Erasmus School of Economics

MSc in Economics and Business

Master Thesis

The use of Text Analytics on Customer Reviews: Identifying Features that can Differentiate between Satisfied and Dissatisfied Customers

Author	Youri Dullens
Student number	432703
Supervisor	V. Landsman-Schwartz
Second assessor	F. Frasincar
Date final version:	22-07-2019

Abstract

The introduction of online word-of-mouth (e-WOM) has been an increasingly important influence in the customer decision journey. One of the most important aspects of e-WOM are customer reviews. Prior research has shown that reviewers tend to be either very positive or very negative about their product experience, measured by rating. However, a clear overview of what textual features drive people’s emotions is still missing. This study aims to detect features from the review text to explain differences between satisfied and dissatisfied reviewers. I gathered product review data about the Google Chromecast HDMI Streaming Media Player from an Amazon customer reviews dataset of home entertainment products. The logistic regression has been used to measure the effect of textual features on the binary outcome variable product rating. The results suggests that the emotions Joy and Disgust differentiate the best between positive and negative reviews. As a consequence, the topic modeling approach Latent Dirichlet Allocation, has been used to find hidden topics in reviews with a large share of joyful and disgustful words. This has provided many insights in the strengths and weaknesses of the product. Managerial implications are provided based on these results in combination with prior literature, which are applicable to a broad range of product categories.

Keywords: Product Reviews, Machine Learning, Text Analytics, Sentiment Analysis, Customer Decision Journey.

Table of contents

1. Introduction

The number of internet users has risen with 1750% since the start of this century [CITATION Int19 \l 1043]. This had led to a huge shift in consumer behavior. Businesses started to find new channels to offer their products on. The innovative retailers started their own web shop and integrated an omni-channel strategy, both online and offline. Customers were using the online channel for different purposes. These varied from finding product information, looking for the best deal or avoiding the need to travel[CITATION Nyx17 \l 1043]. Nowadays, the internet has become an essential part in many people's lives. For example, 90% of the Dutch people access the internet every day[CITATION CBS15 \l 1043].

The internet is a great opportunity for retailers to sell their offer through, because it provides better possibilities to reach a larger audience relative to physical shops. A physical shop is dependent on its location, limited to opening hours and space in the shop. These are weaknesses that do not harm web shops. Web shops can also automatically collect data and attract people from other countries with similar preferences. However, there are also several downsides of online shopping. The internet is not so credible and trusted as physical stores[CITATION Kim03 \l 1043]. For example, a large share of the active internet users have experienced some sort of scam or fake websites (Scamwatch, 2019). Therefore, people have become very skeptical whether to believe what certain sites show them. As a consequence, people have more trust in what fellow consumers think about a product than what the seller puts in its description[CITATION Pra17 \l 1043].

The internet makes it easy to offer products to a wide range of people. Moreover, low initial investment costs increase competition. Therefore, it is not the product or service itself that drives people to make a purchase. More important factors are the brand name, trust in the brand, product reviews or ratings, the status of the platform the retailer is selling on and the perceived value for money of the deal[CITATION Dav09 \l 1043]. Especially product reviews have become an important influence in the purchase decision. For example, products with a rating below four face a huge drop in sales relative to higher rated products in the same industry[CITATION Pra17 \l 1043]. The same accounts for product offers with a low amount of reviews relative to other products on the page.

Due to the relative novelty of text analytics, a majority of businesses do not have the right tools to understand their customers' opinions in product reviews or on social media. Moreover, inadequate measures of popularity, such as number of followers on Twitter or Instagram, are overestimated[CITATION Mer \l 1043]. Businesses would rather benefit from a more specific approach that informs how satisfied reviewers perceive a product compared to dissatisfied reviewers and extract information about the most positive and negative arguments.

The literature about text analytics on product reviews is extensive and used for different purposes. Berezina, Bilgihan, Cobanogly & Okumus (2016) used text mining on reviews to understand satisfaction drivers in a hotel stay. They found that satisfied customers referred to intangible aspects of their hotel stay, whereas dissatisfied customers mention the tangible aspects more frequently. Moreover, McAuley & Leskovec (2013) used topic modeling techniques on text reviews to predict whether a user will read Harry Potter. They found hidden topics in the reviews and matched these with individuals to create a personalized profile for their recommender system.

As mentioned, the importance of good product reviews and ratings has increased in driving customer purchases. The product rating influences multiple steps in the customer journey[CITATION Var17 \l 1043]. For example, products with a relatively high rating will have a larger likelihood to be in people's initial consideration set, as customers use rating as a broad filtering mechanism. Moreover, the rise of social media and e-commerce platforms has enabled to obtain information about customer's opinions from customer reviews. A deeper understanding of customer preferences regarding a product or service is essential to implement the right product changes and improve customer experience. Therefore, I propose a tool that is composed of several analytical methods and can help managers utilize the information in reviews to better understand positive and negative aspects of their product. This leads to the following research question:

How can different features from customer reviews differentiate between satisfied and dissatisfied customers?

To start with, the paper analyses whether the content of product reviews can be helpful in explaining the product rating. The first step is to clean the review text from punctuations and misspellings and remove the most frequent and infrequent words. The second step is to add a combination of words, bigrams, sentiment and factors to the model as predictors. Then, a logistic regression analysis will be conducted. This regression intends to find textual features that differentiate well between the most positive and negative reviews, measured by a five star and one star product rating. Especially the most negative features are interesting, as they could provide information about the weaknesses of the product. The next step elaborates on the most meaningful variables from the logistic regression, using the topic modeling technique Latent Dirichlet Allocation. This model intends to detect hidden topics in the reviews that could help managers in different ways. For example, a distribution of hidden topics could reveal issues that certain groups of users have with the product or the customer perception regarding after sales service.

The paper is organized as follows. The next chapter provides the literature review, which describes theoretical implications about Word-of-mouth, the perceived credibility of reviews and the role of reviews in the customer journey. The third chapter presents the underlying data preparation steps and variable creation from the review text. The fourth chapter, the methodology, captures the methods that will be used to provide an answer on the research question. The fifth chapter elaborates on the main results from the analysis and will be supported by figures and tables. The sixth chapter discusses the main results and provides an answer on the research question. Moreover, it discusses the managerial implications of the results and several limitations of the paper followed by suggestions for further research.

2. Theoretical Framework

This chapter discusses the main theories related to product reviews and serves as the theoretical background of the paper. The first section describes the development of traditional word-of-mouth into online word-of-mouth. The second and third section provide reasons for people to contribute to word-of-mouth and describes the perceived credibility of online reviews respectively. Section four explains the steps in the customer journey and the role of reviews in this process. The next two sections describe the role of ratings and sentiment in the reviews. The chapter concludes with the economic value of product reviews and states the contribution of this paper to the research field.

2.1 Word-of-mouth vs online word-of-mouth

Word-of-mouth(WOM) is a buzz word to describe interpersonal communication, where people share experiences in their social circle (Godes & Mayzlin, 2004). In other words, WOM is a conversation between consumers about a product (Bambauer-Sachse & Mangold, 2011). It proves to be a strong consumer-dominated marketing communication channel, as people tend to follow the advice from a fellow customer much more quickly than from an advertisement on internet.

The rise of the internet together with social media has reduced the restrictions of WOM remarkably[CITATION Tru09 \l 1043]. What used to be sharing your personal experiences face-to-face, has evolved in posting them on social media or retailer websites. This has led to the emergence of online word-of-mouth (e-WOM), in which a post is visible for users all over the world. As a result of this development, the role of e-WOM has become much more important in the customer purchase decision[CITATION Dua08 \l 1043]. This is, because managers believe that a product's success is related to the valence of the WOM that it creates[CITATION God04 \l 1043]. Moreover, interpersonal communication tends to increase brand awareness and persuades individuals to try new products[CITATION Kle09 \l 1043]. The transition from traditional WOM to online mediums has benefited both customers and managers[CITATION Bro07 \l 1043]. They state that individuals have easier access to the opinion of fellow customers, whereas managers can encourage its customers to write a review on their site concerning the characteristics of a product or service.

The growing importance of online word-of-mouth can also be found in the literature. Many researchers have investigated the impact of e-WOM on product sales. Bughin, Doogan & Vetvik (2010) argued that e-WOM has an impact on between 20% to 50% of all purchasing decisions. They also state that WOM in general generates more than twice the sales compared with paid advertising. Duan, Gu & Whinston (2008) found that the volume of online WOM has a major influence on the product sales. The volume can be referred to as the number of reviews per product offer. This suggests that a product with relatively more reviews tends to have more sales. The valence of e-WOM is another frequently studied factor. Chevalier & Mayzlin (2006) describe valence as an individual's perception towards a product, which can be positive or negative. They have found that reviews with a more positive sense have a positive relation to product sales. Furthermore, more positive valence of reviews also increases the number of reviews, because the people that decided to buy the product based on positive e-WOM are also more likely to write a positive review themselves[CITATION God04 \l 1043].

2.2 reasons to contribute to e-WOM

The impact of e-WOM on customers and product sales have been made clear. However, reasons for customers to write a product review and contribute to e-WOM are yet to be discussed. Sundaram, Mittra & Webster (1998) provide two reasons for individuals to write a product review. First, reviewers could have the desire to help the company. A customer might have a close relation with the company's employees or is very loyal to the brand. The second reason is that people could have benefitted from product purchases in the past and feel the need to recommend this to others. Hennig, Gwinner, Walsh & Gremler (2004) argue that altruism is also a reason to engage in e-WOM, as people might feel the desire to help others without expecting anything in return. Moreover, customers tend share e-WOM to shape the impressions others have of them[CITATION Ber \l 1043]. By recommending certain products or services to others, people are able to display their knowledge about a product or gain attention. For instance, an extensive review of the newest Apple product may be written to signal both welfare and intelligence to other readers.

There are also reasons to contribute to a negative review. An individual might want to take revenge for a product purchase that has led to a negative experience (Sundaram, Mittra & Webster, 1998). The person feels disadvantaged by the company and decides to write a negative review. Similarly, a customer that has had a negative experience might want to warn others not to make the same purchase (Sundaram, Mittra & Webster, 1998).

2.3 perceived credibility of reviews

Review credibility usually derives from a combination of trustworthiness, reliability, content quality and previous beliefs[CITATION Gre94 \l 1043]. As discussed, e-WOM has become a very influential source of communication to potential customers. Product reviews have gained credibility compared to a retailer's product descriptions for multiple reasons. First, reviews are written by individuals with experience of the product being considered, without having any interest in selling the product themselves (Park, Lee & Han, 2007). The customer will therefore perceive reviews as a more objective description of the product. The second reason is the level of product details in combination with the level of reviewer agreement. Jimenez & Mendoza (2013) argue that the more different reviewers agree on detailed aspects of the product, the higher the perceived credibility. This is in line with the fact that customers have less trust in product descriptions from the seller, because there is no second opinion to confirm the statements in the product description.

