Discovering Aggregate Customer Preferences from Unstructured Reviews

Noah van Roekel
email: 448056nr@eur.nl

July 2021

Abstract

"With the rise of the quantity of user generated content (UGC), a lot of information on customers' preferences is freely available to companies. Finding a way to discover the aggregated customer preferences from UGC can help companies improve products and marketing efforts. Current literature focuses primarily on aspect-based sentiment analysis, learning product features and finding aggregated customer preferences from structured UGC. This paper proposes a novel method to uncover the aggregated customer preferences from unstructured UGC, using state-of-the-art techniques. Our approach requires minimal input from the practitioner which makes it widely applicable. The method that we propose uses Latent Dirichlet Allocation for attribute mining, a simple sentiment analysis step and a multinomial logistic regression to find the aggregated customer preferences. This paper suggests two informative visualizations. The final method is demonstrated using a dataset with almost 30,000 phone reviews from `Amazon`."

Contents

List of Figures

List of Tables

1 Introduction

Knowing customers' preferences can be very valuable, e.g., for marketing managers and product developers. With this knowledge, marketing managers can highlight specific attributes, and product developers know what to focus on with new products. Since knowing these preferences is valuable, much effort, in both research and the industry, has been put into finding these preferences. Consumer preferences are often discovered using survey-based methods, e.g., Adaptive Conjoint Analysis (ACA) (Johnson, 1985). However, new Natural Language Processing (NLP) techniques allow us to extract much information from User Generated Content (UGC). With the rise of the internet and e-commerce in the 21st century, the amount of UGC proliferated. A well-known version of UGC is customer reviews. These reviews contain a lot of information (Jin, Liu, Ji, & Kwong, 2019). This paper focuses on finding aggregated consumer preferences using UGC in the form of full-text online consumer reviews.

The current standard in discovering consumer preferences is using survey-based methods like conjoint analysis. However, Survey-based methods have some problems. One of the problems is that one needs respondents to fill in the survey. Another problem is that the practitioner needs to set out the product attributes in advance. A high amount of expertise about a product is required to know all the relevant attributes. However, it is not always clear what consumers consider product features in real life, let alone what features they find important. In practice, practitioners of conjoint analysis most often use "expert judgment" or "group interviews" to decide upon the features that they include in their study (Cattin & Wittink, 1982). Product attributes that are unknown or unclear to management or the interviewed group will be missing in the survey, and the practitioner will not know the importance of these attributes.

New NLP techniques allow us to extract much information from UGC, such as a consumer's review. A lot of the existing research into consumer-generated online reviews is done in sentiment analysis (SA) (e.g., Bagheri, Saraee, & De Jong, 2013; Garcia-Pablos, Cuadros, & Rigau, 2018; Song, Wen, Xiao, & Park, 2021). The goal of sentiment analysis is to discover whether the consumer sentiment, i.e., feeling, about a product or a product attribute is positive or negative. Knowing the consumer's sentiment towards a product attribute is already informative, e.g., for marketing managers. For example, it can help find what attributes of your product are not perceived as positively as you might want. This paper takes this a step further and takes the consumers' sentiments to find their preferences.

New modern methods of preference discovery are needed because of the rise of new NLP methodologies and UGC and the problems associated with survey-based preference measurements. In this paper, we, therefore, answer the following research question, whereby the outcomes are relevant for both managers and research:

– *How can we determine and visualize aggregate consumer preferences based on unstructured, full-text, online customer reviews?*

To be able to correctly interpret the results from the proposed method, it is vital that we know what drives people to write reviews and to know what types of attributes reviewers write about. To the best of our knowledge, current literature seems to disregard the possible effects of different motivations for writing reviews in the interpretation of results. The second step in the proposed process is determining the relevant product attributes of our product. This paper focuses on unsupervised methods because it is important that both researchers and marketing managers with no to little technical knowledge of the product can use this method. Using unsupervised

methods ensure that no labelled data is needed. The step after determining the product attributes is discovering the customers' sentiment regarding each attribute in their reviews. Fourth, we need to calculate the aggregate consumer preferences concerning the different attributes. Finally, we have to visualize the aforementioned aggregated consumer preferences structurally and informatively so that it is immediately clear to the user of our method what the preferences are. We can summarize this into the following sub-questions:

- *Why do people write reviews and what do they write about?*
- *How to determine what the relevant product attributes are?*
- *How to determine the sentiment of a review about an attribute?*
- *How to determine what the aggregate consumer preferences concerning the different attributes are?*
- *How to visualize the aggregated consumer preferences?*

In the next section, we give a theoretical framework. After that, we describe our proposed method in detail. In the section thereafter, we describe the data that we use to test our proposed method, and we share the results of the model on the data. Finally, we conclude the paper and give recommendations for further research.

2 Theoretical Framework

Before one can interpret the results of the model that we propose, it is important to know who writes reviews, why they write reviews and what the things are that they write about.

2.1 Why Do People Write Reviews?

The large number of reviews available online can be somewhat paradoxical. Writing reviews can take much time. Most of the reviews are written by laypeople who generally do not get an immediate reward for writing them. To interpret the results of our model, it is essential to know why these laypeople take the time to write the reviews. With that, it is vital whether and how that would influence the outcomes of the model.

Since people generally do not get an immediate reward for writing reviews, such as a monetary reward, the motivation for spending time to write a review has to come from something else. Multiple motivations have been introduced in literature for people participating in word-of-mouth (e.g., Hennig-Thurau, Gwinner, Walsh, & Gremler, 2004; Kovács & Horwitz, 2018). Reviews can be seen as a form of word-of-mouth (WOM), namely electronic word-of-mouth (e-WOM). eWOM and WOM are conceptually very close (Hennig-Thurau et al., 2004). The main differences between the two are that e-WOM is no longer face-to-face, synchronous and private as WOM is (J. Berger & Iyengar, 2013; Porter, 2017). The motivations found for participating in e-WOM are somewhat similar to the incentives for participating in WOM.

For the sake of summarizing and completeness, we split the motivations into two buckets in this paper; egocentric motivations and altruistic motivations.

One of the egocentric motivations for writing a review is conspicuous reviewing, as described by Kovács and Horwitz (2018). They argue that a reason for consumers to write a review is similar to the traditional conspicuous consumption of status-signalling goods introduced by Veblen (1899). According to Kovács and Horwitz (2018), people write reviews to showcase their purchases for an online audience whereby others can see the reviewer's status based on what they bought. This is a modern-day equivalent to parking an expensive car on the driveway instead of in the garage where somebody would typically park their vehicle. Lampel and Bhalla (2007) also suggest that status-seeking is a motivation of why people post their reviews. Dichter (1966) talks about a similar reason for people to engage in positive word-of-mouth, and he calls it self-enhancement whereby a person can get attention, show connoisseurship or seek encouragement or reassurance. This motivation for WOM seems to also apply to e-WOM.

Another egocentric motivation for participating in e-WOM is post-purchase advice-seeking from other consumers on the web (Hennig-Thurau et al., 2004). People might participate in e-WOM to acquire the skills needed to correctly understand, operate, improve, use, or patch up the product they review. People might believe that participating might help solve their problems better than anonymously reading the comments or reviews.

People might write reviews because they want to be heard by the company (Whiting, Williams, & Hair, 2019). This can be for three main reasons: wanting the company to listen to them, being heard, or wanting the company to understand them. In line with this, Harrison-Walker (2001) show that people post complaints about products on forums because the ease of using and identifying it compared to contacting the producer is much higher. The latter might also be the case for consumers writing negative reviews, hoping that the sellers try to resolve the problem.

Another egocentric motivation for writing a review might be to enhance the understanding of a topic (Peddibhotla & Subramani, 2007). A reviewer can sharpen her thoughts by writing the review and articulating her thoughts about a product, and that might lead to a better understanding of a topic or a product (Peddibhotla & Subramani, 2007).

People might also write reviews to become part of an online community. Reviewers might see this as a social benefit (Hennig-Thurau et al., 2004; McWilliam, 2000). Plant (2004) defines online communities as:

> *a collective group of entities, individuals or organizations that come together either temporarily or permanently through an electronic medium to interact in a common problem or interest space.* (p. 54)

Participating in review writing might help people become part of this community by enhancing their presence on the websites.

According to Wu (2019), people write reviews for enjoyment. Reviewers might get pleasure from writing reviews, primarily because of the opportunity of self-expressing. This reason for posting a review seems to be especially motivating for people who recently began writing reviews. Wang and Fesenmaier (2004) found that enjoyment was one of the most important reasons for travellers to participate in electronic word-of-mouth. Litvin, Goldsmith, and Pan (2008) state that people enjoy sharing their experiences and see it as a part of the pleasure of travelling. This effect can also be the case for people who buy products that they like, and that might lead to reviewing behaviour.

Peddibhotla and Subramani (2007) found that another reason for an individual to write a review is to develop their writing skills. An individual might want to become

better at writing copy, and writing a review can be an excellent way to develop these skills.

The last egocentric motivation for participating in WOM that we want to highlight is described by Sundaram, Mitra, and Webster (1998) and regards the need for customers to restore balance in their lives. The balance can be distorted by getting a strong positive or negative consumption experience. A reviewer can restore their internal balance by either venting negative emotions or expressing positive emotions, and this is thereby also applicable to e-WOM.

Besides egocentric motivations, people might write reviews because of altruistic motivations. A reviewer might care about other customers and potential customers and therefore might want to share their experiences because of internal values, regardless of other reinforcements (Dichter, 1966; Sundaram et al., 1998). These motivations come from the theory of motivations for participating in WOM but also apply to e-WOM as Peddibhotla and Subramani (2007) and (Hennig-Thurau et al., 2004) found.

According to Price, Feick, and Guskey (1995), market involvement also influences the likelihood of somebody helping others. This can be in the form of writing a review and is an altruistic motivation. Market involvement can cause a person to feel empathy for another 'shopper', and the person might be able to reduce the distress by helping, e.g. writing an experience. Being involved in a market might also increase a person's feeling of being an expert since involvement and expertise are related according to (e.g. Bloch, Sherrell, & Ridgway, 1986). A person is more likely to help another if she feels competent to help (Harris & Huang, 1973). This, in turn, likely means that a person is more likely to write a review if she is more involved in a market. A reviewer might write the review to reduce the distress of others.

Another altruistic reason might be because a consumer wants to *help the company* Sundaram et al. (1998) when the customer is pleased about the product. Customers want to give something back when the experience is strongly positive. She might therefore try to create a positive image about the product by, in the online world, writing a positive review about it.

A similar motivation to the motivation of wanting to help the company is the will to help employees (Whiting et al., 2019). This motivation can be divided into two main reasons, that is, wanting to help the employee receiving either compensation or recognition.

