, Karen Sparck. 1993a. “Discourse Modeling for Automatic Text Summarizing.” Technical Report.

Jones, Karen Sparck. 1993b. “What Might Be in a Summary?” Information Retrieval 93: 9–26.

Kang, Yin, and Lina Zhou. 2013. “Extracting Product Features from Online Consumer Reviews.”

Koomsap, Pisut, and Risdiyono. 2013. “Design by Customer: Concept and Applications.” *Journal of Intelligent Manufacturing* 24 (2): 295–311.

Kulak, Osman, Selcuk Cebi, and Cengiz Kahraman. 2010. “Applications of Axiomatic Design Principles: A Literature Review.” *Expert Systems with Applications* 37 (9): 6705–17.

Kupiec, Julian, Jan Pedersen, and Francine Chen. 1995. “A Trainable Document Summarizer.” In Proceedings of the 18th Annual International Acm Sigir Conference on Research and Development in Information Retrieval, 68–73.

Lee, Jay, Hung-An Kao, and Shanhu Yang. 2014. “Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment.” *Procedia Cirp* 16: 3–8.

Lee, Thomas, and Eric T Bradlow. 2007. “Automatic Construction of Conjoint Attributes and Levels from Online Customer Reviews.” University of Pennsylvania, the Wharton School Working Paper.

Lewis, David D, Yiming Yang, Tony Russell-Rose, and Fan Li. 2004. “Rcv1: A New Benchmark Collection for Text Categorization Research.” Journal of Machine Learning Research 5 (Apr): 361–97.

Liu, Ang, and Stephen C-Y Lu. 2016. “A Crowdsourcing Design Framework for Concept Generation.” CIRP Annals 65 (1): 177–80.

Liu, Bing, Minqing Hu, and Junsheng Cheng. 2005. “Opinion Observer: Analyzing and Comparing Opinions on the Web.” In Proceedings of the 14th International Conference on World Wide Web, 342–51.

Lutters, Eric, Winnie Dankers, Ellen Oude Luttikhuis, and Jos de Lange. 2014. “Network Based Requirement Specification.” *CIRP Annals* 63 (1): 133–36.

McAfee, Andrew, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. 2012. “Big Data: The Management Revolution.” *Harvard Business Review* 90 (10): 60–68.

McDonald, Ryan, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, et al. 2013. “Universal Dependency Annotation for Multilingual Parsing.” In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 92–97.

Misuraca, Michelangelo, Alessia Forciniti, Germana Scepi, and Maria Spano. 2020. “Sentiment Analysis for Education with R: Packages, Methods and Practical Applications.” arXiv Preprint arXiv:2005.12840.

Mohammad, Saif M. 2017. “Challenges in Sentiment Analysis.” In A Practical Guide to Sentiment Analysis, 61–83. Springer.

Mudambi, Susan M, and David Schuff. 2010. “Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon. Com.” MIS Quarterly, 185–200.

Nellore, Rajesh. 2001. *Managing Buyer-Supplier Relations: The Winning Edge Through Specification Management*. Routledge.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko. 2012. “Mine Your Own Business: Market-Structure Surveillance Through Text Mining.” *Marketing Science* 31 (3): 521–43.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. “Universal Dependencies V2: An Evergrowing Multilingual Treebank Collection.” arXiv Preprint arXiv:2004.10643.

Paice, Chris D. 1990. “Constructing Literature Abstracts by Computer: Techniques and Prospects.” Information Processing & Management 26 (1): 171–86.

Pang, Bo, and Lillian Lee. 2008. “Opinion Mining and Sentiment Analysis.” Information Retrieval 2 (1-2): 1–135.

Pavlou, Paul A, and David Gefen. 2004. “Building Effective Online Marketplaces with Institution-Based Trust.” Information Systems Research 15 (1): 37–59.

Popescu, Ana-Maria, and Orena Etzioni. 2007. “Extracting Product Features and Opinions from Reviews.” In Natural Language Processing and Text Mining, 9–28. Springer.

Qi, Jiayin, Zhenping Zhang, Seongmin Jeon, and Yanquan Zhou. 2016. “Mining Customer Requirements from Online Reviews: A Product Improvement Perspective.” Information & Management 53 (8): 951–63.

Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. 2009. “Expanding Domain Sentiment Lexicon Through Double Propagation.” In IJCAI, 9:1199–1204. Citeseer.

Radev, Dragomir R., and K. McKeown. 1998. “Generating Natural Language Summaries from Multiple on-Line Sources.” In.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reisenbichler, Martin, and Thomas Reutterer. 2019. "Topic Modeling in Marketing: Recent Advances and Research Opportunities." Journal of Business Economics 89 (3): 327–56.

Ren, Jingye Wang Heng. 2007. "Feature-Based Customer Review Mining."