The quantity of reviews also has an impact on the perceived credibility. Zhang & Watts (2003) have shown that information consistency among reviews has a positive effect on knowledge adoption in the online community. This implies that more similar reviews about a certain characteristic of the product tend to improve perceived review credibility. Cheung, Luo, Sia & Chen (2009) agree with the theory of Zhang & Watts (2003) and state that customers are more likely to adopt a more consistent point of view across most reviews. On the other hand, customers are more sceptical towards an opinion that is only shared by a relatively small number of reviewers. Moreover, Gavilan, Avello & Navarro (2018) differentiate the impact of review quantity on credibility between high and low product ratings. They state that when the rating is high, the perceived credibility depends on the number of reviews. However, if the rating is low, the quantity of reviews has no effect.

Another influence of perceived credibility is product type. Bae and Lee (2011) compared review credibility between search and experience goods. Search goods are characterized by attributes for which full information can be acquired prior to purchase[CITATION Nel70 \l 1043]. Experience goods mainly contain attributes from which information search is more costly than direct product experience. Bae and Lee (2011) found that the perceived credibility of a review is higher for experience goods than for search goods. This could be due to more information asymmetry with the seller for experience goods.

At last, Kamins & Lawrence (1988) argue that the level of sentiment, in terms of review sidedness, also has an effect on review credibility. A review can be one sided or two sided. One-sided refers to having either a positive or negative sentiment in the review, whereas two sided means that a review could contain both positive and negative elements. Lim, Lwin & Whee (1995) state that two-sided messages are generally perceived as more credible than one-sided reviews. This could be due to biasedness in the review. A one-sided review is more likely to be positively or negatively biased, whereas the bias of a two-sided review could be more balanced.

There are also researchers, who claim that online reviews are less credible compared to other information sources, like information provided by the seller. Literature provides two returning reasons. First, the identity of the source of an online review is not verified nor is the content of the message (Johnson and Kaye, 2002). Therefore, the reviewer can write anything without being monitored by authorities. From this, anonymous reviewers do not have to take their reputation into account, which tends to reduce credibility. Moreover, Wathen and Burkel (2002) have found a negative relation between the review date and the level of perceived credibility. This implies that individuals tend to perceive an outdated review as less accurate and of lower quality than a more recent review.

2.4 customer decision journey

The rise of the internet has triggered customers to interact with companies through numerous touch points of different channels and media. Not only the traditional media channels, just as television, radio or newspaper, but also online media channels have become a part of customer experience[CITATION Lem16 \l 1043]. Online media channels are referred to as webshops, social media and blogs. The online media channels are all two-sided, which allows the sender of the message to receive direct feedback from the reader. For instance, a company posts a video on social media about their newest product and people can instantly provide the company with feedback about the product. However, the two-sidedness of online media channels has also led to a change in customer behaviour compared to the era of traditional media [CITATION Her10 \l 1043]. The seller used to be the party that determined the product selection process of its own store. Then visited the store and bought products. However, the effectiveness of this approach has decreased in the era of online media channels [CITATION Jia13 \l 1043]. The introduction of webshops has had a major impact on the changing customer behaviour. Jun, Yang & Kim (2004) state that service convenience features, such as ease of use, the depth and richness of information and interactivity with the seller are major components for the adoption of webshops in the customer journey.

Traditionally, all pre-purchase touch points in the customer journey were offline. However, the emergence of e-commerce channels has created new touch points for customers in the evaluation phase of their journey. As mentioned by Duan, Gu, & Whinston (2008), the role of e-WOM, in the form of product reviews, has become much more important in the customer purchase decision. However, before we can touch upon the role of product reviews in the evaluating process of customers, the touch points in their decision journey have to be explained.

The customer journey is a dynamic process where the customer evaluates multiple product offers over time across multiple touch points[CITATION Lem16 \l 1043]. Consistent with prior research (Neslin et al. 2006; Pucinelli et al. 2009) the customer journey can be conceptualized in three stages, pre-purchase, purchase and post-purchase.

The pre-purchase stage includes all customer interactions with brands, informative sources and initial considerations. It starts with a trigger that initiates demand for a product or service. For instance, this could be a broken shoe. The initial consideration set of a customer consists of the brands that customers initially take into consideration. The initial set could be formed based on previous experience with a brand or social influences. Thereafter, the active evaluation phase starts in which the customer compares product descriptions, brands and shops. This evaluation can be done completely offline, online or omni-channel.

The purchase stage encompasses all interactions with the seller during the purchase itself. Thereafter, a purchase trigger occurs that will lead to the moment of purchase. This could be a special discount or deal.

The post-purchase stage refers to the after sales period and unconsciously influences customers in their next purchase decision[CITATION Fen09 \l 1043]. A very positive experience could drive the customer to the loyalty loop. This means that the customer will have a larger probability to buy from your brand than from other brands in the future. At the other hand, a negative experience will force the individual to start the customer decision journey over again.

2.4.1 role of reviews in customer journey

As mentioned, the importance of good product reviews and ratings has increased in driving customer purchase decisions. Reviews are merely used during the pre-purchase stage, when individuals collect information on products in their consideration set. A review functions as a source of information that helps to reduce the information gap between them and the seller. As Dholakiya (2017) mentioned, individuals have more trust in how fellow customers perceive the product than what companies put in their product description. Therefore, reviews are used to validate the main advantages and/or disadvantages of the product. A review usually consists of two components, the rating and review text. The role of these components in the customer journey differs and will be discussed in the next sections.

2.5 The role of ratings in reviews

The literature from section 2.4 touches upon multiple components of e-WOM, but mainly focuses on review text and product rating. Therefore, it is important to gain an understanding of factors that influence product rating and the relation between the product rating and purchase behaviour.

Early research on consumer behaviour has shown that consumers feel uncertain about the outcome of a product purchase. A possible explanation is information asymmetry between buyer and seller. This describes the information advantage of seller over the buyer regarding characteristics of the product or service[CITATION Mis98 \l 1043]. The information asymmetry in e-commerce is even larger than in physical shops[CITATION Mav12 \l 1043]. Physical shops allow people to touch the product, ask questions about its characteristics or try it and check whether it fits. However, webshops do not have these benefits. Therefore, people must find other tools to reduce the risk on an online purchase. The product rating could function as such a tool. Dholakiya (2017) has shown that consumers tend to give more weight to the opinion of fellow customers than of the sellers themselves. A reason for this fact could be that customers do not have an interest in selling products themselves and are therefore perceived as more objective (Park, Lee & Han, 2007).

other researchers argued that the average ratings are biased and would not prefer to use the rating as a proxy for product quality. Hu, Pavlou & Zhang (2006) found that 53% of the product ratings have a bimodal distribution, meaning that ratings are only very positive or negative for most of the products. This suggests that the average rating is not an unbiased representation of the average reviewer, but is rather affected by the opinions of customers that were either very satisfied or dissatisfied with the product. Li & Hitt (2008) support the statement of Hu, Pavlou & Zhang (2006). They found that very positive or negative customers are more likely to give a product rating compared to individuals with a more neutral opinion. Mackiewicz (2007) explored the ratings of 640 online products and found that nearly 50% of all product ratings received five stars, the highest possible rank. This suggests that a high rank product rating is necessary to be considered as a purchase option. However, it does not represent better quality relative to other products, when these also have a high rank star rating. Zhang, Lee and Zhao (2010) also found that the average product ratings are generally high and argued that the purchasing bias could influence this effect. This bias suggests that people, who actually purchased the product, are the main group of reviewers. These people only buy products that they perceive positively, which leads to a higher ranked rating. Furthermore, individuals that write a review, after they purchased the product, are influenced by previous reviews from other customers[CITATION Moe11 \l 1043]. This suggests that products with a better rating tend to positively influence a new reviewer to also give a high rating. This paper uses rating to compare the very satisfied and dissatisfied customers. The reviews with a five star rating are perceived as satisfied and with one star as dissatisfied. Therefore, the average product rating is not taken into account, due to the argued biases above.

Moreover, another potential problem with using the average rating as a proxy concerns the information contained in the review[CITATION Nik11 \l 1043]. Economic theory tells us that products have multiple attributes and different consumers might weigh different levels of importance to these attributes [CITATION Ros74 \l 1043]. However, just using the average product rating implicitly assumes that people which the same rating have the same preferences regarding the product. Thus, unless the consumer that reads a review has similar preferences as the reviewer, an average product rating might not be sufficient to extract all relevant information for insights in the purchase decision.

As mentioned, this paper will compare the most satisfied and dissatisfied group of reviewers according to the rating. Therefore, it is important to obtain an understanding of how individuals perceive positive or negative reviews to be helpful in evaluating products.

Forman, Ghose & Wiesenfeld (2008) examined the relation between rating and sales and found that reviews with neutral ratings were considered less helpful to find information then reviews with very positive or negative ratings. This implies that consumers perceive the content of one-sided reviews as more helpful than balanced reviews with a neutral opinion. One-sided reviews provide the strengths and weaknesses of the product. Hence, providing more complete information, which could ultimately reduce information asymmetry (Cheung, Lee & Rabjohn, 2008). Pavlou & Dimoka (2006) demonstrate consistent findings.

When comparing positive and negative reviews, in general, customers perceive very negative reviews as more evident than very positive reviews [CITATION Mah90 \l 1043]. A larger quantity of positive reviews may play a role. The prospect theory of Kahneman & Tversky (1992) is in line with the findings of Maheswaran & Sternthal (1990). According to this theory, people tend to give more utility to a loss than to a gain of the same amount, where a loss could refer to a bad experience after purchase. It states that individuals are loss averse, meaning they are determined to avoid losses. negative ratings are perceived as more valuable than positive ratings, but the ratings play a small role for people that actually look for product characteristics.

2.6 the role of sentiment in reviews

Sentiment analysis is a common term in the field of natural language processing and detects sentimental words from text. As mentioned, e-WOM has influenced the customer purchase decision. Section 2.4.1 argues that product reviews are mainly used in the pre-purchase stage during active evaluation of a set of products. However, what could be the effect of sentiment in reviews on the customer decision making journey? Moreover, is the effect of sentiment different from the effect of product rating on the customer?

Research about consumer decision making has shown that customers faced with complex choices, tend to decrease their cognitive effort and use simplifying ways to make a decision (Kahneman & Tversky, 1974;Bettman, Luce & Payne, 1998). Ghose & Ipeirotis (2007) explored the impact of different features from text reviews on the sales of a product. They argue that the numeric review ratings did not fully capture the polarity of information in the text reviews, where polarity is the level of positive and negative words in a review. This suggests that a five star rating not only contains positive words, but also elements that point to negative sides of the product.