We expect that different motivations lead to different ratings. A person that writes because of conspicuous reviewing reasons probably also gives a higher rating to signal to the reader that they made a good decision with their purchase. A person that wants to be heard by the company because she has a complaint probably gives a lower rating on average because she has a complaint and perhaps also to catch the attention of the company. If a reviewer writes for balance seeking reasons, she probably rates according to the overall satisfaction that she wants to balance out. If a product is unexpectedly bad, she will write a negative review and thereby probably also give a bad rating, and vice versa for an unexpectedly high satisfaction. A reviewer that writes reviews because she cares for others probably rates the product as she perceived the quality of the product since it is most helpful for the reader. If a person writes because she wants to help the company or want to help the employees of a company, she is probably more likely to rate the product or service higher because that sends out a positive signal and is thereby more helpful to the company or employees in improving their image. If a person writes a review because she wants to help the company to make changes, she might give a lower rating to signal to the company that there is a problem. The

expectation that we have for certain motivations are summarized in table 1.

	Expected Effect on Rating
Conspicuous reviewing	↑
Want to be heard by the organization	↓
Balance seeking	↓ or ↑
Help the company (make changes)	↓ or ↑
Help employees	↑

Table 1: Expected Effect of Motivations on the Rating

To answer the question *why do people write reviews?*, we can conclude that people write reviews based on different motivations that we can divide into egocentric motivations and altruistic motivations. Table 2 gives an overview of the motivations. We expect that the different motivations lead to different ratings, and this leads to the need for extra caution with interpreting the final model since the rating might not always reflect the true satisfaction of a customer.

2.2 Who Writes Reviews?

Prior literature shows that a large part of the reviews are written by a small group that write many reviews, and that small contributions are prevalent (Peddibhotla & Subramani, 2007), Juran (1992) calls this pattern of contributors *the vital few and the useful many.* Marwell and Oliver (1993) proposed the theory of the critical mass. This first pattern is in line with this theory, suggesting that this group is the critical mass (Peddibhotla & Subramani, 2007). This group is not only very active in making contributions, but they also make the most valuable contributions. We can see reviews on a product as a collective good in the online setting since they benefit everybody

Table 2: Motivations for Writing Reviews

Egocentric	**Altruistic**
conspicuous reviewing / self-expression *(Dichter, 1966; Kovács & Horwitz, 2018; Lampel & Bhalla, 2007; Peddibhotla & Subramani, 2007)*	Care about others *(Dichter, 1966; Peddibhotla & Subramani, 2007; Sundaram et al., 1998)*
Post-purchase advice seeking *(Hennig-Thurau et al., 2004)*	Reduce distress of others due to market involvement *(Price et al., 1995)*
Want to be heard by the organization *(Harrison-Walker, 2001; Whiting et al., 2019)*	Help the company (make changes) *(Sundaram et al., 1998; Whiting et al., 2019)*
Enhance understanding of topic *(Peddibhotla & Subramani, 2007)*	Help employees *(Whiting et al., 2019)*
Become part of online community *(Hennig-Thurau et al., 2004; McWilliam, 2000)*	
Enjoyment *(Litvin et al., 2008; Wang & Fesenmaier, 2004; Wu, 2019)*	
Developing writing skills *(Peddibhotla & Subramani, 2007)*	
Balance seeking *(Sundaram et al., 1998)*	

in the group, the people reading them, it is hard to exclude individuals, and it is non-rivalrous, as one person using them makes it not impossible for another to make use of them. As stated by Marwell and Oliver (1993), there is usually a mismatch between individual interests and group interests for groups with a common interest in a collective good. One problem might be an efficacy problem when no one is able to create enough common benefit individually to make acting worthwhile. Another problem might be a free-rider problem when individuals in the group expect others to act for the common interest. The critical mass is needed in a group. The critical mass sets collective action in motion to overcome the aforementioned problems, e.g., efficacy and free-riding (Marwell & Oliver, 1993).

According to Peddibhotla and Subramani (2007) the median number of useful votes per review, that is how helpful other users find the review, for `Amazon.com` top 1000 most prolific reviewers is almost four times as large as the median for all other reviewers.

Interestingly, according to Peddibhotla and Subramani (2007) over 50% of the reviews written by the top 1000 most prolific reviewers are among the first ten reviews available for a product. They, overall, make early-period contributions with few prior contributions by others. This is also in line with the critical mass theory and provides empirical evidence for the 1000 most prolific reviewers to be considered the critical mass (Peddibhotla & Subramani, 2007). The critical mass theory states that the critical mass will contribute their resources to collective action earlier than others (Marwell & Oliver, 1993). According to Peddibhotla and Subramani (2007), there is strong empirical evidence that the 1000 most prolific reviewers show characteristics of the critical mass that is suggested by the critical mass theory because this group write more reviews, are early-contributors, and write more helpful reviews. The fact that this group write more helpful reviews might also be an effect of the fact that their reviews are often among

the first ten reviews.

Interestingly, the most often self-mentioned reason for the top 1000 prolific reviewers for writing reviews is for altruistic reasons, which is the case for nearly 40% of these reviewers, according to Peddibhotla and Subramani (2007). Peddibhotla and Subramani (2007) also found that for people of the top 1000 most prolific reviewers, mentioning altruistic reasons for writing reviews is significantly correlated with a higher quality review. They also found that they write fewer reviews, but this correlation was not significant. The second and third most mentioned reasons are, respectively, to be a part of a social community, as described by Hennig-Thurau et al. (2004); McWilliam (2000), and self-expression, as described by Dichter (1966); Kovács and Horwitz (2018); Lampel and Bhalla (2007); Peddibhotla and Subramani (2007). Peddibhotla and Subramani (2007) found that both aforementioned reasons had significant correlations with the number of reviews that are written. The self-expression motivation has a positive correlation, and the "wanting to be a part of a social community" - motivation has a negative correlation with the number of reviews that an individual wrote. Summarized, egocentric motivations for writing reviews are mostly positively related to quantity, whilst altruistic motivations are mostly positively related to quality (Peddibhotla & Subramani, 2007). According to Whiting et al. (2019), the most important reason for participating in e-WOM, in general, seems to be for altruistic reasons. This result is the same as for the `Amazon` top 1000 prolific reviewers.

To summarize, the people who write reviews are, to a certain extent the, more experienced and more vocal, critical mass. The critical mass starts the collective action of writing reviews according to Marwell and Oliver (1993). Besides that, many reviews are written by people who do not post reviews as often as the people in the critical mass. As stated by Juran (1992) *the vital few and the useful many.* The most important

motivation for both groups of reviewers seems to be the altruistic motivation (Peddibhotla & Subramani, 2007; Whiting et al., 2019). For the top 1000 prolific reviewers, Peddibhotla and Subramani (2007) found that egocentric motivations are mostly positively related to the number of reviews, and altruistic motivation are mostly positively related to the quality of the reviews.

2.3 What Do People Write About?

The reviews consist of product attributes and the reviewer's opinion about said attribute. Besides that, a reviewer might also write about her general feeling towards the product or brand without specifying a specific attribute.

It is good to make a distinction between different attribute types. A well-known model for segmenting product attributes for the perceived quality of a product is the Kano model by Kano (1984). According to Kano (1984), consumers value the attributes according to four requirement levels for the perceived quality.

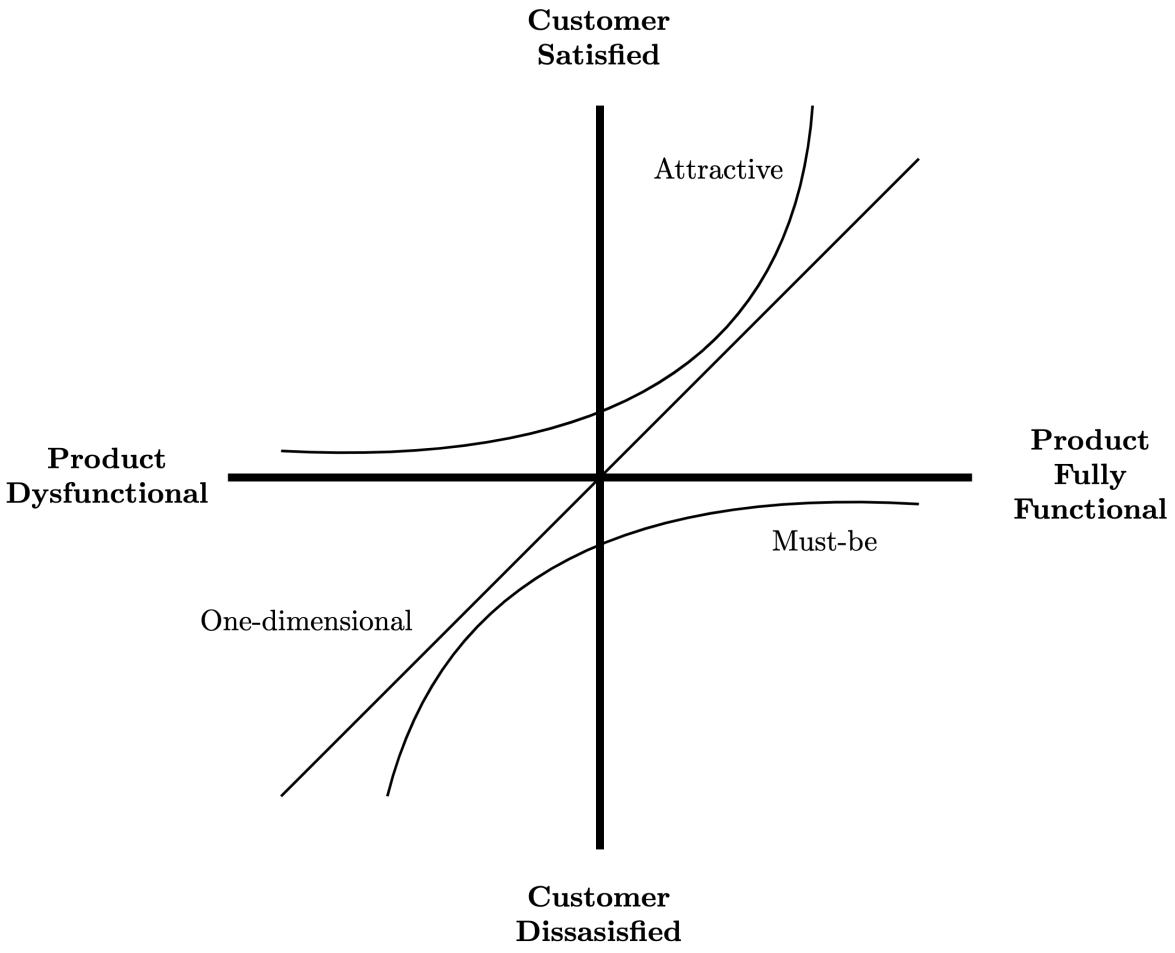

Figure 1: Kano Diagram (C. Berger et al., 1993)

The first type is the *one-dimensional* attributes. These attributes follow the 45° line in figure 1 (C. Berger et al., 1993). These are comparable to the traditional interpretation of quality whereby it is thought that the customers' satisfaction is simply a linear function of the product's functionality. We can assume that these represent performance-related attributes, they are known and specified to the customer, and they are measurable and often technical (Aune, 2000). We might expect the memory to be a one-dimensional attribute in a cellphone since it is measurable, known beforehand and technical.