Rinker, Tyler W. 2018. textstem: Tools for Stemming and Lemmatizing Text. Buffalo, New York. http://github.com/trinker/textstem.

Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. "Automatic Keyword Extraction from Individual Documents." Text Mining: Applications and Theory 1: 1–20.

RStudio Team. 2020. RStudio: Integrated Development Environment for R. Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Schuh, Guenther, Till Potente, Rawina Varandani, and Torben Schmitz. 2014. "Global Footprint Design Based on Genetic Algorithms–an 'Industry 4.0' Perspective." CIRP Annals 63 (1): 433–36.

Shah, Denish, Roland T Rust, Ananthanarayanan Parasuraman, Richard Staelin, and George S Day. 2006. "The Path to Customer Centricity." *Journal of Service Research* 9 (2): 113–24.

Straka, Milan, Jan Hajic, and Jana Straková. 2016. "UDPipe: Trainable Pipeline for Processing Conll-U Files Performing Tokenization, Morphological Analysis, Pos Tagging and Parsing." In Proceedings of the Tenth International Conference on Language Resources and Evaluation (Lrec'16), 4290–7.

Tait, John Irving. 1982. "Automatic Summarising of English Texts." University of Cambridge, Computer Laboratory.

Tan, Luke Kien-Weng, Jin-Cheon Na, Yin-Leng Theng, and Kuiyu Chang. 2011. "Sentence-Level Sentiment Polarity Classification Using a Linguistic Approach." In International Conference on Asian Digital Libraries, 77–87. Springer.

Timoshenko, Artem, and John R Hauser. 2019. "Identifying Customer Needs from User-Generated Content." *Marketing Science* 38 (1): 1–20.

Tontini, Gerson. 2007. "Integrating the Kano Model and Qfd for Designing New Products." *Total Quality Management* 18 (6): 599–612.

Tseng, Mitchell M, and Xuehong Du. 1998. "Design by Customers for Mass Customization Products." *Cirp Annals* 47 (1): 103–6.

Tseng, Mitchell M, and Jianxin Jiao. 1998. "Computer-Aided Requirement Management for Product Definition: A Methodology and Implementation." *Concurrent Engineering* 6 (2): 145–60.

Tseng, Mitchell M, Jianxin Jiao, and M Eugene Merchant. 1996. “Design for Mass Customization.” *CIRP Annals* 45 (1): 153–56.

Tseng, Mitchell M, Torsten Kjellberg, and Stephen CY Lu. 2003. “Design in the New Ecommerce Era.” *CIRP Annals* 52 (2): 509–19.

Turney, Peter D. 2001. “Mining the Web for Synonyms: PMI-Ir Versus Lsa on Toefl.” In European Conference on Machine Learning, 491–502. Springer.

Turney, Peter D, and Michael L Littman. 2002. “Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus.” arXiv Preprint Cs/0212012.

Ulrich, Karl T. 2003. *Product Design and Development*. Tata McGraw-Hill Education.

Urban, Glen L, Philip L Johnson, and John R Hauser. 1984. “Testing Competitive Market Structures.” *Marketing Science* 3 (2): 83–112.

Wang, Y, and MM Tseng. 2008. “Incorporating Probabilistic Model of Customers’ Preferences in Concurrent Engineering.” *CIRP Annals* 57 (1): 137–40.

Wang, Yue, Daniel Y Mo, and Mitchell M Tseng. 2018. “Mapping Customer Needs to Design Parameters in the Front End of Product Design by Applying Deep Learning.” *CIRP Annals* 67 (1): 145–48.

Wang, Yue, and Mitchell M Tseng. 2011. “Integrating Comprehensive Customer Requirements into Product Design.” *CIRP Annals* 60 (1): 175–78.

Yang, Bai, Ying Liu, Yan Liang, and Min Tang. 2019. “Exploiting User Experience from Online Customer Reviews for Product Design.” *International Journal of Information Management* 46: 173–86.

Yin, Dezhi, Samuel D Bond, and Han Zhang. 2014. “Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews.” MIS Quarterly 38 (2): 539–60.

Zhou, Haochen, and Fei Song. 2015. “Aspect-Level Sentiment Analysis Based on a Generalized Probabilistic Topic and Syntax Model.” In The Twenty-Eighth International Flairs Conference