Hu, Koh & Reddy (2014) believe that the sentiment in a review provides customers with extra information on a product in addition to product ratings. They argue that the online evaluation process of different product sets can be complex and customers use different elements of information in different phases of the decision journey. This suggests that easily comparable information that does not require much effort to process, such as product ratings, may be used to reduce the consideration set. On the other hand, detailed information about the sentiments or experiences from previous customers is more difficult to evaluate and might be used to make the final decision[CITATION Hua09 \l 1043]. For instance, two products score excellent on rating, but the individuals choose the product with reviews that are more comparable to their preferences.

2.7 Economic value of product reviews

How can consumer generated content help managers to better understand the position of their products in the market? This section provides a summary of recent papers that created tools that could be helpful to provide an answer on this business question.

Netzer, Feldman, Goldenberg & Fresko (2012) utilized large-scale consumer-generated data from consumer forums to understand how consumers' compile their initial consideration set of products and the corresponding market structure insights. Note that they have not focused on a single entity but rather on a large number of entities from a certain product category. Therefore, their purpose was to find textual features that could provide information on the implied market structure. They have proposed that a manager could use this approach to monitor the changing market position over time at a higher resolution and lower costs relative to more traditional market structure methods i.e. quantitative research. Lee and Bradlow (2011) have presented another method to provide market structure insights. They have also utilized the textual features of online customer reviews to automatically detect product attributes and visualize the brand's position relative to its competitors in the market. However, they defined similarity based on the attributes that were mentioned in combination with a particular product. Such a similarity approach is more likely to appear in more structured text sources, such as product reviews. Managers could visualize the insights in a perceptual map to create an image of the perceived position of their products in the market. The axis could represent characteristics, such as price, durability or design. A possible insight could be that their product is perceived as more expensive than a close competitor, but has a greater design.

Decker and Trusov (2010) presented an econometric framework that can be applied to turn individual consumer opinions from product reviews into aggregate consumer preference data. They estimated the effect of product attributes and brand names on overall product evaluation relative to competitors. Furthermore, they have claimed to be helpful in reputation analysis, by comparing the aggregate sentiment level between different products. A manager could compare the aggregate sentiment level between two products to identify the perceived image of customers towards its products. Archak, Ghose & Ipeiroti (2011) argued that the textual content of product reviews is an important determinant of a customer's purchase decision, over and above the valence and rating of reviews. their method incorporated review text in a consumer choice model, by decomposing textual reviews into segments describing different product features. They have claimed that managers can use their approach to learn about a consumer's relative preferences for different product features. Ghose, Ipeirotis & Li (2012) proposed an approach that used the consumer's differentiated product preferences from user generated content to estimate demand and generate a ranking system. The model ranks products according to the level of aligned with a consumer's preferences. A manager could use this approach to estimate the relative demand between products and analyse what type of customers would be more likely to purchase a certain type of product.

At last, McAuley & Leskovec (2013) have used topic modeling techniques on review text for their recommender system regarding Harry Potter books. Topic modeling was used to uncover the implicit tastes, which each user has revealed in their reviews. Their approach could have several useful advantages for managers. Firstly, they were able to obtain highly interpretable textual labels, which helped to justify the given rating of a reviewer with the features from the review text. Secondly, the discovered topics in the review text could be useful to automatically detect genre preferences and identify representative reviews per genre.

2.8 summary literature

The literature focused on customer reviews is constantly evolving. Traditional word-of-mouth has made the transition to online word-of-mouth. This has triggered researchers to write about new subjects, such as perceived online review credibility or reasons to contribute to e-WOM. Researchers have also written extensively about the changing customer journey, due to the rise of e-commerce transactions and e-WOM. This has led to literature about new evaluation metrics, such as product rating and review text. Consequently, product ratings were perceived as more useful in the general phase of the customer journey, whereas the sentiment level in a review has been considered in circumstances closer to the actual purchase decision. Furthermore, the economic value of product reviews has been addresses in which researchers have provided several tools that could help improve a manager's understanding of the position of their product or service in the market.

However, the literature about product reviews could be improved. Many researchers have investigated effects related to the average product rating. However, this does not provide a thorough understanding of people's perception about a product, as the average product rating is generally high (Zhang, Lee & Zhao, 2010). Furthermore, literature that has addressed the limitations of the average product rating could also be improved. This paper contributes a tool, which is composed of several analytical methods that could help managers utilizing the information in reviews to better understand positive and negative aspects of their product. This paper combines the predictive power of logistic regression with the interpretational benefits of topic modeling technique Latent Dirichlet Allocation. Logistic regression will be conducted to find features that differentiated well between satisfied and dissatisfied customers. Latent Dirichlet Allocation will then detect substantiated opinions regarding the important features. Hence, this approach could be used as an exploratory tool for managers, who have not performed extensive research into customer preferences regarding their products.

3. Methods

This section will discuss the technical implications of the two methods that are used for the analysis, logistic regression and Latent Dirichlet Allocation. These two methods will be discussed separately.

3.1 Linear vs Logistic regression

The goal of this paper is to analyse the impact of different textual features on a dichotomous outcome. Here dichotomous means that the dependent variable is either classified as one of the only two possible outcomes[CITATION Pen10 \l 1043]. An example of a dichotomous outcome is whether students will pass or fail their final exam. Traditionally, these tasks were fulfilled by an ordinary least squares regression (OLS). OLS is a generalized linear method to estimate the unknown parameters that are related to the dependent variable [CITATION Jam131 \l 1043]. Its goal is to minimize the sum of squared error, which is the difference between the observed and predicted outcomes of the independent variables. However, OLS assumes that the response variable is quantitative and increases/decreases linearly with the coefficients of the predictors[CITATION Jam131 \l 1043]. This makes OLS less accurate on a classification task, where the goal is to predict probabilities. Moreover, OLS faces an interval problem. Probabilities normally range between 0 and 1, whereas OLS also produces negative and very positive outcomes, due to extreme values of the predictors. These estimations are not sensible, as it suggests a negative probability of an event[CITATION Jam131 \l 1043].

The following example clarifies the statements made above. In case of a logistic regression on one continuous predictor X(number of study hours) and a dichotomous outcome variable Y(the student's chance to pass the exam), a figure will result in two parallel lines, one for each outcome of the Y variable (Figure 4.1).

Figure 3.1: Logistic regression output of Reading score on the probability to pass the exam

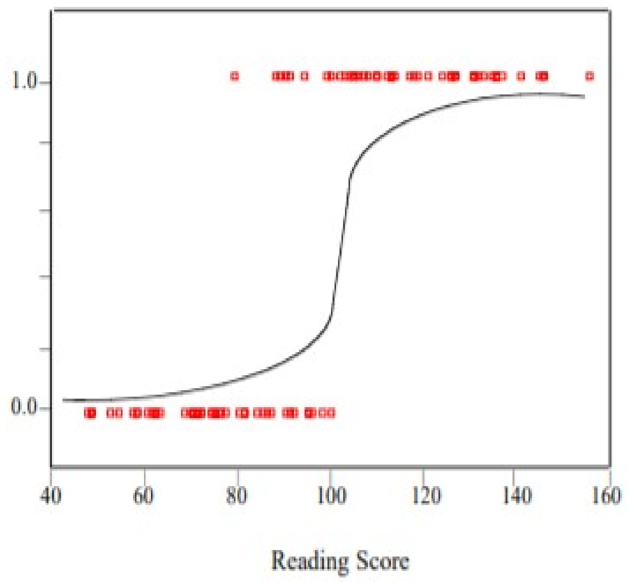


Figure 3.1 shows that the category means are linear in the middle, but curved at the start and the end. This is referred to as sigmoidal or S-shaped and difficult to describe with a linear regression for two reasons. Firstly, the extremes of the curves are not linear. This means that the probability of success -increases more quickly when to category means increase from 40 to 60, than for an increase from 140 to 160. Secondly, the errors are neither constant across the entire range of data nor normally distributed (Peng, Manz, & Keck, 2001). This implies that the predicted errors are larger for the category means closer to the average than to the ends of the curve.

Logistic regression is an alternative that does well in describing the relationship between a categorical outcome variable and one or more continuous or categorical predictors[CITATION Pen10 \l 1043]. The mathematical concept that underlies logistic regression is the logit, the natural logarithm of odds of Y. the odds are the ratios of probabilities (π) of Y happening to probabilities (1 – π) of Y not happening and can be easily derived from a 2 × 2 contingency table (table 3.1).

Table 3.1 Contingency table with fictional sample data about the exam results separated by gender

exam result	**gender**		
	boys	girls	total
pass	92	88	180
fail	8	12	20
total	100	100	200

Consider the situation, from Table 3.1, where a dichotomous outcome variable (students that either passes or fails the final exam) is paired with a dichotomous predictor variable (gender). The odds that a boy passes the exam is 11.5 (92/8) and a girl is 7.33 (88/12). The odds can take any value between 0 and infinity. Values close to zero and infinity indicate very low and high probabilities of success, respectively. Alternatively, one might prefer to compare a boy's odds of passing the exam relative to a girl's odds. The resulting odds ratio is 1.57, which suggests that boys are 1.57 times more likely to pass the exam compared with girls. The odds ratio is derived by dividing the male's odds of passing the exam by the girl's odds (92/8 for boys and 88/12 for girls). Therefore, odds larger than 1 represent a relatively larger likelihood for males to pass the exam. Furthermore, the natural logarithm of the odds ratio is a logit, [i.e., ln(1.57)] and equals 0.45. The value of 0.45 is the regression coefficient of the predictor *gender*.

The logistic function is technically described as

$$Logit(Y) = natural \log(odds) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X \quad (1)$$

The inverse logaritm of equation 1 on both sides derives an equation to predict the probability of the occurrence of the outcome of interest as follows:

$$\pi = Probability \left(Y = outcome\ of\ interest \middle| X = x, a\ specific\ value\ of\ X \text{¿} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \right. \quad (2)$$

Here π is the probability of the outcome of interest given a specific value of X, α is the Y intercept, β is the regression coefficient, and e = 2.718 is the base of the system of natural logarithms.

The predictor variables X can be categorical or continuous, but the outcome variable Y is always categorical. Following Equation 1, the relationship between logit (Y) and X is linear. However, Equation 2 presents a non-linear relationship between the probability of Y and X. this is because the amount that π changes due to a one-unit change in X depends on the current value of X. Figure 1 shows that a one-unit change in X increases the probability more, if the initial number of study hours is 30 than when it is 100. Therefore, it is required to take the natural log transformation of the odds in Equation 1, to make the relationship between a categorical outcome variable and its predictor(s) linear. However, the direction of the relationship between X and the logit of Y does not depend on the current value of X. If β is greater than zero, a larger value of X will be associated with a larger logit of Y i.e. an increasing probability π. Furthermore, if β is smaller than zero, a larger value of X will be associated with a smaller logit of Y i.e. a decreasing probability π.