The second type of attributes is *Must-be* attributes. These attributes are the attributes that the customer assumes to be present. These attributes can be seen as the basic requirements of a product, they are often implied, self-evident, not mentioned in the advertisements, and rather importantly, they are taken for granted (Aune, 2000). With a mobile phone, we can expect, e.g., the possibility of calling with it to be a must-be attribute as it is implied, not mentioned and often taken for granted. Bad performance of these attributes will lead to more dissatisfied customers, but better performance will never lead to a more satisfied customer than neutral.

The third kind of attributes is *attractive* attributes. Attractive attributes increase the customers' satisfaction when the attribute is more functional, but the customers' satisfaction will not be dissatisfied when the attribute is not functional. We can see these attributes as surprises to the customers; they are often not expressed, customer-tailored, and often rather exciting (Aune, 2000). With mobile phones, we can expect this could, e.g., by the operating system's quality, to somebody unfamiliar with it.

The fourth and last type of attributes is *indifference* attributes. These attributes do not affect the satisfaction of the customer. This line will be plotted on the horizontal axis in figure 1.

There might also be an interaction between what is mentioned on the product page, what people expect because of that and what they talk about. One-dimensional products that are mentioned on the product page might not be essential to write about, Xiao, Wei, and Dong (2016) found that for their research, attributes that were well described on the product page did not seem to influence the rating of the product much because people already knew how these attributes would be.

Knowing what different types of attributes exists helps us in understanding what a person might talk about. When a person is happy about the product and writes a review, she most likely not write about *must-be* attributes on average since these attributes did not increase her satisfaction, and since she is satisfied, the attribute is good enough. Similarly, when a review is negative, we expect it to not mention *attractive* attributes on average. *One-dimensional* attributes are probably discussed in both ratings with a high as well as a low rating. We expect that reviewers do not mention *indifference* attributes in their review because these attributes do not influence their satisfaction.

We can also expect some differences in what people write about depending on their motivation for writing a review. If a person writes a review because she wants to help other people, she likely writes about a more expansive arrangement of attributes belonging to all attribute types than when a reviewer writes because she wants to show off what she bought as with the conspicuous reviewing motivation. As Peddibhotla and Subramani (2007) found in their paper that, for the top 1000 prolific reviewers, egocentric motivations are positively related to quantity, and altruistic motivations are positively related to quality and a better helpfulness score. People who write reviews with the motivation of helping others will also probably not write a lot about attributes that are already known because it is not very helpful to repeat such known features, such as the physical memory of a phone. While, people who write, e.g., for conspicuous reviewing reasons, might, for example, write about the physical memory of a phone, that is known beforehand, if that signals a message to others, as a phone with a large physical memory might be very expensive. A person who writes for altruistic reasons in wanting to help others will probably also write more about attributes that one needs to experience before knowing the true quality than about attributes that are known

beforehand because that will help others more by reducing the risk that others take. People who write for the reason of wanting to be heard by the company or seeking advice about, e.g., a problem, might also not write about all the attributes that are important to them or other customers but only about things that they dislike or want to resolve because for some reason they want the company to hear their complaints.

	One-dimensional	*Must-be*	*Attractive*	*Indifference*
Present	↑	–	↑	–
Missing	↓	↓	–	–

Table 3: Effect of Different Attribute Types on Satisfaction (Aune, 2000)

We expect that reviews are about different attributes of a product. We expect that the attributes that a review is about differ according to the satisfaction of a reviewer and according to the motivation that a person has for writing a review. Knowing that these differences exist helps interpret the final results. For example, when a review is positive, it probably does not mention a *must-be* attribute, but that does not mean that said attribute is not needed for the customer.

3 Methodology

The method that we use in this paper is separated into four stages. Each stage answers a sub-question. The rest of this chapter first describes how to determine the relevant product attributes. Next, we discuss how we determine the sentiment of a review about an attribute. After that, we discuss how we determine the aggregate consumer preferences concerning the different attributes. Lastly, we describe how we visualize these preferences.

3.1 How to Determine What the Relevant Product Attributes Are?

It is crucial to make the distinction between explicit and implicit attributes. Explicit attributes are found more easily, and an example could be: *"I like the camera on this phone"*, whereby the attribute would be *camera*. Implicit aspects are a bit harder to find. An example of an implicit attribute would be *battery life* in the sentence: *"I can use this phone all day without charging."*.

Different ways of finding and mining product attributes are proposed in the current literature. To give an overview of the most prominent methods, we will divide them into two groups based on the methodology. The first group are the methods that use frequency-based or statistical methods to find the attributes, and the second group are methods that use rule-based and other methods.

One of the most well-known methods to mine product attributes is proposed by Hu and Liu (2004). To find the explicit product attributes in unstructured reviews, they use the association mining algorithm that was proposed by Agarwal and Srikant (1994). This is an example of a frequency-based method. Hu and Liu (2004) detect opinion

words by looking at adjacent words to the explicit aspects. The implicit aspects are found by using the opinion words that they detected. Bafna and Toshniwal (2013) built on this method and introduced a probabilistic approach to remove feature candidates that are not features. The first group domain synonym nouns or noun phrases using the method proposed by Zhai, Liu, Xu, and Jia (2011), they then perform part-of-speech (pos) tagging using the Stanford Parser proposed by Toutanova, Klein, Manning, and Singer (2003), after that, they use the proposed Associative Rule Mining Technique to find domain-specific features. Popescu and Etzioni (2007) propose a method called OPINE that uses a statistical method with point-wise mutual information (PMI) and information from the web using web-PMI to find aspects from a list of potential aspects that is created by taking all noun phrases with a certain frequency threshold. Another method, proposed by Bagheri et al. (2013) makes use of a bootstrapping algorithm to find the explicit aspects. After this process, it uses the generative opinion lexicon created before the bootstrapping to find the implicit aspects. Using opinion words to find implicit aspects is similar to what Hu and Liu (2004) do for finding implicit aspects.

Bancken, Alfarone, and Davis (2014) use a rule-based algorithm to find the attributes. Their method is called SPECTATOR. A set of rules, based on linguistics, is used to find attributes after finding the dependency tree using the Stanford Dependency Parser (De Marneffe, MacCartney, & Manning, 2006; Klein & Manning, 2003). Another paper that uses a rule-based algorithm is by Poria, Cambria, Ku, Gui, and Gelbukh (2014), they propose to use implicit aspect clues (IAC) to focus on the implicit aspects in the reviews. The aforementioned methods work well when aspects are formed by a single noun, but they result in many attributes when attributes are formed by many low-frequency words or when the attributes are formed by abstract terms (Brody & Elhadad, 2010).

In this paper, we make use of another technique wherefore we make an assumption, which is; **we assume that every sentence in a review is about an attribute or topic and about one topic only**. In the review domain, this often holds true; whilst reviews are about multiple attributes, any particular sentence concern only one topic (Büschken & Allenby, 2016). With this assumption in place, we can use topic modelling on the sentence level, similar to the method by Brody and Elhadad (2010). We assign a topic to each sentence using most of the information in the sentence. Another advantage of using this topic modelling is that we can use more information from a sentence. The verbs that are used, for example, might indicate that a specific sentence is about a specific topic. To give an example in the context of phone reviews, if a person uses the verb **touching** in a sentence, she is more likely to talk about the screen than about, for example, the price. Topic modelling on a sentence level gives us another advantage over the other methods, and that is the fact that we no longer have to differentiate between explicit and implicit aspects since we can find both while assigning a topic to a sentence.

A well-known method for topic modelling is Latent Dirichlet Allocation (LDA) and was introduced by Blei et al. (2003). We use LDA on a sentence level, meaning that we see each sentence as a document in the corpus. We assign a topic to each sentence that has the highest probability according to LDA. We do this because people seem to write about different topics across sentences but not within sentences (Büschken & Allenby, 2016). The main goal is thus finding latent topics in our data that represent the product's attributes.

Brody and Elhadad (2010) also use an LDA model on a sentence level which they call a local LDA. They state that using LDA on a sentence level solves the problem of getting global topics instead of rateable topics that one might get using LDA on the

entire review. Büschken and Allenby (2016) also apply a form of LDA on a sentence level which they call sentence constrained LDA (SC-LDA), whereby they assume that all words in a sentence belong to the same topic, which slightly increases the ease of interpreting the topics. In this paper, we use a different method to interpret the topics based on relative word importance, which we discuss later. Other extensions of LDA for topic mining have been introduced in literature, but these methods require extra information and are not suitable for our method (e.g., Andrzejewski, Zhu, & Craven, 2009; Titov & McDonald, 2008).

For this section we use the following notation. A sentence is a sequence of N words and represented by an N-dimensional vector. Since the words are one-hot encoded with vectors of length V, which represents the total vocabulary size we can denote the sentence as $\mathbf{s} = \{(s_1, s_2, ..., s_N) | s_i \in \mathbb{N} \cap [1, V] \text{ for each } i\}$. The set of sentences is called a corpus and it consists of M sentences in total. We can denote the corpus as $D = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_M\}$. In total there are K latent topics, which we need to specify. $\mathbf{z}$ represents a vector of the topics for the words in a sentence,$\{(z_1, z_1, ..., z_N) | z_i \in \mathbb{N} \cap [1, K] \text{ for each } i\}$. β represents a $K \times V$ matrix, where the assumption is made that each row is independently drawn for an exchangeable Dirichlet distribution with η as scalar parameter. α represents the Dirichlet prior on the per-document topic distributions. θ represent the topic distribution and is drawn from $\theta_s \sim Dirichlet(\alpha)$, such that $\theta_{s=1...M}$ is a K-dimensional vector which values sum up to 1. To illustrate the model we present a visual representation in figure 2.

