Gazi University

Journal of Science

PART B: ART, HUMANITIES, DESIGN AND PLANNING

http://dergipark.gov.tr/gujsb

Extracting Data-Driven User Segments and Knowledge by Using Online Product Reviews

Serkan GÜNEŞ[1,*]

[1] *0000-0003-4377-528X, Gazi University Design Application and Research Center, Ankara/Turkey*

Article Info

Received: 21/02/2023
Accepted: 27/03/2023

Keywords

Big Data,
Decision making,
Design knowledge,
Industrial design,
User behavior

Abstract

With the growth of e-commerce, consumer reviews are becoming more widely available and influential. These valuable online product reviews (OPR) show any product issues and contain unique and hidden user information fragments that designers can use in decision-making. OPRs are often unstructured, massive, disorganized, and highly detailed. OPRs are voluntary production and are available in large numbers, publicly available, and accessible. These features increase the number of samples and save money and time for designers to understand the user. The analysis of OPRs is done with AI-supported text analysis tools, especially if many reviews are to get through. In this study, user demographics and opinions about the product are extracted through text mining and statistical methods through the OPRs of a sample product. The data analysis results provided valuable information about the users and had the potential to develop new knowledge and generate new ideas for the design process. By arguing for the merit of adding Big Data analysis to the design process, first, valuable user information content contained in OPRs has been revealed. Secondly, it was possible to express user stacks as clusters with similar characteristics. Finally, it has been revealed that demographic user clusters become homogenized after the product experience, and the initially disjointed clusters begin to resemble independently from the demographic clusters due to independent product/aspect evaluations.

1. INTRODUCTION

The user's primary position in the design field assigns a central role at all stages (Margolin, 1997; Turner and Turner, 2011). Revealing the potential users with limited data is insufficient to frame the design problem. Therefore, the issue of who could be the potential product user is a problem that requires a qualified and continuous information flow. Yet, designers have always been concerned with how design is used and by whom (Cooper, 2011). Widespread Internet usage has allowed users to present valuable product views, and access to a mass of publicly available data has become possible. This data is valid only if a designer knows how to make design decisions to increase user engagement and become more receptive to the users' wants and needs. Hence, Big Data (BD) should be considered an area that helps develop new information about the design process and generate new insights, as it provides valuable information about end users. The sources of information to identify the potential user vary (Oygür, 2018), and there is no right way to create personas (Almaliki, Ncube, and Raian, 2015). The effective use of this information is highly dependent on the designer's work experience, subjective judgment, and the ability to access and interpret varying resources. While efforts to identify users have been extensive, the integration of data analytics into design processes has been slower than expected. Despite that, BD can be used in the early design phase to summarize user characteristics quickly, efficiently, and accurately and correlate user characteristics and experiences (Güneş, 2020). BD analytics uses various data sources, such as social media networks, product discussion platforms, or online product reviews (OPR). This helps build the intellectual core of the potential user, analyze their future direction and show a wealth of possible information. Information from OPRs serves two narrative purposes, similar to the process of persona representation (Miaskiewicz and Kozar,

[*] Corresponding author: serkangunes@gazi.edu.tr

2011). The first is to find and cluster the real people behind the data content. The second is to present a vivid story about a real person's product needs and experiences.

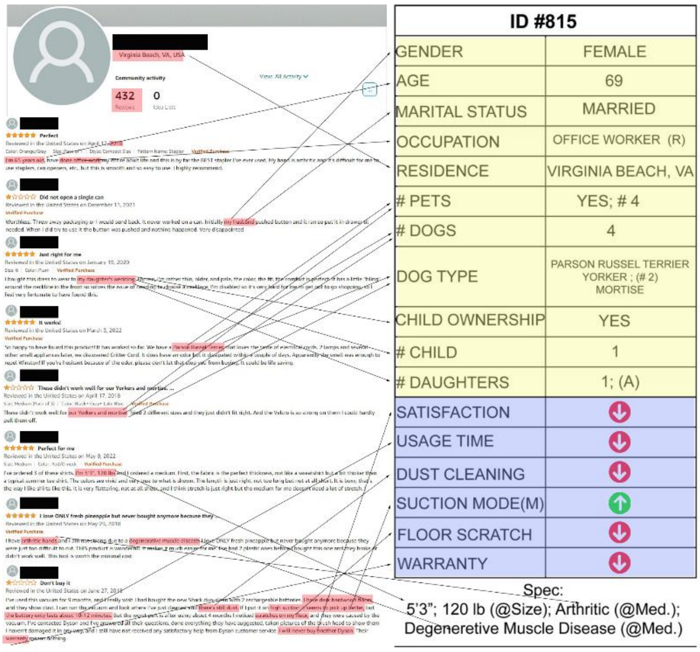
The user data usually include static demographic data and dynamic OPR. Static data are the primary user data to draw a general user portrait. Dynamic data reveal the user's product experience and the actual demand. User data sources are diverse: traditional corporate data based on reporting, interactive data from IoT-based devices, internet data such as OPRs from e-commerce, and data about users from social media. E-commerce platforms allow potential customers to monitor and share users' experiences, and communicate to make purchasing decisions. As informative and recommending, OPRs are essential in purchasing decisions and sales (Park, Lee, and Han, 2007) and provide insights into business intelligence to capture user information and feedback used in conceptual design and personalization (Zhan, Loh, and Liu, 2009) for product decision-makers. BD often does not refer to any particular data type; it is highly diverse and dispersed to be managed and analyzed with conventional methods. The issue is not the data size but how to access it, what to do with this information, and what kind of resource to use to gain these insights to enable users to participate directly in product design and development. Designers who want to take advantage of the possibilities of BD should develop an ability to analyses data and internalize and take pre-processing into account to understand their hidden meanings.