The simple logistic regression can also be extended to a model with multiple predictors as follows:

$$Logit(Y) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta 1\, X 1 + \beta 2\, X 2 + BiXi$$

(3)

From this, the equation to predict the probability of occurrence of the outcome of interest is as follows:

$$\pi = Probability \left(Y = outcome\ of\ interest \middle| X 1 = x 1, X 2 = x 2 \text{¿} = \frac{e^{\alpha+\beta 1 X 1+\beta 2 X 2+\beta iXi}}{1+e^{\alpha+\beta 1 X 1+\beta 2 X 2+\beta iXi}} \right. \quad (4)$$

Here π is still the probability of the outcome of interest given a specific value of X, α is the Y intercept, βs represent the regression coefficients, and the X's are a set of predictor variables. The maximum likelihood method is typically uses to estimate the parameters of α and the βs. Haberman (2014) and Schlesselman

(1982) preferred this method over the weighted least squares approach, due to its statistical properties. The basic intuition of maximum likelihood is as follows. We fit the parameters α and the βs such that the predicted outcome is as close as possible to the individual's observed outcome for each individual. In other words, the optimal model yields a number close to one for all individuals that passed the exam and a number close to zero for who did not pass. This can be formalized in the likelihood function:

$$£(B0, B1) = \prod_{i:yi=1} p(xi) \prod_{i':yi'=0} \left(1 - p\left(xi'\right)\right) \tag{5}$$

The aim is the maximize this function, such that the likelihood of a correct prediction is maximized.

3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an unsupervised machine learning method to soft-cluster large quantities of discrete textual data[CITATION Ble03 \l 1043]. The model is particularly useful to automatically find latent structures in a collection of documents, where latent refers to hidden patterns in a text that were not visible before the analysis. LDA assumes that textual documents consist of topics and that these topics consist of words from the list of unique words in the documents. According to Brett (2012) the topics are a recurring pattern of co-occurring words. This implies that certain groups of words that tend to occur relatively more together in sentences, have a larger probability to be detected as a topic in documents.

Moreover, LDA is a mixed membership model, meaning that each document exhibits all topics with different proportions (Blei, 2012). The distributions, of term per topic and topic per document, are expressed by a vector of continuous non-negative latent variables that sum to 1 [CITATION Air15 \l 1043]. For instance, in a three topic model, the topic distribution for document *n* could be (0.5,0.49,0.01). This implies that the first two topics correspond to 99% of the document and that the content of the third topic can be disregarded for interpretation. Similarly, every word partially belongs to all topics with varying probabilities. The unique word list usually comprises numerous words. Therefore, the term-distribution for topic *k* will contain many terms close to zero and just relatively a few with larger proportions. Therefore, the modeling output of LDA is a list of words with their probabilities per topic and the different proportions of these topics per document. A list of the most probable words per topic (or topics per document) will then be used to understand the content of the topics (documents).

To achieve this, LDA estimates the posterior distribution of hidden variables, by performing data analysis on the joint probability distribution over hidden and observed variables[CITATION Rei19 \l 1043]. This paper computes LDA with a built-in function, LDA, from the *Topicmodels* package in R[CITATION Grü18 \l 1043]. Therefore, the mathematical derivation of the joint probability distribution is not included. However, this function still requires optimization of two parameters, the number of topics *k* and the controller of sparseness *α*.

I have chosen the first parameter, the number of topics *k,* based on the lowest perplexity on the validation set. Perplexity is a measurement of how accurate a probability distribution predicts a sample. The data has been split in a training and validation set. The aim is to pick *k* with the lowest perplexity on the validation set. The validation set is used to avoid the problem of overfitting, meaning that the model performs too well on the training data. This means that the noise or random fluctuations in the training data are learned as concepts by the model[CITATION Bro16 \l 1043]. However, the problem is that these fluctuations do not appear the same in new data. This will negatively impact the model's ability to generalize to unseen data. Therefore, *k* that has the lowest perplexity on the validation set is preferred.

The second parameter *α* affects topic sparsity through the Dirichlet distribution. The technical details of the Dirichlet distribution will help understanding how different values of *α* can control sparseness.

$$For\ p\ \ Dirichlet(\alpha 1, \alpha 2, \ldots \alpha k)$$

$$Expectation\ pk = E[pk] = \frac{\alpha k}{\sum_j \alpha j} \tag{6}$$

And

$$Variance\ pk = Var[pk] = \frac{E[pk](1 - E[pk])}{1 + \sum_j \alpha j}$$

(7)

Here $\sum_j \alpha j$ controls the sparseness. According to equation 6, a lower value of α leads to larger variance (more sparseness), which means more deviations between different documents. In this setting the model differentiates well in its prediction. Therefore, the probability that a document will be predicted to be just about topic A or B will be larger than that the document will be explained by a mixture of topics. On the other hand, increasing the α reduces the variance (less sparseness) and predicts the document to be about a mixture of topics. A high sparsity, therefore a low α, is preferred, as it is more informative to know what topic is mainly represented in the document rather than having 10 topics that all have modest input. I decided to choose the value of α based on the lowest perplexity on the validation set.

3.3 Principal Component Analysis

This paper applies Principal Component Analysis (PCA) to add factors as predictors in the model. PCA reduces the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information of the original dataset[CITATION Shl14 \l 1043]. It converts a series of observations of possibly correlated variables into a set of linearly uncorrelated variables, called orthogonal principal components or factors[CITATION Jam13 \l 1043]. These set of factors reduce the risk of multicollinearity between variables[CITATION Far67 \l 1043]. This implies that variables are more independent, because their effect is less affected by other variables i.e. they are linearly uncorrelated. Moreover, PCA aims to put as much information as possible in the first component, which is the factor with the largest variance explained[CITATION Jaa19 \l 1043]. Furthermore, all variables have their own loadings on every factor. These loadings represent the correlation between the variable and the corresponding factor and can be positive or negative. The importance of a factor is given by its eigenvalue. The eigenvalue represents the explained variance per factor. A larger eigenvalue is preferred, as more variance is explained by that factor. One could then select the number of factors based on the relative eigenvalues between components. This suggests that the cut-off point for the number of factors to use in the model should be based on a large difference in eigenvalue between two consecutive factors.

4. Data

This chapter provides an overview of the data and methods that are used in the analysis. The chapter will start with a brief description of the source and collection process of the data. The next section explains the data cleaning and preparation steps. Then I will discuss the variable creation process in which four categories of variables can be distinguished. The chapter ends with a data description section that summarizes the descriptive statistics related to a review. Furthermore, it provides an overview of the commonly used words in the review text.

4.1 Data Source & Collection

The dataset for this paper originates from a larger dataset on product reviews, which is called *amazon_reviews_us_Home_Entertainment.* This dataset consists of 362,297 reviews and 15 variables about different products in the home entertainment industry from July 2013 till August 2015. The data was made available by Amazon Customer Reviews and is a rich source of information for academic researchers in the fields of machine learning and natural language processing. This paper employs a dataset with reviews about the Google Chromecast HDMI streaming media player. This device streams internet content from a mobile device or personal computer on a television or audio system through mobile and/or web apps that support the Google Cast technology [CITATION Goo19 \l 1043]. Moreover, content from a Google Chrome web browser can be mirrored from a personal computer to another device. Furthermore, over 30 million units have been sold globally since the launch in 2013 till the end of 2014, making the Chromecast the most sold streaming device in the United States in 2014 [CITATION Mar15 \l 1043].

The dataset contains 9,001 reviews and three variables. First, *product rating* functions as the dependent variable and differentiates between satisfied and dissatisfied reviewers. The average rating is 4.1 out of 5 and is left-skewed, as 57% of all the reviewers have given a 5 star rating. Second, *verified_purchase* filters all reviews without a verified purchase. This will increase the review credibility, because I assume reviewers that actually purchased the product are more credible and provide more information about their perceived product experience. Third, the variable *review_body* contains the review text. This is the most important variable, as features from the review text will function as explanatory variables of the product rating.

4.2 Data Cleaning & Preparation

This section describes a stepwise approach of data cleaning and preparation for analysis. The data requires cleaning, as many reviews contain spelling mistakes, punctuation errors or capital letters. The variable *product rating* is a factor with five levels ranging from one to five. However, I intend to use rating as a proxy to satisfaction. Therefore, I have chosen to divide reviewers into satisfied and dissatisfied according to their rating. The people with a rating of one are perceived as dissatisfied and people with a five star rating as satisfied. The reviewers with a rating between two and four are removed from the dataset. This is, because I assume that comparing the most positive and negative review ratings will yield the most differentiating results. For example, an ordered variable with 5 levels assumes that the textual features from a 3 star review are more positive than the features from a 2 star review. This assumption is ambiguous, as both reviewers could be very dissatisfied with the product but have different perceptions regarding the rating system. One might perceive a 2 star rating as negative and 3 stars as neutral, whereas another individual might perceive all ratings below 4 stars as negative. This ambiguity problem will less likely occur in a model with just the most positive and negative ratings, as the valence of the features are assumed to be more one-sided. Therefore, removing the reviews with a rating between 2 and 4 tends to reduce the ambiguity regarding the alignment between product rating and the valence of the features. This results in a binary rating with two classes, a zero for all 1 star reviews and a one for all 5 star reviews.

Next, I removed punctuation, capital letters and smileys. For example, the sentence “what can I say. I Love it!!!!! I Love it -:)” changed to “what can i say i love it i love it”. Capital letters were removed, as a machine interprets “Love” differently than “love”, whereas humans interpret them similarly.

Furthermore, the reviews contain many stop words that have to be removed from the dataset. Stop words are a set of commonly used words, that do not add interpretational value to the review. Examples of stop words are "the", "and" or "is". Such words appear multiple times per review. This is a problem, as many analytical models tend to allocate much weight to the most occurring words [CITATION Jam131 \l 1043]. The amount of stop words in reviews is substantial. Therefore, I have used the built-in function *stop_words* in R, which contains a list of 1149 stop words, to automatically remove all the stop words from reviews. This has removed 61% of the words in the dataset. In addition, it is also required to apply word stemming in a review. Stemming is the process of reducing a word to its root. A human interprets the words *love* and *loving* similarly, but a computer does not. Therefore, the words *love* and *loving* have to be stemmed to their root, *lov,* such that a computer will now interpret the words similarly.

The next step is to remove frequent and infrequent words from the review text. Infrequent refers to words that occur in less than 0.5% of the reviews. I have chosen to remove these words for two reasons. First, infrequent words are less likely to have co-occurrences, meaning that these words have a smaller probability to occur with other words in reviews or single sentences. This is because co-occurrences between words only originate from non-sparse entries[CITATION Lak17 \l 1043]. For example, a word that occurs in 10 of the 9000 reviews will only be non-sparse in the 10 reviews it appears in. On the other hand, words that occur in 1000 of the 9000 reviews will have many non-sparse entries and thereby have a larger probability to co-occur with words. Therefore, the infrequent words only add sparsity to the data without having a substantial amount of co-occurrences. Second, the removal of infrequent words automatically removes most of the misspelled words, as such words will not occur in more than 0.5% of the reviews. Hence, this saves much cleaning time. Furthermore, the most frequent words tend to appear in almost every review. For example, the word *Chromecast* describes the product name. Such words do not provide useful information to differentiate between satisfied and dissatisfied reviewers, as both groups use these words.