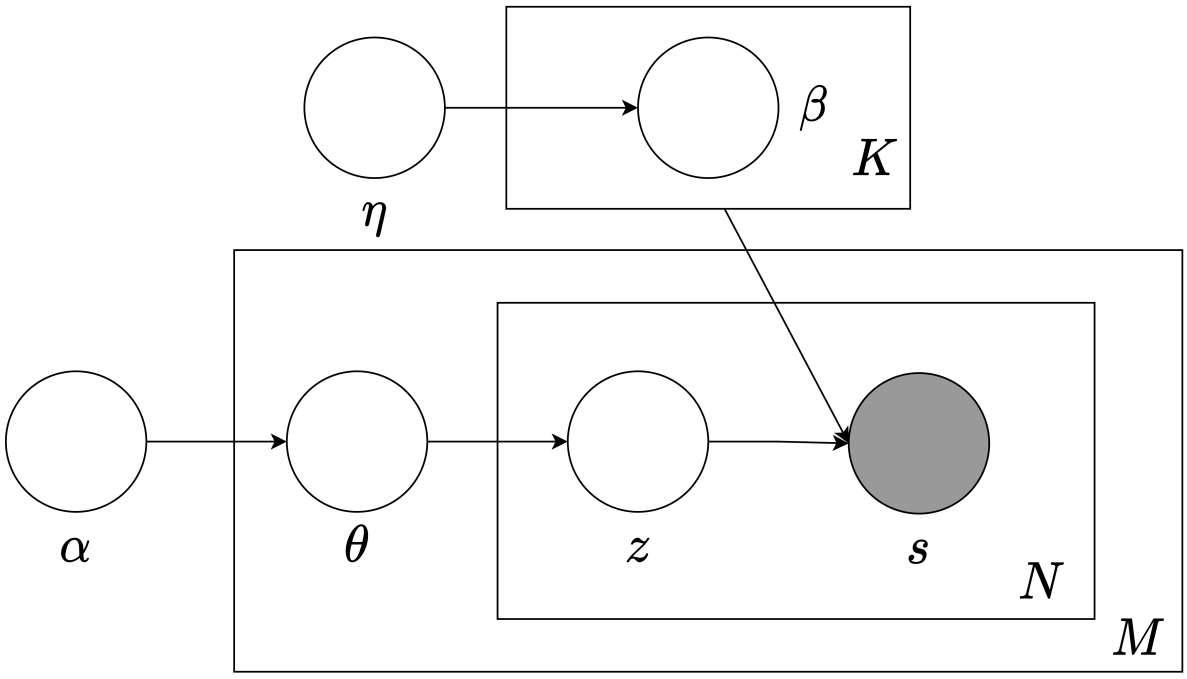

Figure 2: LDA Diagram (Blei et al., 2003)

To use LDA we need to compute the posterior distribution of the hidden variables for a sentence. This inferential problem poses as follows and returns an intractable distribution:

$$p(\theta, \beta, \mathbf{z}|\mathbf{s}, \alpha, \eta) = \frac{p(\theta, \beta, \mathbf{z}, \mathbf{s}, |\alpha, \eta)}{p(\mathbf{s}, \beta|\alpha\eta}$$

To find the hidden variables we have to use another technique to approximate the posterior distribution. Because the data used for problems that this paper tries to solve are often rather large we need to use an efficient method to do this. We use a method called Online Variational Inference for LDA that was suggested by Hoffman, Bach, and Blei (2010). This method is incorporated in the Python `gensim` package

(Řehůřek & Sojka, 2011). It is similar to standard variational inference in that we try to approximate the posterior distribution with a simpler distribution, $q(\mathbf{z}, \theta, \beta)$ which is indexed by free parameters that we set to maximize the Evidence Lower BOund (ELBO) (Hoffman et al., 2010): $\log p(\mathbf{s}|\alpha, \eta) \geq L(\mathbf{s}, \phi, \gamma, \lambda)$. By maximizing the ELBO, we minimize the KL divergence between the true posterior distribution and the simpler distribution, $q(\mathbf{z}, \theta, \beta)$. Hoffman et al. (2010) propose a fully vectorized distribution of q: $q(z_{si} = k) = \phi_{sw_{si}k}$; $q(\theta_s) = Dirichlet(\theta_s; \gamma_s)$; $q(\beta_k) = Dirichlet(\beta_k; \lambda_k)$. Whereby $z_{si} =$ represents the topic of the ith word in sentence s and w_{si} represents the ith word w in sentence s, and k represents the topic k which is an integer between 1 and K.

To perform the Online Variational Inference, first, the locally optimal values of γ and ϕ are found by iteratively updating them until convergence while keeping λ fixed. After that, $\tilde{\lambda}$ is computed, given the values that we found in the previous step, which is λ that would be optimal if the entire corpus was made up of M times a single document. $\tilde{\lambda}$ is used to update λ, in combination with a weighted average of the previous value of λ. To reduce noise in the process, multiple sentences are considered per update. This step keeps repeating.

To find the number of topics that we should use, we can use the coherence value. We use the coherence value as proposed by Röder, Both, and Hinneburg (2015). The optimal number of topics is found by comparing and selecting the LDA model with the highest coherence value.

The K topics that are uncovered need to be labelled manually. We look at the words within each topic with the highest probability of that word belonging to said topic and decide on the label. Some topics can have mostly words with no specific meaning with respect to product attributes. To discover what attributes these topics describe, we can

look at, what we call, the relative probabilities, whereby we divide the probability of the word by the number of times that word appears in the corpus. This way, we can see the discriminative words in the topics that are not general to the corpus.

3.2 How to Determine the Sentiment of a Review About an Attribute?

Now that we know what the different attributes (topics) are that people talk about, it is crucial to know the sentiment of a review towards that attribute. If we know the sentiment, we can use that to find the aggregate consumer preferences in the next step. Since we have the assumption that each sentence is about a single attribute, we can perform sentiment analysis on that sentence to find the sentiment of a reviewer towards that attribute.

To perform the sentiment analysis, we make use of the sentiment analysis tool that is part of the Python `TextBlob` (Loria, 2018) package. This method has a reported accuracy of 75% on the polarity dataset by Pang and Lee (2004) and works quickly, which makes it attractive for our case where we often have many sentences to process. The `TextBlob` sentiment function uses a lexicon with a polarity score for each word and an intensity score for each word. Polarity is a value between -1 and 1 and represents the sentiment. When a word is considered a modifier, such as *very*, the intensity is used to modify the polarity of the text by multiplying it. Negations are also used in the sentiment analysis step. Whenever a negation appears in a text, the polarity is multiplied by -0.5. If there is a negation in a sentence and a modifier, the polarity is multiplied by the inverse of the intensity of the modifier instead of by the intensity itself. The final polarity score is a value between -1 and 1. Whereby -1 is very negative, and 1 is very positive.

The polarity of a reviewer towards a topic is thus calculated as above. If a reviewer does not write about a specific topic, the polarity towards said topic is assumed to be 0. If a person writes multiple sentences about the same topic, we take the average of the polarities as the polarity towards that attribute because we assume that when a person writes two positive sentences about an attribute, she does not necessarily like that attribute twice as much as when she writes one sentence about it. Taking the average of the polarities also makes sense when a reviewer writes both a positive and a negative sentence about an attribute. We thereby assume that overall the person is neutral about that attribute.

Now that we have clear how we can determine what the relevant product attributes are and the sentiment towards these attributes, we describe how we can find aggregate consumer preferences and what attributes are most important to a customer.

3.3 How to Determine What the Aggregate Consumer Preferences Concerning the Different Attributes Are?

According to Engler, Winter, and Schulz (2015), the online product rating represents the customer's pre-purchase expectations and actual product performance. We can therefore assume that the rating represents the customer's satisfaction and not the actual product quality. Since it represents satisfaction, we can assume that the rating is reflective of the utility that a customer got from consuming the product. With this information in mind, we try to model the rating.

To determine the most important attributes and see the aggregate consumer preferences, we use a multinomial logistic regression. Another regression that we considered is an ordinal logistic regression, this regression has a restrictive assumption of propor-

tional odds, which does not hold quite often but is more parsimonious. To make our method more widely applicable, we propose only the multinomial logistic regression. We model the rating using the sentiment towards the topic. Besides that, we also include the brand as a control variable. We can use a multinomial logistic regression because the dependent variable, rating, is a multinomial variable. For a multinomial logistic regression, we must have a large enough sample size, Hosmer Jr, Lemeshow, and Sturdivant (2013) propose to use at least 400 observations. It is often easy to gather more than enough observations to use this method in the case of reviews.

We take a rating of 1 as the reference. Therefore we formally model;

$$p(Rating_i = 1) = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\beta_j \cdot \mathbf{X}_s}}$$

$$\forall\, r \in \{2, ..., 5\}$$

$$p(Rating_i = r) = \frac{e^{\beta_r \cdot \mathbf{X}_i}}{1 + \sum_{j=2}^{5} e^{\beta_j \cdot \mathbf{X}_i}}$$

Where each β_j represents a vector of the coefficients for each rating j, we incorporate the polarity towards each topic and the brand into the model. The vectors β_j thus have a size of K + number of brands in the data. X_i represent the model vector for each review.

The final model is fit, whereby the coefficients are set to minimize the negative log-likelihood of the model.

To see what coefficients are significant we make use of the Wald χ^2 statistic which

we can calcuate as follows (Bewick, Cheek, & Ball, 2005):

$$\frac{\text{coefficient}}{\text{Standard Error}_{\text{coefficient}}}$$

To finally compare which coefficients are most important for the rating, we have to look at the standardized effects. We standardize the coefficients by multiplying them with the standard deviation for said variable.

Multinomial logistic regressions are often used to categorize by selecting the outcome with the highest predicted probability. In this paper, we multiply the probability for each rating with the rating to get the predicted rating for each review. With this predicted rating, it is possible to calculate the R^2 of our model as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{M}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{M}(y_i - \bar{y})^2}$$

,

where y_i represents the actual rating for review i, $\hat{y}_i$ represents the predicted rating and $\bar{y}$ represents the average rating.

Lastly, we describe how we can visualize the findings of the aforementioned method.

3.4 How to Visualize the Aggregated Consumer Preferences?

To visualize the aggregated consumer preferences, we consider methods to visualize the outcomes of our model. The simplest and most intuitive way to visualize what attributes matter the most for consumers in their rating behaviour and thereby their

satisfaction is by creating a plot of the standardized coefficients of our model. Each topic will be represented by 4 bars representing the effect of a one standard deviation increase of the variable on the log odds of being that rating versus the reference rating for which we took a rating of 1.

Another visualization that we use is based on the method by Fox and Andersen (2006). Fox and Andersen (2006) call this visualization an effects display. The effects display is easy to interpret and immediately shows us the effects of the individual attributes on the probabilities of the ratings. For this plot, we look at the predicted probabilities of ratings for different combinations of predictors. These results are then plotted for each topic to give an overview of their effect on the rating, whereby the X-axis represents the polarity towards said attribute.