8. Appendix

Table 19. All FSPs extracted through Dependency Parsing

great sound	soild Bass	decent quality	good experience
good quality	weak force	good comfort	easy bluetooth
decent cancellation	average head	gone connection	faulty pair
better quality	hot ears	impressive sound	faulty set
good cancellation	comfy pads	tedious adjustment	able ones
great quality	separate power	tedious adjustment	good battery
great Battery	soft Pads	tedious adjustment	great battery
perfect fit	durable body	good equipment	quick panel
excellent quality	unprecedented battery	great cancellation	excellent function
low battery	crafted one	solid app	high layout
awesome sound	attractive frame	outstanding clarity	good system
good produce	smooth control	seamless experience	active feature
excellent cancellation	cheap ones	awesome quality	respectable Battery
tight bass	good connection	crisp quality	customizable app
tight reinforcement	great charge	great TV	quick charging
mellow mids	wonderful sound	comfortable feel	smooth sound
pleasant audible	comfy Headphones	problematic connection	strong bass
insulative material	same battery	sensitive control	sleek design
great functionality	comparable sound	amazing Separation	much way
comfortable cushion	brighter light	crisp hats	simplistic far.Packaging
fantastic quality	vibrate phone	nice thing	fine fits
better build	good product	good Ns	more Weight
alright Sub-bass	dead battery	available lot	superior quality
less vocals	awesome battery	enough issues	similar comfort
passable Clarity	able earbuds	unforgivable must.2	smooth adjustment
unimpressive soundstage	ok fit	good sound	smooth adjustment
great connectivity	non-existent line	mediocre quality	similar openings
alright ANC	obnoxious mid-tone	stable connection	responsive controls
negative review	high volume	horrible assist	dead time
soft cushions	present tones	moist hands	buggy app
tactile Media	nice product	shinier predecessors	ready product
solid Construction	great cable	better version	subtle defect
soft cups	little cord	oriente isexpansive	roomy cushions
average quality	tough plastic	great headset	wide cushion
good bass	potent bass	great range	good form
greatest cancellation	comfortable cushions	apparent bass	outstanding cancellation
better case	beautiful color	comfy earpads	slimmer case
neat bass	fine mower	beat genre	inconvenient ones
amazing battery	nice packaging	soft Cushioning	responsive controls

Table 20. Distribution of Manually Annotated Polarities

Aspect	Num. Of Features	Positive (%)	Negative (%)	Neutral (%)
Battery life	30	86.67	6.67	6.67
Comfort	45	22.22	77.78	0.00
Connectivity	53	28.30	56.60	15.09
Controls	45	77.78	22.22	0.00
Earpads	12	58.33	33.33	8.33
Noise Cancellation	55	81.82	12.73	5.45
Phone App	14	50.00	50.00	0.00
Sound	103	81.55	16.50	1.94
Sound Features	61	60.66	37.70	1.64
Voice Assistance	8	62.50	25.00	12.50

Table 21. Distribution of Sentence-based Sentiment Analysis

Aspect	Num. Of Sentences	Positive (%)	Negative (%)	Neutral (%)
Battery life	46	67.39	32.61	0.00
Comfort	70	80.00	10.00	10.00
Connectivity	95	64.21	25.26	10.53
Controls	58	63.79	31.03	5.17
Earpads	58	46.55	34.48	18.97
Noise Cancellation	95	38.95	41.05	20.00
Phone App	48	47.92	43.75	8.33
Sound	171	73.68	16.37	9.94
Sound Features	48	54.17	37.50	8.33
Voice Assistance	23	65.22	30.43	4.35

Table 22. Distribution of Aspect-based Sentiment Analysis

Aspect	Num. Of Features	Positive (%)	Negative (%)	Neutral (%)
Battery life	12	58,33	25.00	16.67
Comfort	13	76,92	7.69	15.38
Connectivity	8	75,00	25.00	0.00
Controls	7	71,43	14.29	14.29
Earpads	8	87,50	12.50	0.00
Noise Cancellation	13	92,31	0.00	7.69
Phone App	3	66,67	33.33	0.00
Sound	14	92,86	0.00	7.14
Sound Features	18	61,11	16.67	22.22
Voice Assistance	1	0,00	100.00	0.00

Table 23. Example Manual Annotation

Term	Feature	Polarity
Great sound quality - TERRIBLE usability. Great sound quality, great noise cancellation, finally bose try to catch up with the new technologies (ambiance amplification and touch)!On the Flip side:1 - Heavy and get uncomfortable if you wear them for a long time. The top of the head hurts They fall-off my head a lot. Can’t use them in the gym, can lean your head back!2- They fall off your head!!! I cant use them at the gym, I cant lay down on the plane, I cant tilt my head... WHAT ARE THEY MADE FOR THEN?3 - VERY user-unfriendly!!!! I never turn them off correctly!!! They never turn off or on when I need them!4 - Surprised you cant edit music themes or equalize it5 - Pairing them is a hassle!6 - They are not easy to carry around! My sennheisers and Sonys are folded and get fit into my bag, these ones are way too inconvenient to carry around.I would say this is an overall fail for BOSE.	ambiance amplification	1
	comfort	-1
	fit	-1
	noise cancellation	1
	pairing	-1
	sound quality	1
	sound quality	1
	storage	-1
	touch controls	1
	usability	-1
	usability	-1
	weight	-1