Furthermore, I added the variable *Nr_of_words*, which counts the number of words in a review. Then, I filtered the reviews on a minimum word count of two, which improves the quality of the results. A review with one word lacks the ability to co-occur with other words in the review, as the review does not contain any other words. Therefore, it is per definition not possible to have co-occurrences in reviews with just one word, after data cleaning steps.

The final step is to create Document-Term Matrices, one for single words and one for bigrams. The Document-Term matrix will be used as input data for both the regression and topic modeling models. It assumes a bag-of-words model. This means the model interprets the words to be isolated and ignores information about the order or location of the words relative to other words in a review[CITATION Bro17 \l 1043]. Therefore, it only matters whether words will co-occur within reviews and the more reviews that contain this co-occurrence, the larger the strength of the association between the words. Note that co-occurrences can only occur from non-sparse entries[CITATION Lak17 \l 1043]. Hence, shorter reviews contain more sparse entries, as they consist of less unique words. This implies that a review with more sparse entries is less likely to have co-occurrences. A Document-Term matrix is a matrix where each row represents one review, each column one unique word and each value the number of appearances of that unique term in the particular document[CITATION SIl17 \l 1043]. For example, a column value of two in row three for the word *fun* means that *fun* occurs twice in the third review. A column value of zero means the term is sparse and thus not mentioned in the review.

4.3 Variable creation

All the variables that are used as predictors in the model are features from the review text, after cleaning. The variables can be split into four categories. The first category is single words. This category consists of the 75 most occurring words, apart from the frequent words that were deleted in the cleaning process. The most occurring words are more informative than less frequent words, as they are less sparse and have a larger probability of co-occurrence[CITATION Lak17 \l 1043].

The second category is bigrams. A bigram is a sequence of two adjacent words on any place in the review text[CITATION Bro17 \l 1043]. For example, the bigram *easi instal* might refer to easy installation of the product. Recall that the bag-of-words model does not account for context around a word in a review. A bigram is a sequence of two words and thus adds more context than a single word. Moreover, I removed the bigrams that appeared in less than 1% of the reviews, to avoid inclusion of thousands of bigrams in the dataset, This remains 29 bigrams in the dataset.

The third category contains emotions. People that give an opinion about a product use emotional words such as *appreciate* or *disappointing*. Such emotional words are very informative to obtain an understanding of how reviewers perceive a product. Emotion is added to the reviews as follows. Saif Mohammad (2019) has provided the NRC emotion lexicon. This is a list of English words and their association with eight emotions. The eight emotions are anger, fear, anticipation, trust, surprise, sadness, joy and disgust. The lexicon also contains two sentiments, negative and positive. The words from the review text are then inner joined with the words from the NRC lexicon. For example, a word that matches a word from the Joy emotion of NRC gets counted as a joyful word. This process repeats itself for every word in the review. The counts are therefore conducted per review. For example, a review with the text “Text analytics is fun” contains the emotional word *fun*. This word appears in the list of three emotions, joy, anticipation and positive. Therefore, this review has a count of one for joy, anticipation and positive and a count of zero for the other emotions.

The fourth category contains factors. Factors can be interpreted as dimensions in which all the unique words have a certain factor loading. Words that appear many times together will be perceived as more similar than words that only appear a few times together. Hence, the more similar words will have a larger component loading on the same factor. The larger the relative loading, the larger the correlation of the feature in the particular factor.

4.4 Data description

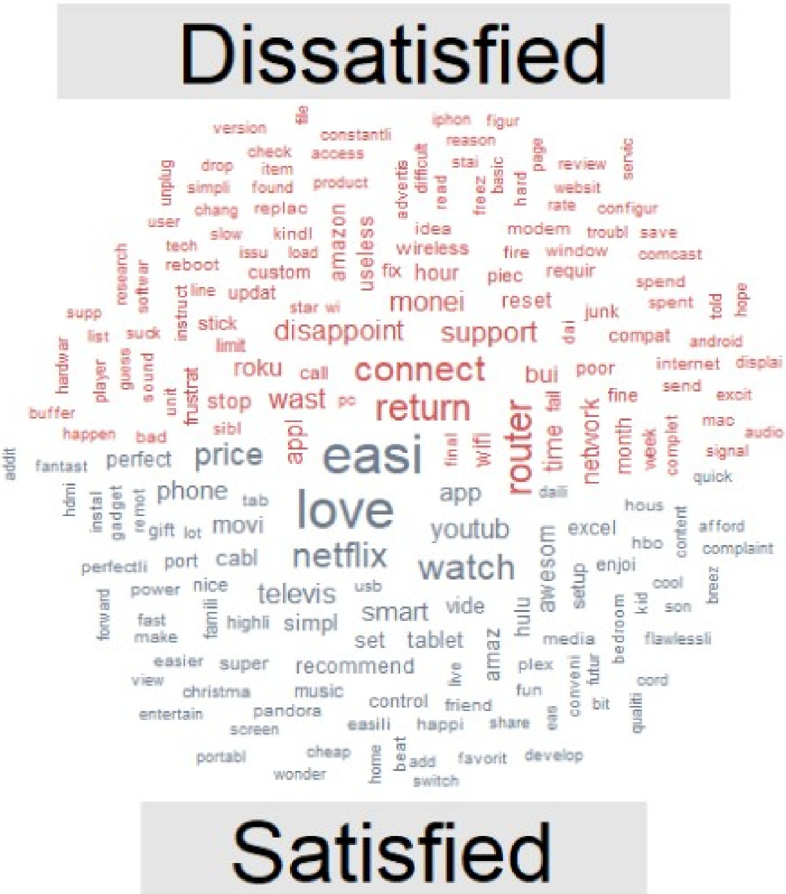
This section helps the reader to get more familiar with the data and focuses on the differences between satisfied and dissatisfied reviewers. First, it provides an overview of some descriptive statistics. Thereafter, I will describe the most frequent words to be mentioned in satisfied and dissatisfied reviews.

The data has a sample size of 8,938 and originates of respondents from the United States. As mentioned, rating is used to differentiate between satisfied and dissatisfied reviewers and therefore has a binary design. The rating distribution is not divided equally, as only 14.9% of the reviews have a one star rating, whereas 85.1% a five star rating. Moreover, the number of words in a cleaned review range from 2 to 99 with an average of 10.7 words per review.

Figure 4.1 shows a comparison cloud with the most common words in the dataset. As the name suggests, it compares the word count of a word between two classes i.e. satisfied and dissatisfied. The cloud shows the 200 most frequent words from the review text. The words differ on two aspects, colour and size. For example, the word *connect* is red and has a relatively large size. This means that *connect* could occur relatively frequent in both satisfied and dissatisfied reviews, but relatively more in dissatisfied reviews. However, the place of the word relative to other words in the figure do not have additional meaning. The figure shows that positive words, *love, easi, perfect, recommend, perfect, gift* and *fast* are all mentioned relatively more in satisfied reviews. furthermore, the negative words,

Disappoint, return, wast, useless, junk and *frustrat* are relatively used more in dissatisfied reviews. It is also expected that positive words are represented more in satisfied reviews and negative words in dissatisfied reviews. However, a closer look into the distribution of more neutral words is also interesting. The words *connect, return, support, time, network* and *money* are relatively more mentioned in dissatisfied reviews. This could imply that dissatisfied reviewers cannot connect the device or think it is not worth the money. The words *watch, Netflix, youtub, price* and *televis* are more common in satisfied reviews. These reviewers might appreciate the price of the product or like to use the product to watch Netflix or YouTube.

Figure 4.1: Comparison cloud with the 200 most common words. The red words occur relatively more in dissatisfied reviews and the blue words relatively more in satisfied reviews. Larger words occur more often in all reviews combined compared to smaller words.



5. Results & interpretation

This chapter provides the reader with an outline and interpretation of the main results that were used to answer the research question. First, I will discuss the logistic regression analysis along with an interpretation of the results. The next section describes the features from linear regression that were positively related to the emotions joy and disgust. The chapter ends with a topic modeling approach, to identify hidden topics in the review text.

5.1 logistic regression analysis

Recall that logistic regression performed well in describing the relationship between a categorical outcome variable and one or more continuous or categorical predictors. The outcome variable for this model has two classes, satisfied and dissatisfied. The reviews with a five star rating were classified as satisfied and reviews with one star as dissatisfied. The purpose was to find features that drive people to be either very positive or negative about the product. Therefore, reviews with modest ratings, between two and four stars, were excluded from the sample. This group of reviewers tend to be more neutral and therefore less likely to differentiate well between either positive or negative. As mentioned in section 3.3 Variable creation, there were four different types of predictors. These were single words, bigrams, emotions and factors. The model included 75 single words, 29 bigrams, 10 emotions and 10 factors. The number of factors in the model were chosen as follows. According to section 4.3, the number of factors (dimensions) usually depends on a scree plot of relative eigenvalues. A lower amount of factors eases the interpretation, if the goal is to visualize the dimensions of the data. However, I intended to use factors as predictors in the model, where each factor contained a group of words with large correlations. A model with only two factors groups all words according to these two factors. Such a model tends to lose information, as it does not distinct separate groups of words in a separate factor. Therefore, choosing 10 factors leads to more separated groups of words i.e. facilitates more interpretable factors.

Table 5.1 below presents the output of the significant variables of the logistic regression. Remember that the dependent binary variable is 0 for dissatisfied reviews and 1 for satisfied reviews. This implies that a negative coefficient tends to increase the probability of a more dissatisfied reviewer. I have only used the sign and not the magnitude of the coefficients for interpretation. The magnitude is not useful, as the effect of a single word on the outcome variable is relatively low and depends on the other words in the review. For example, when you compare two identical reviews, but one review contains one extra word, then the probability that a reviewer perceives the product experience as positive would increase/decrease with the log(odds) of the coefficient. Furthermore, this paper is only interested in finding features that differentiate between satisfied and dissatisfied reviewers and does not attach added value to the effect size of an increase/decrease in probability of a certain word.

Table 5.1 logistic regression results for the relationship between the significant textual features and rating in Google Chromecast reviews (sample size 8,938)

	Dependent variable
	Rating
	Logistic regression
Variable	**Coefficient**
intercept	0.022
factor 6	0.008*
anticipation	(0.014)**
disgust	(0.105)**
fear	0.023*
joy	0.097**
sadness	(0.024)*
negative	(0.005)**
support	(0.003)*
roku	0.003*
android	0.520*
hbo	(0.360)**
love+easi	0.992*

Note. The coefficients of the independent variables are probabilities; * $p<0.05$, ** $p<0.01$.