4 Data

To test the methods in this paper we use reviews that are scraped from `amazon.com`, `amazon.co.uk` and `amazon.ca`. The total cleaned dataset consists of 29,891 phone reviews. Cleaning the data is rather important for methods like these to work properly. Firstly all non-English reviews were deleted from the dataset. Secondly, we keep only reviews that were written by reviewers whose purchase was verified by `Amazon` to be more sure that the reviews were not review spam. Besides this step, we also deleted duplicates from the text to minimize the effects of review spam. The focus of this paper is not on review spam, and we will therefore not focus on other state-of-the-art review spam detection machines. The reviews that we use have at least 40 words and consist of approximately 120 words on average. To only include brands in our dataset that have been reviewed more often, we kept only brands with at least 500 reviews.

The next step in cleaning was replacing emoticons with their meaning in words to include the meaning of the emoticons in our analysis. After that we made sure that brands were written correctly, so that e.g., *Samsung Electronics* and *Samsung* were both *Samsung*. The dataset consists of 15 brands. Table 4 summarizes the brands and the number of reviews per brand after cleaning.

Table 4: Number of Reviews per Brand

Brand	**Number of Reviews**
Samsung	5,797
Motorola	4,257
Nokia	3,544
Blackberry	2,701
Blu	2,356
Huawei	1,649
Sony	1,639
Xiaomi	1,452
Blackview	1,222
HTC	1,210
LG	996
Google	858
Oneplus	657
Microsoft	591
Oukitel	541

In the cleaning process, we also expand contractions. All contractions are expanded to their full notation, e.g., "aren't" $\rightarrow$ "are not", to ensure that all negations and important verbs are found.

After this step, we had to split the reviews into sentences, and we had to split these sentences into words, which we can also call tokens. The sentence splitting is done by using the `Punkt` algorithm by Kiss and Strunk (2006). We use a pre-trained version of their model that is trained on data from the Wall Street Journal that is

part of the Penn Treebank dataset (M. Marcus, Santorini, & Marcinkiewicz, 1993; M. P. Marcus, Santorini, Marcinkiewicz, & Taylor, 1999). To perform the tokenization step, which is the process of splitting the sentence into tokens, we use the pre-trained model *en_core_web_trf* of *Explosion.ai*'s `Spacy` tokenization pipeline (Honnibal, Montani, Van Landeghem, & Boyd, 2020). The pre-trained model is by Honnibal et al. (2020) and has a reported accuracy of 100% for the tokenization process; it is trained on more than 1,000,000 documents Weischedel et al. (2013). This model also performs part-of-speech tagging and dependency-labelling, for which it has a reported accuracy of 98% and 94% respectively, which we need in the next steps of our cleaning process.

Using the algorithm by Norvig (2007), all words are also checked on spelling and corrected when needed. For the same reason as synonym replacement, it is important that words are written correctly. The algorithm takes a Bayesian approach. The word that is considered the best spelling is chosen using: $argmax_{c \in candidates} P(c) \cdot P(w|c/P(w)$, where w is the word in case, c is the correction, $P(c)$ is the probability that c appears in an English text. In this paper, $P(c)$is calculated as the number of occurrences of c divided by the total length in words of a text. The candidates are selected by editing the original word and allowing either one or two edits whereby an edit can be either a deletion of a letter, a transposition of two letters, a replacement of a letter or an insertion of a letter. Since $P(w)$ is the same for each word, we can ignore it. We also assume that if w is in our list of words, which is needed to calculate $P(c)$, it is infinitely more probable than any $c \in candidates$, and we will only consider w a candidate here. We make a similar assumption, of being infinitely more probable, about candidates with one edit as compared to candidates with two edits. Lastly, we assume that edits with two edits are infinitely more probable than w if w is unknown. With these assumptions we can dismiss $P(w|c)$, since all candidates that are selected are equally probable,

which reduces the function to $argmax_{c \in candidates} P(c)$. This paper uses a document with approximately 32,000 unique words to calculate $P(c)$. The document consists of words from the Gutenberg Project (*Project Gutenberg*, 1972) expanded with the most frequent words from Wiktionary (Wikimedia Foundation, 2002) and the British National Corpus (Oxford Text Archive, 2009) and with domain-specific words such as e.g., "GPS" and brand names.

Next, we made sure that words that consist of two separate words are combined. To do this, all two consecutive nouns in the text are combined with an underscore. This means, for example, that *"speaker unit"* becomes *"speaker_unit"*. This is also needed for the next step wherein we use WordNet (Miller, 1998), which uses a similar notation.

The next step in the cleaning process consists of finding and replacing synonyms. Since synonyms have the same meaning and our method is based on word count, it might increase the performance if all terms with the same meaning are written similarly. After the sentences are tokenized, all nouns that are the subject, according to the dependency path that is found with Python `Spacy's` dependency parser (Honnibal et al., 2020), are collected and considered product features. For each of these nouns, their relevant synonyms are found using WordNet (Miller, 1998), a large electronic Lexical Database for the English language. All synonyms in the reviews are replaced with the same word, which reduces the number of unique words.

To further reduce the number of unique words, we perform a process called *stemming*, which is a method of word normalization. In this paper, we use the Porter Stemmer by Van Rijsbergen, Robertson, and Porter (1980). It is a straightforward method and uses a set of rules to replaces word endings with a typical ending to ensure that the words become similar. Examples are that plurals become singulars and that

certain verbs are shortened to a shape that specific shapes become similar words, e.g., *failing* and *failed* become *fail*.

All words that appeared less than 20 times in the entire cleaned corpus were also deleted because these words were often misspelt or not actual words. Besides that, all the stop words were also deleted because they carry no value in our case. The negations were kept in the review because they are essential for the sentiment of the sentence. The sentence that remained and had less than three words are deleted because they often carry no relevant information.

The final shape of the data is a matrix with a row for each sentence of the reviews. Each sentence is made up of the cleaned words.

The reviews all have a rating. The distribution of these ratings is shown in table 5. It is visible that a rating of 5 is the most occurring rating.

Table 5: Distribution of Ratings

Rating	**Number of Reviews**
1	5,288
2	2,349
3	2,797
4	4,841
5	14,195

5 Results

5.1 Latent Dirichlet Allocation

The first step in our methodology is to perform the LDA. After that, we will perform sentiment analysis on the individual sentences, and after that, we create our multinomial logistic regression.

To find the optimal number of topics, we selected the model with the highest coherence value. In our case, the optimal number of topics is 10. The coherence value of our final LDA model is 0.5632.

The top 25 words for the individual topics are shown in table 7. Because this is still a bit messy it is interesting to look at the words that are relatively the most important, with respect to their occurrence. The top 10 words with the highest relative importance are shown in table 8. In these tables we can see what the topics are about and we manually labeled them as follows and we henceforth refer to them with their label:

Table 6: Topic Labels

Topic	**Label**
Topic 1	Service
Topic 2	Design
Topic 3	Sim Card
Topic 4	Hardware
Topic 5	Purchase
Topic 6	Camera
Topic 7	OS[1]
Topic 8	Price
Topic 9	Battery
Topic 10	Memory

[1] Operating System

service refers to the overall service that the brand or seller gives. This also includes warranty problems. *design* refers to the overall design of the phone. The *sim card* topic refers to the sim, provider, and reception topic of the phones. The *hardware* topic is about the phone hardware and hardware included in the box. *Purchase* is about the way somebody purchased the product and their purchase experience. *Camera* is about the camera and the pictures it takes. *OS* is about the operating system and the software on the phone. *price* speaks for itself and is about the price of the product. *battery* is about the battery and the battery life. *Memory* is about the available physical memory on the phone.

Table 7: LDA Top 25 Most Important Words

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
not	use	onli	work	new	screen	veri	go	batteri	need
one	good	samsung	no	buy	camera	get	also	charg	app
would	great	old	time	bought	more	ani	price	day	take
back	like	best	well	issu	much	android	want	even	come
make	realli	sim	seem	year	better	problem	purchas	last	set
could	look	sim_card	tri	first	fast	thing	happi	few	say
call	still	mobil	everyth	month	easi	recommend	expect	most	far
receiv	love	unlock	turn	amazon	feel	updat	featur	though	lot
start	nice	g	howev	case	bit	think	nokia	way	product
seller	devic	differ	came	replac	littl	review	mani	long	googl
know	find	model	without	two	hand	run	money	give	abl
put	overal	found	box	never	excel	enough	quit	hour	gb
sure	perform	canada	fine	week	big	disappoint	reason	alway	made
peopl	connect	chang	open	see	amaz	keep	£	full	quick
help	almost	plu	everi	return	pretti	read	less	play	instal
text	absolut	previou	stop	sinc	qualiti	wife	upgrad	around	right
ever	especi	smart	htc	another	clear	slow	definit	easili	sdcard
second	simpl	version	origin	order	actual	blackberri	top	bad	etc
anyth	beauti	galaxi	touch	drop	photo	quickli	valu	thought	noth
contact	design	took	perfectli	someth	decent	respons	includ	hold	yet
send	fact	alreadi	button	arriv	compar	notic	pay	end	extra
fix	solid	smart_phon	longer	brand	pictur	mean	function	ok	download
check	fingerprintread	said	let	may	impress	support	although	game	storag
tell	awesom	switch	show	pleas	size	softwar	cheap	star	side
either	internet	dual	black	soni	light	email	basic	miss	move

Table 8: LDA Top 10 Relative Most Important Words

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
contactlist	public	canada	suspici	unvex	lightphoto	freedom	goplu	halfday	lesson
legal	securityfeatur	canadian	chargingcord	waterdamag	doubtatom	behav	linephon	temporarili	memoryspac
repaircentr	speakervolum	roger	pour	fond	tripl	touchwiz	pixela	unexpectedli	bake
pretend	pink	north	unlockcod	outright	distanc	downfal	op	wifesign	ex
deaf	cameraqu	senior	wallplug	granddaught	landscapemod	gwife	umidgi	wind	adressbook
author	tooth	virgin	chargec	promptli	qualityimag	dec	cheer	monster	unwant
code	gold	spanish	evid	vendor	distort	nontheless	pricediffer	appus	mmaudio
garag	videoplayback	samsungpay	fuss	chair	glassback	nought	rangephon	screentim	cloudstorag
disclos	fussi	latin	drove	glassprotector	nightshot	internetaccess	phoneon	chargingtim	corrupt
expert	badg	exyno	appropri	paint	mono	anybodi	grate	turncharg	babybutton