Table 5.2 PCA results of the words with largest component loadings of factor 6

connection	**laptop**	**setup**	**easy**	**wireless**	**network**	**issue**	**port**	**USB**	**cable**
0.37	0.33	0.30	0.30	0.27	0.25	-0.15	-0.14	-0.12	-0.11

Factor 6 has a positive coefficient of 0.008. Table 5.2 shows the words and component loadings from factor 6. Recall that a positive component loading indicates a positive correlation with the factor and a negative loading a negative correlation. The terms *connect, laptop, setup, easy, wireless* and *network* are all positively correlated to factor 6. This means that reviews mentioning these words tend to be more positive about their product experience than reviews that do not mention these words. For example, a reviewer that uses the words *easy connection* or *easy setup* tends to be more positive than reviewers that do not mention these words. The terms with a negative loading, *issue, port, USB* and *cable* are slight negatively correlated, indicating reviews that have mentioned these words tend to be slightly more likely to have a negative product experience.

The single word *android* has a positive coefficient. For example, reviewer number 10,596 mentions “have been using it for several months with my android phone and plasma tv. simply love it.” This suggests that people are satisfied with the fact that the product can connect with android devices. The word *support* has a significant negative coefficient. This indicates that a reviewer that mentions the word support is more likely to be dissatisfied regarding its product experience compared to an identical review that has not mentioned the word support. A reason of this finding could be that people tend to dislike the amount of apps that support the Google Chromecast. There are also significant words, *roku* and *hbo,* which are very hard to interpret individually and require more context for interpretation. This will be discussed in section 5.3.

The emotions joy and disgust are the most revealing variables, as these features have the largest positive and negative coefficients respectively. Moreover, the variables represent emotions, which means that they consist of multiple words from the review text. Reviews that mention words corresponding to the disgust emotion tend to be more dissatisfied. Since disgust is a very negative emotion, it is also expected to be mentioned by unhappier reviewers. Moreover, people that mention joyful words in their review tend to feel happier about their product experience. For instance, reviewer 9,972 mentions *“I am enjoying Netflix through my tv flawlessly. I am looking forward to seeing more apps that will support Google Chromecast.”* Despite the positive sense of this review, the app support of the product is also a point of discussion among positive reviewers.

5.2 Explanation joy and disgust

As mentioned in section 5.1, The emotions joy and disgust significantly differentiate between satisfied and dissatisfied reviewers. However, it is still unclear what features from the review text drive people to feel disgusted or joyful about their product experience. Therefore, the paper uses two techniques to elaborate further on this matter.

The first technique is a linear regression with a measure of joy/disgust as dependent variable and the same features from the logistic regression as independent variables. To do this, I added two new variables that account for the share of disgustful (joyful) words in the review. These variables are called *ratio_disgust* and *ratio_joy* and are measured by dividing the number of disgustful (joyful) words by the number of words in the corresponding review. *ratio_disgust* ranges from 0 to 0.333 and *ratio_joy* from 0 to 0.667. Thus, the variables are ratios, which makes them both continuous. Therefore, I estimated two separate linear regressions with *ratio_disgust* and *ratio_joy* as continuous outcome variables and the same features from the logistic model as predictors.

Table 5.3 below shows the significant regression results of the emotions joy and disgust. The non-significant variables are not used for interpretation and therefore not shown. The paragraph starts with the interpretation of *ratio Joy*. This model has an R-squared of 0.06, which means that 6% of the variance is explained by the variables in the model. The variable *Number of Words* has a coefficient of -0.009 and tends to be negatively related to *ratio Joy*. This indicates that a longer review tends to have a lower ratio of joyful words.

The single words *easy, fast* and *pretty* all have positive coefficients. This was also expected, as these words are associated with positive emotions. The single word *time* has a significantly negative coefficient. For example, a group of reviewers mention that it took them a substantial amount of time to connect the device with their television, whereas others write “this is garbage and not worth your time and money”. At last, factor 6 has a significant positive coefficient. The interpretation of factor 6 in table 3 and its paragraph below also applies to *ratio Joy*.

Table 5.3 Linear regression results for the relationship of the significant textual features and the ratio of joyful and disgustful words in Google Chromecast reviews (sample size is 8938)

	Dependent variable:	
	Ratio Joy	Ratio Disgust
	OLS	*OLS*
Variable	**Coefficient**	**Coefficient**
Intercept	0.053	0.018
Number of Words	(0.009)**	
factor 6	0.037*	
easy	0.023*	
time	(0.037)*	
update	(0.007)*	
fast	0.008*	
advertise	0.007*	
pretty	0.016*	
mirror		0.021*

love		(0.041)*
send		0.003*
power		(0.005)*
bedroom		0.003**
R-squared	0.06	0.03
Adjusted R-squared	0.04	0.01

Note. The coefficients of the independent variables are in percentages; * p<0.05, ** p<0.01

The model of *ratio Disgust* has an R-squared of 0.03. The word *mirror* refers to a recurring issue in the reviews. The Chromecast tends to mirror every action from your mobile phone, including the incoming of messages. Sometimes these messages are private and not intended for others. Many reviewers also mention that Roku, close competitor, does not have this disadvantage. Furthermore, the word *love* has a strong negative effect. This is logical, as people that mention love in their review are less likely to feel negative about their product experience.

5.3 topic modeling on emotions

In addition to low R-squares, the main disadvantage of interpreting these regression results is that it ignores the context around the words. The meaning of a word is typically described be the words around it i.e. the context. More context is especially necessary for the words that are more neutral, such as *bedroom* and *send* from the *ratio Disgust* regression. Therefore, the second technique, Latent Dirichlet Allocation, has provided the reader with more context. Recall from section 4.2 that LDA is unsupervised and designed to find hidden topics in text, where topics are a concentration of words that occur relatively often together in reviews. Furthermore, it follows a Dirichlet distribution, meaning that multiple topics can occur in different documents and words can occur in multiple topics.

The purpose is to identify topics that are discussed among reviewers that feel joyful (disgustful) about their product experience. Therefore, the input data for LDA is a subset of the original dataset and filtered on three criteria. First, the *Ratio Joy* variable must exceed 0.07 in a review, so more than 7% of all the words from a review have to be from the Joy-lexicon. Here, I have chosen 7% to have at least 750 remaining reviews. Second, the one star reviewers that passed the 7% mark were removed, as these reviews were perceived is sarcastic, random or misinterpreted. Third, the words from the Joy-lexicon of NRC were overrepresented in the remaining data and therefore removed. These words, such as love, are not informative, as it is logical that joyful reviews mention joyful words. The same filtering criteria apply to the input data for the Disgust emotion.

Furthermore, LDA requires a Document term matrix as input and two hyperparameters have to be optimized, the number of topics k and a control for topic sparsity α. Table 5.4 below shows the perplexity values on the validation set for different number of topics *k* and topic sparsity *α*. The joy dataset uses 25 topics with an α of 0.06, whereas the disgust dataset selects 25 topics with an α of 0.07.

Table 5.4: LDA perplexity values on validation set for the parameters k and α

k	**Perplexity Validation**		**α**	**Perplexity Validation**	
	joy	*disgust*		*joy*	*disgust*
5	182.69	250.43	0.01	172.41	202.35
10	182.05	240.52	0.02	159.99	190.23
15	182.01	233.56	0.03	156.96	187.11
20	181.97	228.35	0.04	156.53	186.58
25	**180.60**	**227.75**	0.05	154.32	185.35
30	183.60	230.69	0.06	**152.94**	185.03
35	185.69	230.97	0.07	160.72	**184.67**
40	188.26	231.56	0.08	158.55	185.13
45	190.82	232.54	0.09	157.37	185.23
50	193.69	232.87	0.10	158.99	186.69

Note. the lowest perplexity values are in bold.

Table 5.5 shows a selection of the five most informative topics per emotion and its corresponding 8 terms with the largest word-topic probability. A word-topic probability of 15% means that the corresponding word explains 15% of the topic[CITATION Sll17 \l 1043]. Intuitively, all word-topic probabilities sum to 1 for each topic. An important observation about the word-topic distribution is that particular words, such as *watch* in topic 3 & 14, are common within both topics. This is a result of LDA’s soft-clustering approach, as mentioned in section 4.2. Furthermore, the document-topic probability is helpful to detect the documents (reviews) that explain a large proportion of the corresponding topic.

Table 5.5 LDA results with five most interesting joyful topics with the eight stemmed terms with largest topic probability in Google Chromecast reviews

				Topic:					
functionalities		**delivery service**		**moment of purchase**		**connecting devices**		**price perception**	
watch	0.17	fast	0.20	son	0.18	stream	0.12	buy	0.18
netflix	0.16	ship	0.14	gadget	0.09	cable	0.11	worth	0.11
youtube	0.13	product	0.13	christmas	0.08	ipad	0.09	price	0.09
movie	0.11	delivery	0.09	cheap	0.07	apple	0.09	inexpensive	0.08
purchase	0.05	service	0.05	internet	0.07	galaxy	0.06	roku	0.07
stream	0.05	quality	0.04	easy	0.06	service	0.06	play	0.07
fast	0.04	absolute	0.04	connect	0.05	support	0.05	connect	0.03
quality	0.04	quick	0.03	law	0.05	watch	0.04	gadget	0.03

This paragraph functions as an overview of the most interesting topics from the emotion Joy. The topic *functionalities* could explain tasks for which people use the product. the words *watch, Netflix, youtube, movie* and *stream*, could mention that people use the Chromecast to watch Netflix movies or stream via youtube. Moreover, *fast* could refer to the streaming speed. The next topic is *delivery service*. The words *fast, shipping, product, delivery* state that people with a joyful product experience are satisfied about the delivery service. The words *son, gadget, Christmas* and *cheap*, from the topic *moment of purchase* indicate that people tend to perceive the Chromecast as a cheap Christmas gadget for their son. The next topic, *connecting devices,* mentions different devices that can connect with the Chromecast. The words, s*tream, ipad, apple, galaxy* state that people use the Chromecast to stream with their apple or galaxy devices. Moreover, the words, *app* and *support* could either mean that the Chromecast is supported by a substantial amount of apps or that the app support could be improved. For example, reviewer 2,092, who mentions “couldn’t be happier with the chromecast”. Only hope that more apps support it like fox sports go.”, hints to the latter. The final topic, *price perception*, describes whether people perceive the product as good value-for-money. The words, *buy, worth, price, inexpensive,* point to a good value-for-money perception. Moreover, *buy, price, roku* could provide information of how joyful people perceive the Chromecast compared to its close competitor Roku.

Reviewer 10,761, who mentions “goodbye roku and apple tv, hello chromecast. Our third device in the family. Excellent product to buy for entertainment and excellent price point.”, perceives Chromecast as an improvement with a sharp price.