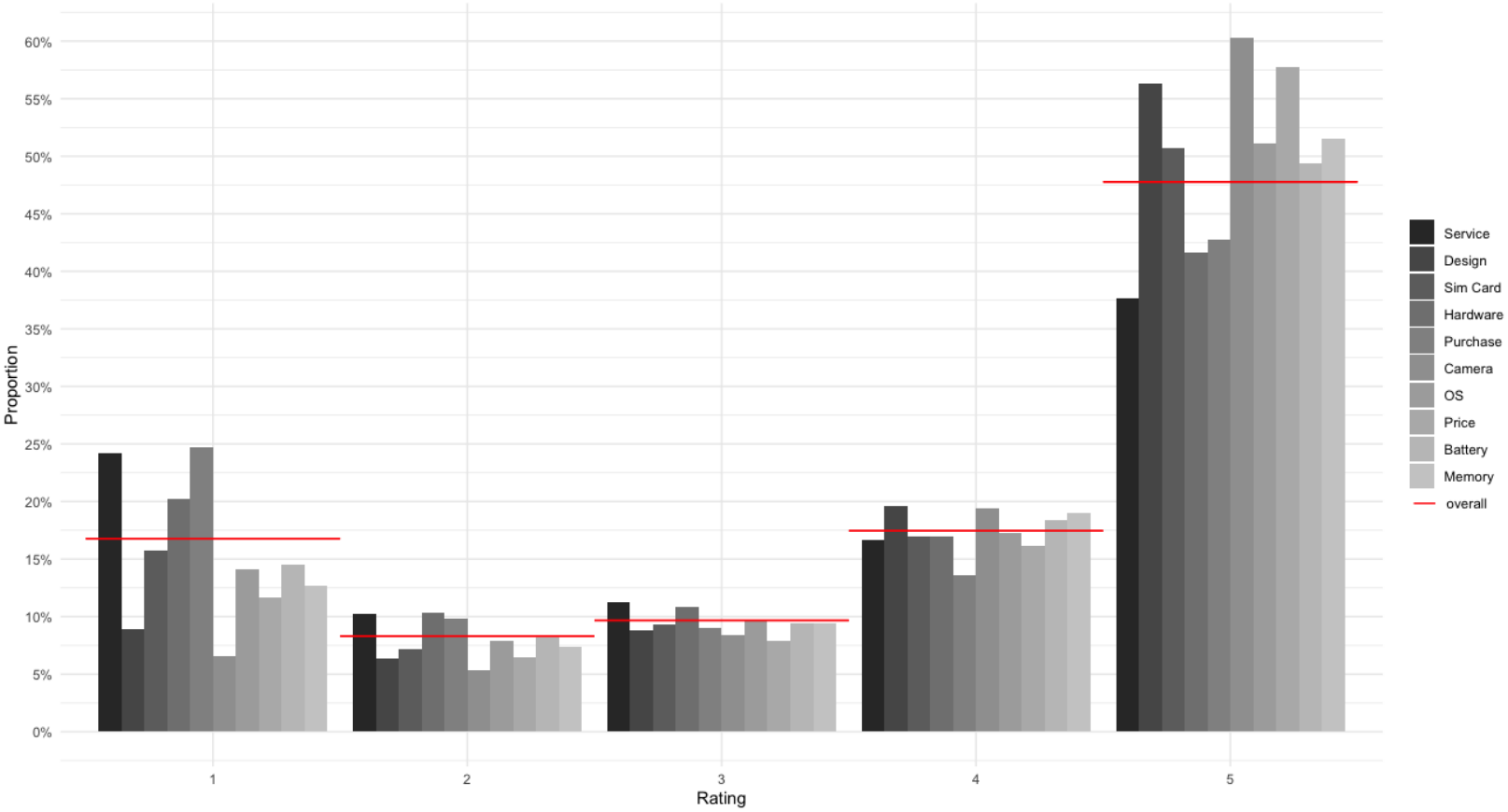

Figure 3: Proportional Distribution of Occurrences of Topics

The proportional distribution of each topic per rating is shown in figure 3. We can, for example, see that approximately 24% of the reviews that mention *service* have a rating of 1 whilst overall approximately 17% of the reviews have a rating of 1.

5.2 Sentiment Analysis

The next step is the sentiment analysis towards the topics defined in the LDA step. In table 9 we give a summary of the polarity towards the different attributes.

Table 9: Sentiment Analysis Summary

Topic	**Reviews**[1]	**Mean**	**Min**	**Max**	**SD**[2]
Service	9,309	0.028	-0.538	1.000	0.099
Design	10,663	0.107	-0.428	0.788	0.110
Sim card	7,347	0.054	-0.500	1.000	0.101
Hardware	8,133	0.043	-0.700	1.000	0.096
Purchase	9,125	0.036	-0.700	0.780	0.092
Camera	11,060	0.082	-0.341	1.000	0.095
OS	8,413	0.042	-0.658	0.750	0.103
Price	8,778	0.071	-0.500	0.803	0.102
Battery	8,418	0.037	-0.700	0.800	0.084
Memory	6,925	0.045	-0.975	1.000	0.088

[1] Number of reviews mentioning the topic

[2] Standard Deviation

5.3 Estimating Aggregate Customer Preferences

As mentioned, we use a multinomial logistic regression for the aggregation step. The final model has a Mean Squared Error of 1.592. The model has an R^2 of 0.340. To check for the significance of each of the independent variables, being all the sentiments towards the topics, and the Brand dummy variables, we use the Wald test, which has

(Wasserman, 2013):

$$H_0 : coefficient = 0$$

$$H_a : coefficient \neq 0$$

The Wald static is calculated as: $\frac{\hat{\beta}}{\hat{se}(\hat{\beta}} \sim N(0,1)$, where $\hat{\beta}$ represents the predicted coefficient. In Appendix A, one can see the p-values for all the coefficients. On a 5% significance level, we can reject the H_0 for all coefficients of the sentiments towards the topics and most of the coefficients for the brand control variables.

Table 10 shows the coefficients of the multinomial logistic regression.

Table 10: Coefficients Multinomial Logistic Regression

	Rating 2	**Rating 3**	**Rating 4**	**Rating 5**
Service	1.925	2.290	4.127	5.870
Design	2.548	6.217	10.583	13.039
Sim Card	1.869	3.595	6.636	10.919
Hardware	2.397	2.568	5.875	8.136
Purchase	1.411	1.398	2.670	6.603
Camera	6.034	10.211	17811	21.846
OS	1.416	3.796	7.936	12.006
Price	2.322	4.601	9.311	13.588
Battery	2.064	4.703	8.819	11.853
Memory	1.972	2.782	8.243	11.824
Blackberry	*reference brand*	-	-	-
Blackview	-1.148	-1.106	-1.006	-0.270
Blu	-0.849	-0.888	-1.005	-1.171
Google	-1.170	-1.118	-1.447	-0.882
HTC	-1.342	-1.370	-1.561	-1.218
Huawei	-1.115	-1.156	-1.240	-0.618
LG	-1.142	-1.187	-1.250	-0.913
Microsoft	-0.619	-0.692	-0.677	-0.458
Motorola	-0.729	-0.689	-0.800	-0.469
Nokia	-0.721	-0.707	-0.746	-0.705
Oneplus	-0.393	-0.749	-0.602	0.174
Oukitel	-0.639	-0.729	-0.911	0.059
Samsung	-1.089	-0.991	-0.1006	-0.560
Sony	-0.0821	-1.144	-1.165	-1.071
Xiaomi	-0.640	-0.561	-0.576	0.106
R^2	34.029%			
MSE	1.592			

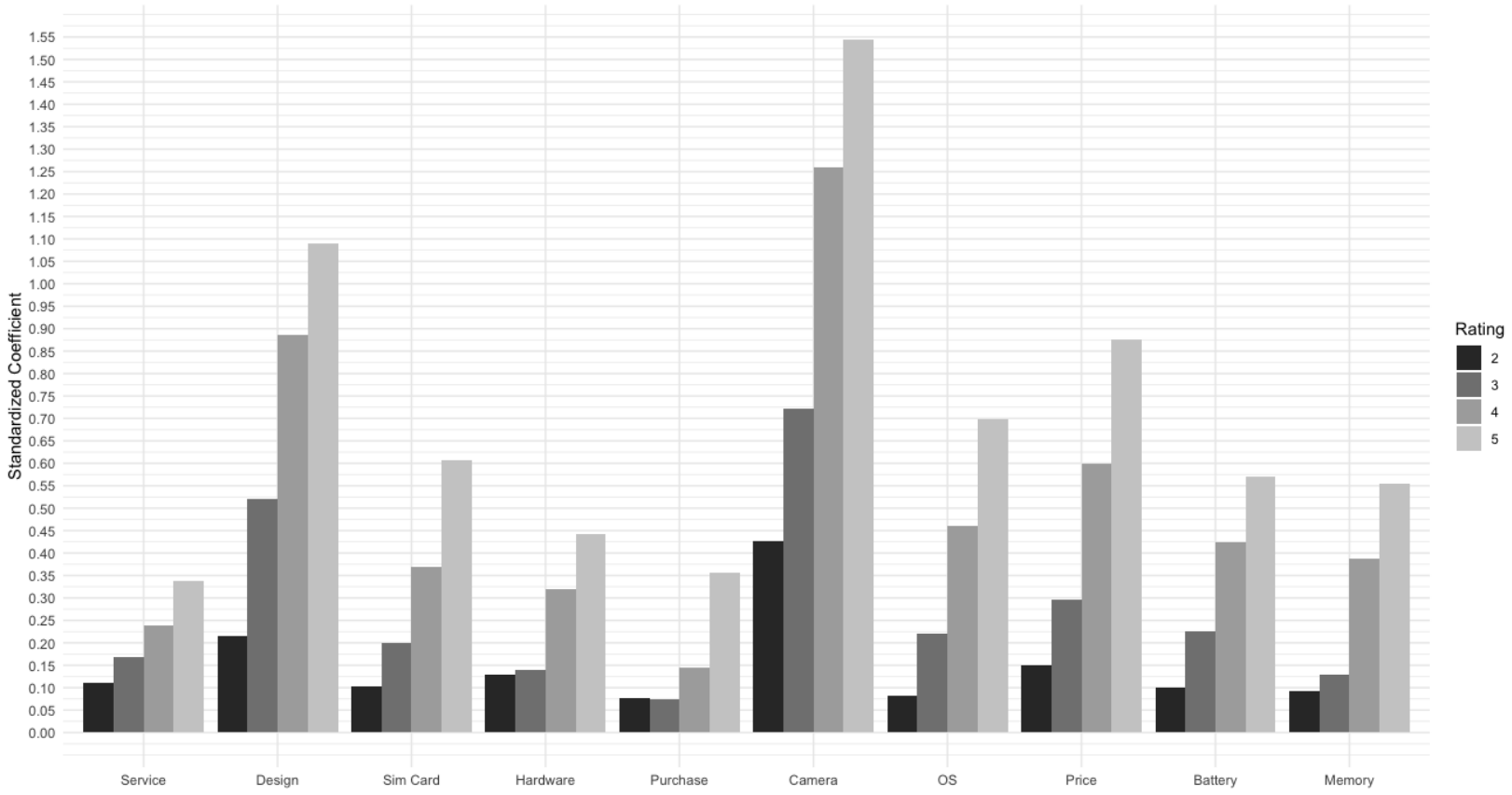

Figure 4: Standardized Coefficients of Multinomial Logistic Regression

5.4 Visualizations of the Estimated Aggregate Customer Preferences

We use two methods to visualize the aggregated consumer preferences calculated with the multinomial logistic regression model. In figure 4, we plot the standardized coefficients for the sentiments towards the topics. In figure 5, we plot the effects display as based on the model by Fox and Andersen (2006).

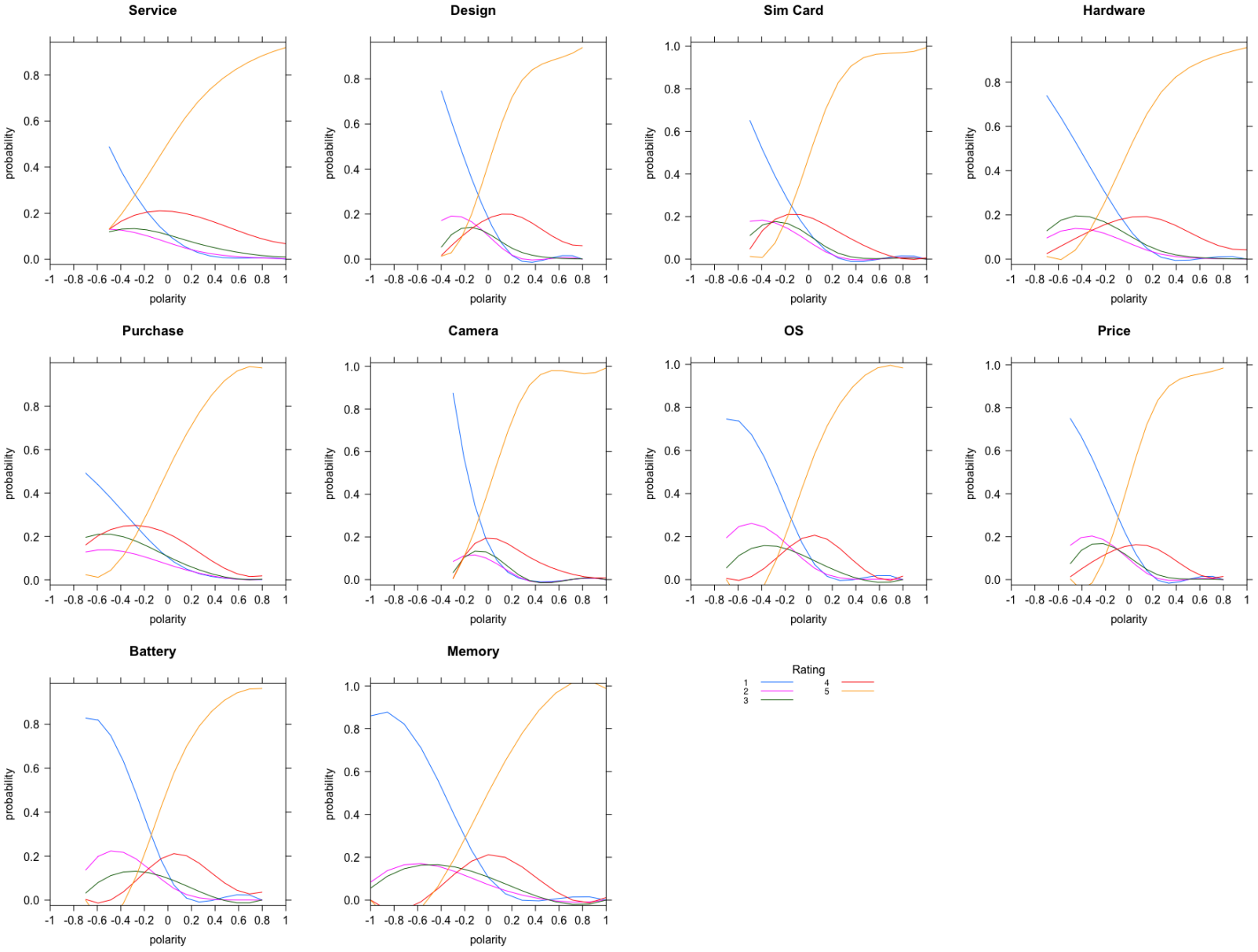

Figure 5: Effects Display Aggregate Customer Preferences

6 Discussion

The final visualizations give us exciting results about the customer preferences in the mobile phone market. We must, however, interpret these results with caution for reasons described in the theoretical framework. People might not write about everything important to them. Well defined and expected features might not be important to write about for people who want to help others because they are already known. Therefore, some critical features might not be incorporated into the model because they are not topics people write about. Besides, people who write for conspicuous reviewing reasons might not write about the attributes that are important to them but only about attributes that signal an image to the reader. People who write for advice seeking or because they want to be heard might also not write about important attributes and express their opinions but alternatively probably writes about a complaint of sorts.

In figure 3 we see that reviews that mention *service*, *hardware* and/or *purchase* are on average more often negative. If we zoom in on the 3,234 reviews that mention both *service* and *purchase*, the proportion of reviews that are rated 1 is even higher with 29.6%. We might expect that people expect a certain level of service and a certain purchase experience with their phones. This expectancy is in line with *must-be* attributes and it therefore seems like at least *service* and *purchase* are attributes of this kind. The rating behaviour with these attributes is also in line with what we expected. We expected that the review is more likely to be negative when mentioned, and we do not expect them to be mentioned often in positive reviews. Because the attributes are still quite broad, we still see them in positive reviews sometimes. An example of this can be a part of an attribute. The warranty that a seller might be unexpectedly good and that might then be an *attractive* attribute in hindsight. This unexpectedly good service might lead to a customer wanting to help the company, wanting to help

the employees, or seeking balance. We expect that these things are the reason that the attributes are also mentioned in positive reviews.

In figure 3, we can also see that reviews that mention *design*, *camera* and/or *price* are on average more often positive. A part of this perhaps has something to do with the conspicuous reviewing motive; we expect that reviews written for this motivation have a higher rating on average, and we expect that people write about certain attributes that signal something to the reader. When somebody writes about the high price of the phone, it conveys that that person might be able to purchase an expensive phone. Besides that, it is possible that the *camera* and the *design* exceeded the expectations, which might lead to the motivations mentioned above; a customer wanting to help the company, wanting to help the employees or is seeking balance. This might explain why these reviews are rated higher on average.

We expect that *memory* is a *one-dimensional* attribute. With the information from figure 3, we have no reason yet to think otherwise. The proportional distribution of the ratings for reviews that mention *memory* is pretty similar to the average line. This means that the attribute is not mentioned inordinately in high or low rated reviews, which is what we expect with an attribute of which effect on the satisfaction holds a linear relationship with its quality. The same seems to be the case for *battery*, which also fits the description of a *one-dimensional* attribute that was given by Aune (2000) in that it is known, specified, measurable and often technical.

In table 9, we see that all the means of polarities are positive. This might be the case because people are more often positive about attributes than negative, but it might also be partly because of a concept called grade inflation (Kwartler, 2017). Grade inflation is a type of response bias whereby a person feels a form of social responsibility to include

some positive words into negative sentences.

The R^2 value of 34.029% indicates that we can explain 34.029% of the variance in the rating with our model. This means that much of the rating is not explained by the sentiments towards the ten identified topics and the brand. This might be because of the aforementioned reasons.

We can, however, still interpret the results of our model with this in mind. The most important reason for using the model was to aggregate consumer preferences. We can look at the different (standardized) coefficients of our model to see which discovered topics/attributes are the attributes that have the most impact on the rating.

In figure 4 we see that *camera* has the highest standardized coefficients for all ratings. A one standard deviation increase in the polarity towards *camera* gives a 1.544 increase in the log odds of the rating being 5 instead of 1. The full table of the standardized coefficients can be found in Appendix B. *design* has the second-highest standardized coefficients, and *price* has the third-highest standardized coefficients.

Interestingly, we see that the standardized coefficient of *memory* for rating 5 is also quite high (0.554). This might be the case due to the conspicuous reviewing motive. We suspect that most people with this motivation write about the attribute *memory* to signal something to the reader. If this is the case, it makes sense that somebody also rates the product high to convey that same signal to the reader. In table 9 we see that *memory* also has the lowest sentiment score that a reviewer has expressed. This might be because it was a lot worse than the person expected and when an individual sees that attribute as a *must-be* attribute, a bad performance lead to high dissatisfaction with the product. This effect is also visible in figure 5, where a decrease in the polarity

increases the probability of the rating being 1 sharply. Besides that, it is also clear that *memory* is the least spoken about attribute in our data, as we can see in table 9.

In table 9, we see that *design* has the highest average polarity. On average, people speak most positively about *design*. This might be because the design of a phone can surprise a customer when using it as opposed to seeing it in a picture. For example, a phone might be more sturdy than one expected, and that can positively influence a customer's opinion towards the phone. As stated before, *design* also has high coefficients whereby a one standard deviation increase in polarity towards the *design* leads to a 1.090 increase in the log odds of the rating being 5 instead of 1.

The final visualization (figure 5 & figure 4) show us a lot of information about the aggregated customer preferences. The method that we propose makes it possible to easily analyze and find these preferences in unstructured reviews.

7 Conclusion

In this paper we present a new combination of state-of-the-art techniques to calculate the aggregated customer preferences from unstructured UGC. Prior literature focuses only on sentiment analysis (e.g., Bagheri et al., 2013), feature extraction (e.g., Hu & Liu, 2004; Li, Qin, Xu, & Guo, 2015) or finding aggregate consumer preferences from structured UGC (e.g., Decker & Trusov, 2010; Xiao et al., 2016).

Our proposed method provides visualizations that are easily interpretable and show what attributes have the most considerable effects. By standardizing the coefficients in the bar plot, we filter out the effect of the more extensive range of polarities towards specific attributes.