Table 5.6 LDA results with five most interesting disgustful topics with the eight stemmed terms with largest topic probability in Google Chromecast reviews

				Topic:					
Recurring issues		**mirror function**		**Discourage**		**disgusted reviewer**		**app support**	
movie	0.08	mirror	0.08	quality	0.10	waste	0.15	apple	0.10
watch	0.07	function	0.08	expect	0.07	money	0.11	support	0.09
constantly	0.05	computer	0.05	bad	0.07	time	0.09	view	0.08
buffer	0.04	android	0.05	lag	0.05	buy	0.07	roku	0.04
reboot	0.04	limit	0.05	picture	0.05	apple	0.07	android	0.04
time	0.04	support	0.04	video	0.04	expect	0.04	issue	0.03
reset	0.04	buy	0.03	recommend	0.04	plug	0.03	absolut	0.02
phone	0.03	stick	0.03	roku	0.03	freeze	0.03	recommend	0.02

The next paragraph provides an overview of the most interesting topics from the disgust emotion from table 5.6. The first topic describes reviews with recurring problems. The words, *movie, watch, constantly* and *buffer* signal that the Chromecast constantly buffers during movies for reviewers that have a large document-topic probability on this topic. Moreover, it tends to take much time for these reviewers to reset and reboot the Chromecast. Reviewer 1,057 says “this was a total waste of money. had to constantly reboot or reset it to get to even work. constantly buffering. the wait time just to watch a movie load up after all of the rebooting and reacting was so frustrating, even though I have a good wifi signal in my house.” The topic, *mirror function*, describes a very specific product characteristic. As mentioned in the last paragraph of section 5.2, the Chromecast has a function that mirrors every action from a mobile device or tablet to the television, including private messages. Moreover, people mention the combination of *mirror* and *function* mainly in the disgustful topics. Therefore, it can be seen as a disadvantage of the Chromecast. The topics, *discourage* and *disgusted reviewer,* are very superficial and intend to warn potential purchasers. The words *expect, bad, quality* and *roku* have a discouraging sense and may point to buy Roku instead. The same accounts for the words *waste, money, time, buy* and *apple,* where the reviewers recommend to purchase an Apple tv. The final topic, *app support*, could relate to the number of apps that support the Google Chromecast. The words *support, view, issue* state that there are issues to view certain apps, as they are not supported by the Chromecast. Moreover, the words *android* and *apple* in combination with *support,* could suggest that both Android and Apple devices have issues regarding the app support. For example, reviewer 9,601 mentions “what were you thinking google? good idea, bad execution. there is very little use for this product, since not that many channels are supported.”

6. Discussion & conclusion

This chapter discusses the main results and answers the research question. Moreover, the implications for managers section will describe what type of results can be used for different practical purposes. The chapter finishes with limitations of the research followed by suggestions for further research.

6.1 main results and answer on RQ

In this study, a number of models regarding the relation between features from customer reviews and the product rating have been analysed and interpreted. This section will highlight and elaborate on the most important findings regarding the Google Chromecast streaming device.

Overall, the emotions Joy and Disgust were perceived as the most important features to differentiate between satisfied and dissatisfied reviews. Furthermore, factor 6 with words as *connection, laptop, setup, easy* and *wireless* has been positively related to rating. The next regression models analysed the impact of the original features on the outcome variables *ratio Joy* and *ratio Disgust.* The terms *time* and *update* were negatively associated with *ratio joy* and factor 6 showed a positive association. Moreover, reviews with the word *mirror* tended to be negatively related to *ratio Disgust,* whereas the word *bedroom* had a slightly positive relation. Lastly, the topic modeling approach LDA has shown that people tend to feel more joyful, because of the product’s delivery service, functionalities and price quality perception. Moreover, the Christmas period was perceived as a great moment to purchase the product and people mainly used the product to watch Netflix movies or stream YouTube videos. Furthermore, reviewers tended to feel more disgustful, because of the product’s mirror function, recurring connection issues and the lack of app support.

6.2 implications for managers

The findings of this study have several practical implications for managers in the marketing field. managers could use the logistic regression coefficients on rating as a first step in the analysis. As mentioned in section 5.1, the impact of a single word on the outcome variable is relatively low in a model with textual features. This suggests that a manager should not use the outcome of the analysis to make bold statements, but rather use it as statistical proof for further analysis. From the logistic regression output, one could conclude that the categories single words and emotions are useful for further analysis, as these categories have multiple significant features. Furthermore, the emotions Joy and Disgust showed the largest positive and negative relation to the outcome variable.

Recall from section 2.6 that Huang, Lurie & Mitra (2009) argued that a review's product ratings and sentiment could have different roles in the customer decision journey. They stated that detailed information about the sentiments or product experiences is more difficult to evaluate than the average product rating. Therefore, emotions could be more useful in later phases of the customer decision journey i.e. closer to the final decision. This suggests that a further analysis on the determinants of a reviewer's emotion could improve insights in drivers of the customer purchase decision. This is essential information for managers, as the new marketing strategy is customer driven i.e. focused on meeting customer preferences [CITATION Jos19 \l 1043]. Therefore, additional information regarding features that tend to differentiate well between several emotions, could help a manager to customize its product offer. For example, a manager could compare the analysis results over time, following the same methodological steps. One could then identify changing customer behaviour, based on different significant words or topics. These insights could help to improve alignment of future versions of the product with the changing customer preferences.

The linear regressions with continuous outcome variables *ratio Joy* and *ratio Disgust* tend to describe a positive and negative emotion respectively. Therefore, managers should look for features that describe either tangible or intangible product characteristics for which the sign of the relation to the outcome variables is previously ambiguous. For example, features such as *easy* and *fast* are not so useful, as these features were expected to have a positive effect on *ratio Joy* or negative effect on *ratio Disgust*. However, the negative coefficient of the term *mirror* on the outcome variable *ratio Disgust* is interesting. The regression output does not provide context other than the word itself. A manager could therefore filter reviews that mention the term *mirror* in the dataset, to identify the reasons behind the negative relation with *ratio Disgust*. Another advantage of the filter approach is that one could easily measure the ratio of negative reviews that mention the word *mirror* relative to the positive reviews. For instance, a large ratio suggests that the majority of reviewers perceive the mirror function as a negative feature. In this paper, multiple reviews explicitly mentioned that the mirror function also streamed private messages from their phone to the screen. Though, others were very positive about the mirror function as a solution for a streaming task that was not supported by an app. This is a great insight for managers in the process of product evaluation and development.

LDA topic modeling has been conducted to identify topics in documents that were shared by a group of reviewers. Recall that textual documents (reviews) have a distribution of topics and these topics a distribution of words from the list of unique words. A manager can use topic modeling to extract information in two steps. Firstly, the distribution of words per topic could reveal certain topics that are either beneficial or harmful for the company. Secondly, the distribution of topics per document could reveal what reviewers tend to share the opinion on the concerned topic. Documents with a large probability on a certain beneficial topic can be clustered and used as promoters of the product, whereas documents on a certain harmful topic can be clustered as distractors. Here a cluster means that reviews in the same cluster are treated similarly. For example, documents with a large topic probability for the harmful topic *app support* could be clustered under the name *distractors*. A manager could then make a profile of the distractors and target them on social media, once new apps are added that support the product. This could improve their product perception or prevent them to spread more negative WOM in their social circle or online. Furthermore, the company can contact these people and ask which apps they would like to be supported by the product in the future. The company could then consider to collaborate with the apps that are in great demand and have a positive net present value.

6.3 limitations & suggestions further research

This section discusses several limitations of the paper followed by suggestions for improvement in further research.

First of all, the input data is not particularly designed for text analytics and specifically to identify the features from review text that differentiate well between satisfied and dissatisfied reviewers. I would suggest managers of e-commerce businesses to adjust the data collection process of reviews in two different ways. Firstly, the process of writing a review could be adjusted to a stepwise approach, to improve the accuracy of the topic modeling task. One could start asking the customer to select a certain amount of emotions from Mohammad's NRC list (2019) that correspond to their product experience. This allows to group reviews based on the individual's emotion. This could ease taking subsets of data for a topic modeling task. The second step is then to ask customers to explain their stated emotion(s) from the first step. By using this two-way approach, managers can self-select the reviews about emotions they are interested in combination with arguments that explain such an emotion. The second way to adjust the data collection of reviews is to add demographic information about the reviewer. As discussed in section 6.2, a manager could cluster documents based on certain beneficial or harmful topics with a large topic probability in these documents. Such a cluster can then be used for targeting purposes. However, it is not feasible to target certain clusters of reviewers, without having detailed information about them. Therefore, I would suggest further researchers, with targeting purposes, to add demographic information about reviewer's age, gender, household size and/or annual income.

As mentioned in the last paragraph of section 4.2,, the models for this paper have a Document-Term matrix with the bag-of-words model as input. The bag-of-words is focused on detecting co-occurrences between words of different reviews. However, many reviews only retained a limited amount of words after data cleaning, which has resulted in many sparse-entries. Therefore, the models have had more difficulties to find co-occurrences in the text, as co-occurrences only arise from non-sparse entries. To tackle this problem, I have chosen to exclude all the reviews that had less than three remaining terms, after cleaning for the topic modeling tasks. A stricter criteria would result in a too low sample size. I would suggest future researchers to take this limitation into account during the data gathering process. For example, the dataset for this paper originally contained 16.000 reviews, but only 237 were retained in the disgust dataset. This suggests that a large dataset does not necessarily mean that all reviews are equally useful. Furthermore, future researchers could investigate whether the meaningfulness of the results tend to improve when the minimum number of words criteria becomes stricter.

Another point of emphasis is that product reviews might be more informative for experience goods than search goods. The Google Chromecast is a typical search good, as full information about the product attributes can be acquired before purchase. The added value of reviews could be larger for experience goods, as these tend to provide more information about the product experience that cannot be found online before purchase. Therefore, it might be interesting to perform the same analysis on an experience good and compare the quality of the results.

The last limitation concerns the logistic regression approach on review text. This is because it is very difficult to estimate the true effect of a variable on the change in probability on the outcome variable. Furthermore, it is hard to compare the impact of different coefficient values on the outcome variable, as the impact of a variable's individual effect within a review also depends on other words in the review. Therefore, I would only suggest to use the output from regression results for exploratory purposes, by using the sign of the coefficients.

Bibliography

Airoldi, Blei, Erosheva, & Fienberg. (2015). Handbook of mixed membership models and their applications. *CRC Press* .

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

Bae, & Lee. (2011). Product type and consumers’ perception of online consumer reviews. *Electronic Markets*, 255-266.

Berger , J. (2016). *Contagious: Why things catch on.*

Berger, & Milkman. (2012). What makes online content viral? *Journal of Marketing Research*, 192-205.

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of consumer psychology*, 586-607.

Bettman, Luce, & Payne. (1998). Constructive Consumer Choice Processes . *Journal of Consumer Research*, 187-217.

Blei. (2012). Probabilistic topic models. Surveying a suite of algorithms that ofer a solution to managing large document archives. . *Commun ACM*, 77-84.

Blei, & Jordan. (2003). Modeling annotated data. *SIGIR’03*.

Brett, M. (2012). *Topic Modeling: A Basic Introduction.* Retrieved from journal of digital humanities: http://journalofdigitalhumanities.org/2-1/topic -modeling-a-basic-introduction-by-megan-r-brett/

Brown, Broderick, & Lee. (2007). Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of Interactive Marketing*, 2-20.