This paper uses a relatively simple sentiment analysis technique, and more advanced methods might lead to more accurate results. These methods, however, often need extra (external) information (e.g., Popescu & Etzioni, 2007) to determine polarities, which is not always easily accessible. Unfortunately, our method cannot consider topic-specific sentiments (long can be good for the battery life but might be bad for downloading). Methods that use domain-specific labelled data (e.g. Song et al., 2021) can be helpful to overcome this problem. Since we want our method to be widely applicable, we use a more straightforward method for sentiment analysis.

Because our method uses a multinomial logistic regression, the final output consists of multiple coefficients per topic, which makes interpreting the results a bit harder than when we get a single set of coefficients. Other methods, such as ordinal logistic regression, might solve this problem because they are more parsimonious. However, restrictive assumptions should be tested before these methods can be used.

The focus of this paper is on finding the aggregate consumer preferences towards different attributes. It might, however, also be interesting for practitioners in the marketing field to use this information to discover market structures. Combining the method in our paper with the proposed visualization in Netzer, Feldman, Goldenberg, and Fresko (2012) can show such structures.

References

Agarwal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proc. of the 20th vldb conference* (Vol. 487, p. 499).

Andrzejewski, D., Zhu, X., & Craven, M. (2009). Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th annual international conference on machine learning* (pp. 25–32).

Aune, A. (2000). *Kvalitetsdrevet ledelse, kvalitetsstyrte bedrifter*. Gyldendal akademisk. (ISBN: 978-82-417-1123-7)

Bafna, K., & Toshniwal, D. (2013). Feature based summarization of customers' reviews of online products. *Procedia Computer Science*, *22*, 142–151.

Bagheri, A., Saraee, M., & De Jong, F. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems*, *52*, 201–213.

Bancken, W., Alfarone, D., & Davis, J. (2014). Automatically detecting and rating product aspects from textual customer reviews. In *Proceedings of the 1st international workshop on interactions between data mining and natural language processing at ecml/pkdd* (Vol. 1202, pp. 1–16).

Berger, C., Blauth, R., Boger, D., Bolster, C., Burchill, G., DuMouchel, W., ... Walden, D. (1993). Kano's methods for understanding customer-defined quality. *Center for Quality of Management Journal*, *2*(4), 3–36.

Berger, J., & Iyengar, R. (2013). Communication channels and word of mouth: How the medium shapes the message. *Journal of consumer research*, *40*(3), 567–579.

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical care*, *9*(1), 1–7.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal*

of machine Learning research, *3*, 993–1022.

Bloch, P. H., Sherrell, D. L., & Ridgway, N. M. (1986). Consumer search: An extended framework. *Journal of consumer research*, *13*(1), 119–126.

Brody, S., & Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 804–812).

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, *35*(6), 953–975.

Cattin, P., & Wittink, D. R. (1982). Commercial use of conjoint analysis: A survey. *Journal of marketing*, *46*(3), 44–53.

Decker, R., & Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, *27*(4), 293–307.

De Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Lrec* (Vol. 6, pp. 449–454).

Dichter, E. (1966). How word-of-mouth advertising works. *Harvard business review*, *44*, 147–166.

Engler, T. H., Winter, P., & Schulz, M. (2015). Understanding online product ratings: A customer satisfaction model. *Journal of Retailing and Consumer Services*, *27*, 113–120.

Fellbaum, C. (1998). *1998, wordnet: An electronic lexical database*. MIT Press.

Filip, J., & Kliegr, T. (2018). Classification based on associations (cba)-a performance analysis.

Fox, J., & Andersen, R. (2006). Effect displays for multinomial and proportional-odds logit models. *Sociological Methodology*, *36*(1), 225–255.

Garcia-Pablos, A., Cuadros, M., & Rigau, G. (2018). W2vlda: almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications*, *91*, 127–137.

Harris, M. B., & Huang, L. C. (1973). Competence and helping. *The Journal of Social Psychology*, *89*(2), 203–210.

Harrison-Walker, L. J. (2001). E-complaining: a content analysis of an internet complaint forum. *Journal of Services marketing*.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? *Journal of interactive marketing*, *18*(1), 38–52.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (pp. 856–864).

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spacy: Industrial-strength natural language processing in python.
doi: doi: 10.5281/zenodo.1212303

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 168–177).

Jin, J., Liu, Y., Ji, P., & Kwong, C. K. (2019). Review on recent advances in information mining from big consumer opinion data for product design. *Journal of Computing and Information Science in Engineering*, *19*(1), 010801.

Johnson, R. (1985). Adaptive conjoint analysis. sawtooth software. *Inc., Idaho*.

Juran, J. M. (1992). *Juran on quality by design: the new steps for planning quality*

into goods and services. Simon and Schuster.

Kano, N. (1984). Attractive quality and must-be quality. *Hinshitsu (Quality, The Journal of Japanese Society for Quality Control)*, *14*, 39–48.

Kiss, T., & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational linguistics*, *32*(4), 485–525.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics* (pp. 423–430).

Kovács, B., & Horwitz, S. (2018). Conspicuous reviewing: affiliation with high-status organizations as a motivation for writing online reviews. *Socius*, *4*, 2378023118776848.

Kwartler, T. (2017). *Text mining in practice with r*. John Wiley & Sons.

Lampel, J., & Bhalla, A. (2007). The role of status seeking in online communities: Giving the gift of experience. *Journal of computer-mediated communication*, *12*(2), 434–455.

Li, Y., Qin, Z., Xu, W., & Guo, J. (2015). A holistic model of mining product aspects and associated sentiments from online reviews. *Multimedia Tools and Applications*, *74*(23), 10177–10194.

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism management*, *29*(3), 458–468.

Loria, S. (2018). textblob documentation. *Release 0.15*, *2*.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank.

Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., & Taylor, A. (1999). Penn treebank-3. *Linguistic Data Consortium, Catalog# LDC99T42*.

Marwell, G., & Oliver, P. (1993). *The critical mass in collective action*. Cambridge

University Press.

McWilliam, G. (2000). Building stronger brands through online communities. *MIT Sloan Management Review*, *41*(3), 43.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Miller, G. A. (1998). *Wordnet: An electronic lexical database*. MIT press.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521–543.

Norvig, P. (2007, Feb). *How to write a spelling corrector.* Retrieved from `http://norvig.com/spell-correct.html`

Oxford Text Archive. (2009, Jan). *British national corpus.* University of Oxford. Retrieved from `http://www.natcorp.ox.ac.uk/`

Pang, B., & Lee, L. (2004). *Polarity dataset v2. 0.*

Peddibhotla, N. B., & Subramani, M. R. (2007). Contributing to public document repositories: A critical mass theory perspective. *Organization Studies*, *28*(3), 327–346.

Plant, R. (2004). Online communities. *Technology in society*, *26*(1), 51–65.

Popescu, A.-M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. In *Natural language processing and text mining* (pp. 9–28). Springer.

Poria, S., Cambria, E., Ku, L.-W., Gui, C., & Gelbukh, A. (2014). A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (socialnlp)* (pp. 28–37).

Porter, M. (2017). Wom or ewom, is there a difference?: an extension of the social communication theory to consumer purchase related attitudes.

Price, L. L., Feick, L. F., & Guskey, A. (1995). Everyday market helping behavior.

Journal of Public Policy & Marketing, *14*(2), 255–266.

Project gutenberg. (1972, Jul). University of Illinois. Retrieved from `http://www.gutenberg.org/`

Řehůřek, R., & Sojka, P. (2011). Gensim—statistical semantics in python. *Retrieved from genism.org*.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth acm international conference on web search and data mining* (pp. 399–408).

Song, W., Wen, Z., Xiao, Z., & Park, S. C. (2021). Semantics perception and refinement network for aspect-based sentiment analysis. *Knowledge-Based Systems*, *214*, 106755.

Sundaram, D. S., Mitra, K., & Webster, C. (1998). Word-of-mouth communications: A motivational analysis. *Advances in Consumer Research*, *25*.

Titov, I., & McDonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization. In *proceedings of acl-08: Hlt* (pp. 308–316).

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 human language technology conference of the north american chapter of the association for computational linguistics* (pp. 252–259).

Van Rijsbergen, C. J., Robertson, S. E., & Porter, M. F. (1980). *New models in probabilistic information retrieval* (Vol. 5587). British Library Research and Development Department London.

Veblen, T. (1899). The theory of the leisure class.

Wang, Y., & Fesenmaier, D. R. (2004). Towards understanding members' general participation in and active contribution to an online travel community. *Tourism management*, *25*(6), 709–722.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., ... Franchini, M. (2013). Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, *23*.

Whiting, A., Williams, D. L., & Hair, J. (2019). Praise or revenge: why do consumers post about organizations on social media. *Qualitative Market Research: An International Journal*.

Wikimedia Foundation. (2002, Dec). *Wiktionary.*

Wu, P. F. (2019). Motivation crowding in online product reviewing: A qualitative study of amazon reviewers. *Information & Management*, *56*(8), 103163.

Xiao, S., Wei, C.-P., & Dong, M. (2016). Crowd intelligence: Analyzing online product reviews for preference measurement. *Information & Management*, *53*(2), 169–182.

Zhai, Z., Liu, B., Xu, H., & Jia, P. (2011). Clustering product features for opinion mining. In *Proceedings of the fourth acm international conference on web search and data mining* (pp. 347–354).

Appendix A: P-Value for Wald Statistic for All Coefficients

Rating	2	3	4	5
Service	0	0	0	0
Design	0	0	0	0
Sim Card	0.03	0	0	0
Hardware	0	0	0	0
Purchase	0.002	0.002	0	0
Camera	0.005	0	0	0
OS	0	0	0	0
Price	0	0	0	0
Battery	0.001	0	0	0
Memory	0.004	0	0	0
Blackberry	N/A	N/A	N/A	N/A
Blackview	0	0	0	0.006
Blu	0	0	0	0
Google	0	0	0	0
HTC	0	0	0	0
Huawei	0	0	0	0
LG	0	0	0	0
Microsoft	0.002	0	0	0.002
Motorola	0	0	0	0
Nokia	0	0	0	0
Oneplus	0.054	0	0.001	0.283
Oukitel	0.002	0	0	0.710
Samsung	0	0	0	0
Sony	0	0	0	0
Xiaomi	0	0	0	0.33

Appendix B: Standardized Coefficients for the Multinomial Logistic Regression

Rating	2	3	4	5
Service	0.111	0.168	0.238	0.338
Design	0.216	0.520	0.885	1.090
Sim Card	0.104	0.200	0.369	0.607
Hardware	0.130	0.139	0.319	0.411
Purchase	0.076	0.075	0.144	0.355
Camera	0.427	0.723	1.259	1.544
OS	0.082	0.221	0.461	0.698
Price	0.150	0.296	0.600	0.876
Battery	0.099	0.226	0.424	0.569
Memory	0.092	0.130	0.386	0.554