Brownlee, J. (2016, March 21). *Overfitting and Underfitting With Machine Learning Algorithms.* Retrieved from Machine learning mastery: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Brownlee, J. (2017, October 9). *A Gentle Introduction to the Bag-of-Words Model.* Retrieved from www.machinelearningmastery.com: https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Bughin, Doogan, & Vetvik. (2010, April). *A new way to measure word-ofmouth marketing.* Retrieved from McKinsey: http://designdamage.com/wp-content/uploads/2009/04/A-new-way-to-measure-WOM-marketing.pdf

CBS. (2015, 03 12). *9 in 10 people access the internet every day*. Retrieved from CBS: https://www.cbs.nl/en-gb/news/2015/11/9-in-10-people-access-the-internet-every-day

Cheung, Lee, & Rabjohn. (2008). The impact of electronic word-of-mouth - The adoption of online opinions in online customer communities. *Internet Research*, 229-247.

Cheung, Luo, Sia, & Chen. (2009). Credibility of electronic word-of-mouth: Informational and normative determinants of online consumer recommendations. *International Journal of Electronic Commerce,* , 9-38.

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

Court, D., Elzinga, D., Mulder, S., & Vetvik, O. (2009, june). *David Court, Dave Elzinga, Susan Mulder, and Ole Jørgen Vetvik*. Retrieved from McKinsey & Company: https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-consumer-decision-journey

Decker, & Trusov. (2010). Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, 293-307.

Dholakiya, P. (2017, 12 15). *23 Facts You Should Know About Online Reviews*. Retrieved from Search Engine Journal: https://www.searchenginejournal.com/online-review-facts/227829/#close

Duan, W., Gu, B., & Whinston, A. (2008). The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry. *Journal of Retailing*, 233-242.

Farrar, D., & Glauber , R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*, 92-107.

Feng, Y., Wu, C.-m., & Deng, W.-j. (2009). The relationships among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services. *Computers in Human Behavior*, 887-896.

Flanagin, Metzger, Pure, Markov, & hartsell. (2014). Mitigating risk in ecommerce transactions: perceptions of information credibility and the role of user-generated ratings in product quality and purchase intention. *Electronic Commerce Research*, 1-23.

Forman, Ghose, & Wiesenfeld. (2008). Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. *Information Systems Research*, 243-396.

Gavilan, D., Avello, M., & Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. *Tourism Management*, 53-61.

Ghose, & Ipeirotis. (2007). Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews. *Proceedings of the ACM International Conference on Electronic Commerce.*

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

Godes, D., & Mayzlin, D. (2004). Using Online Conversations to Study Word-of-Mouth Communication. *Marketing Science*, 469-631.

Google Chromecast. (2019). Retrieved from Google store: https://store.google.com/product/chromecast

Grewal, D., Gotlieb, J., & Marmorstein, H. (1994). The Moderating Effects of Message Framing and Source Credibility on the Price Perceived Risk Relationship. *Journal of Consumer Research*.

Grün, & Hornik. (2018, december 21). *topicmodels.* Retrieved from https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf

Haberman, S. (2014). *Analysis of qualitative data.* Academic Press.

Hennig, Gwinner, Walsh, & Gremler. (2004). Electronic word of mouth via consumer opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, 38-52.

Hennig, Gwinner, Walsh, & Gremler. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, 38-52.

Hernández, B., Jiménez , J., & Martín, J. (2010). Customer behavior in electronic commerce: The moderating effect of e-purchasing experience. *Journal of Business Research*, 964-971.

Hu, Koh, & Reddy. (2014). Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. *Decision Support Systems*, 42-53.

Hu, Pavlou, & Zhang. (2006). Can online reviews reveal a product's true quality?: Empirical findings and analytical modeling of online word-of-mouth communication. *Electronic Commerce* .

Huang, Lurie, & Mitra. (2009). Searching for experience on the web: an empirical examination of consumer behavior for search and experience goods. *Journal of Marketing*, 55-69.

Jaadi. (2019, Februari 28). *A step by step explanation of Principal Component Analysis.* Retrieved from Towards data science: https://towardsdatascience.com/a-step-by-step-explanation-of-principal-component-analysis-b836fb9c97e2

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to statistical learning.* Springer.

James, Witten, Hastie, & Tibshirani. (2013). *An Introduction to Statistical Learning.* Springer.

Jiang, Yang, & Jun. (2013). Measuring consumer perceptions of online shopping convenience. *Journal of Service Management* .

Jimenez, & Mendoza. (2013). Too Popular to Ignore: The Influence of Online Reviews on Purchase Intensions of Search and Experience Products. . *Journal of interactive marketing*, 226-235.

Johnson, & Kaye. (2002). We believability: a path model examining how convenience and reliance predict online credibility. *Journalism and Mass Communication Quarterly*, 865-879.

Joseph, C. (2019, february 12). *Elements of a Customer-Driven Marketing Strategy.* Retrieved from smallbusiness.chron: https://smallbusiness.chron.com/elements-customer-driven-marketing-strategy-3223.html

Jun, Yang, & Kim. (2004). Customers' perceptions of online retailing service quality and their satisfaction. *International Journal of Quality & Reliability Management*, 1149-1174.

Kahneman, & Tversky. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 1124-1131.

Kahneman, & Tversky. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 297-323.

Kamins, M., & Lawrence, M. (1988). An examination into the effectiveness of two-sided comparative price appeals. *Journal of the Academy of Marketing Science*, 64–71.

Kim, D., & Benbasat, I. (2003). Trust-related arguments in internet stores: a framework for evaluation. *Journal of Electronic Commerce Research, 4*, 49-64.

Kleindorfer, Wind, & Gunther. (2009). *The Network Challenge: Strategy, Profit, and Risk in an Interlinked World.*

Lakeland, C. (2017). *Why do we remove frequent and infrequent words when in NLP?* Retrieved from Quora: https://www.quora.com/Why-do-we-remove-frequent-and-infrequent-words-when-in-NLP

Lee, & Bradlow. (2011). Automated Marketing Research Using Online Customer Reviews. *Journal of Marketing Research*.

Lemon, K., & Verhoef, P. (2016). Understanding Customer Experience throughout the customer journey. *Journal of Marketing*, 69-96.

Li, & Hitt. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 456-474.

Lim, Lwim, & Whee. (1995). Word-of-mouth communication in Singapore: With focus on effects of message-sidedness, source and user-type. *Asia Pacific Journal of Marketing and Logistics*, 5-36.

Mackiewicz. (2007). Customer knowledge management and e-commerce: The role of customer perceived risk. *Association for Business Communication Annual Convention.*

Maheswaran, & Sternthal. (1990). The Effects of Knowledge, Motivation, and Type of Message on Ad Processing and Product Judgments. *Journal of Consumer Research*, 66-73.

Manchanda, P., Dubé, J.-P., Goh, K., & Chintagunta, P. (2006). The Effect of Banner Advertising on Internet Purchasing. *Journal of Marketing Research*, 98-108.

Martonik, A. (2015, January 29). *Chromecast is now the No. 1 streaming device in the U.S., users have casted 1 billion times.* Retrieved from Android central: https://www.androidcentral.com/chromecast-now-number-1-streaming-stick-us-users-have-casted-1-billion-times

Mavlanova, Koufaris, & Fich. (2012). Signaling theory and information asymmetry in online commerce. *Information & management*, 240-247.

McAuley, J., & Leskovec, J. (2013). Hidden Factors and Hidden Topics:. *RecSys*, 165-172.

Mers, J. (2014, October 20). *Quality Over Quantity: The Overblown Importance Of Likes And Followers.* Retrieved from Forbes: https://www.forbes.com/sites/jaysondemers/2014/10/20/quality-over-quantity-the-overblown-importance-of-likes-and-followers/#26d6c9d84d44

Mishra, Heide, & Cort. (1998). Information Asymmetry and Levels of Agency Relationships. *Journal of Marketing Research*.

Moe, & Trusov. (2011). The Value of Social Dynamics in Online Product Ratings Forums. *Journal of Marketing Research*.

Mohammad, S. (n.d.). *NRC Word-Emotion Association Lexicon.* Retrieved from www.saifmohammad.com: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Nelson. (1970). Information and consumer behavior. *Journal of Political Economy*.

Neslin, S., Grewal, D., Leghorn, R., Shankar, V., Teerling, M., Verhoef, P., & Thomas, J. (2006). Challenges and Opportunities in Multichannel Customer Management. *Journal of Service Research*.

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

Nyxone. (2017). *Importance of E-Commerce and online shopping and why to sell Online*. Retrieved from Medium: https://medium.com/@nyxonedigital/importance-of-e-commerce-and-online-shopping-and-why-to-sell-online-5a3fd8e6f416

Park, Lee, & Han. (2007). The effect of online consumer reviews on consumer purchasing intention: the moderating role of involvement. *. International Journal of Electronic Commerce*, 125-148.

Pavlou, & Dimoka. (2006). The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation. *Information Systems Research*, 327-444.

Peng, Lee, & Ingersoll. (2010). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 3-14.

Peng, Manz, & Keck. (2001). Modeling categorical variables by logistic regression. *American Journal of Health Behavior*, 278-284.

Puccineli, N., Goodstein, R., Grewal, D., Price, R., Raghubir, P., & Stewart, D. (2009). Customer Experience Management in Retailing: Understanding the Buying Process. *Journal of Retailing*, 15-30.

Reisenbichler, M., & Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, 327-356.

Rosen, S. (1974). *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.* Harvard University.

Schlesselman, J. (1982). *Case-Control Studies: Design, Conduct, Analysis.* Oxford University Press.

Shlens, J. (2014). A Tutorial on Principal Component Analysis. *Cornell University*.

SIlge, J., & Robinson, D. (2017). *Text Mining with R.* O'reilly.

stats, Internet world. (2019). *INTERNET GROWTH STATISTICS*. Retrieved from Internet world stats: https://www.internetworldstats.com/emarketing.htm

Sundaram, Mitra, & Webster. (1998). Word-Of-Mouth Communications: a Motivational Analysis. *Association for Consumer Research,* , 527-531.

Trusov, Bucklin, & Pauwels. (2009). Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site. *Journal of Marketing*, 90-102.

Varkaris, E. (2017). The influence of social media on the consumers' hotel decision journey. *JOURNAL OF HOSPITALITY AND TOURISM TECHNOLOGY*, 100-118.

Wathen, & Burkell. (2002). Believe it or not: Factors influencing credibility on the Web. . *Journal of The American Society for Information Science and Technology*, 134-144.

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in human behavior*, 634-639.

Zhang, & Watts. (2003). Knowledge adoption in online communities of practice. . *International Conference on Information Systems*, 81-102.

Zhang, Lee, & Zhao. (2010). Understanding the Informational Social. *International Conference on Information Systems.*