From the perspective of enriching data based on real users as a critical designer data source, the study explores ways to create user profiles and make their feedback visible through OPRs. The first part of the study will discuss OPRs as metadata and their potential. The following sections will present preprocessing, analysis, and statistical methods with different layers based on BD through a sample product to give new insight into the design process.

2. OPS AND POTENTIALS

E-commerce facilitates consumers to provide and read OPRs and voice their complaints and opinions on various goods (Boush and Kahle, 2001). These giant platforms are electronic word-of-mouth communication that gathers reviewers and potential customers around the OPRs. The importance of OPRs to the audience stems from the need for self-expression and search recommendations from outside sources (Rogers, 2003). The information content of OPRs reveals their importance to the design field, as negative comments may allow decision-makers to take corrective actions. In this study, the OPRs on Amazon.com are discussed in terms of static and dynamic. User profile information and personal information fragments included in all profile comments and shared by the reviewer were used to generate static user data. Dynamic OPR data were used to determine the reviewers' opinions on the sample product.

A standard OPR consists of several sections. Metadata includes username/nickname, residence, date, and rating; if available, verified the purchase mark and helpfulness score. The rest is the title and the main review. The reviewer's profile can access the user's public profile and previous product reviews. These profiles, if made public, include user information, impact value, and badges. There was no standard format for OPRs. Reviews can be one-word evaluations, pros, and cons based on detailed product analyses, post-sale experiences, and updates. Each OPR may also contain direct/partial information or clues about the user (Table 1). For example, a reviewer may state her age as 35, or if she uses the phrase My husband it may infer that the reviewer is married. In individual OPR, this information is often limited and scarce. Therefore, it is challenging to produce a general framework about who the reviewer is. However, obtaining a wealth of content clearly defining the user is possible when combined with the personal information scrapped in the reviews. Metaphorically, this process is like a jigsaw puzzle where each information content is treated as a piece. Then, combining it with other information may solve the puzzle. Since reviews have an extensive historical range, each information content, such as age, and marital status, must be evaluated on its terms to be up-to-date and valid.



[redacted]
Virginia Beach, VA, USA

Community activity

View: All Activity

432 Reviews

0 Idea Lists

[redacted]
★★★★★ Perfect
Reviewed in the United States on April 12, 2018
Color: Orange/Gray | Size: Pack of 1 | Style: Compact Size | Pattern Name: Stapler | Verified Purchase
I'm 65 years old, have done office work my entire adult life and this is by far the BEST stapler I've ever used. My hand is arthritic and it's difficult for me to use staplers, can openers, etc., but this is smooth and so easy to use. I highly recommend.

[redacted]
★☆☆☆☆ Did not open a single can
Reviewed in the United States on December 11, 2021
Verified Purchase
Worthless. Threw away packaging or I would send back. It never worked on a can. Initially my husband pushed button and it ran so put it in drawer til needed. When I did try to use it the button was pushed and nothing happened. Very disappointed

[redacted]
★★★★★ Just right for me
Reviewed in the United States on January 19, 2020
Size: 6 | Color: Plum | Verified Purchase
I bought this dress to wear to my daughter's wedding. For me, I'm rather thin, older, and pale, the color, the fit, the comfort is perfect. It has a little "bling" around the neckline in the front so solves the issue of needing to choose a necklace. I'm disabled so it's very hard for me to get out to go shopping, so I feel very fortunate to have found this.

[redacted]
★★★★★ It works!
Reviewed in the United States on March 3, 2022
Verified Purchase
So happy to have found this product! It has worked so far. We have a Parson Russell Terrier that loves the taste of electrical cords. 2 lamps and several other small appliances later, we discovered Critter Cord. It does have an odor but it dissipated within a couple of days. Apparently the smell was enough to repel Winston! If you're hesitant because of the odor, please don't let that stop you from buying. It could be life saving.

[redacted]
★☆☆☆☆ These didn't work well for our Yorkers and mortise. ...
Reviewed in the United States on April 17, 2018
Size: Medium (Pack of 3) | Color: Black+Gray+Lake Blue | Verified Purchase
These didn't work well for our Yorkers and mortise. Tried 2 different sizes and they just didn't fit right. And the Velcro is so strong on them I could hardly pull them off.

[redacted]
★★★★★ Perfect for me
Reviewed in the United States on May 9, 2022
Size: Medium | Color: Red/O-neck | Verified Purchase
I've ordered 3 of these shirts. I'm 5'3", 120 lbs and I ordered a medium. First, the fabric is the perfect thickness, not like a sweatshirt but a bit thicker than a typical summer tee shirt. The colors are vivid and very true to what is shown. The length is just right, not too long but not at all short. It is boxy, that's the way I like shirts like this. it is very flattering, not at all sheer, and I think stretch is just right but the medium for me doesn't need a lot of stretch. I

[redacted]
★★★★★ I love ONLY fresh pineapple but never bought anymore because they ...
Reviewed in the United States on May 29, 2018
Verified Purchase
I have arthritic hands and I am not strong due to a degenerative muscle disease. I love ONLY fresh pineapple but never bought anymore because they were just too difficult to cut. THIS product is wonderful. It makes it much easier for me. I've had 2 plastic ones before I bought this one and they broke or didn't work well. This tool is worth the minimal cost.

[redacted]
★☆☆☆☆ Don't buy it
Reviewed in the United States on June 27, 2018
I've used this vacuum for 9 months, and I really wish I had bought the new Shark duo-clean with 2 rechargeable batteries. I have dark hardwood floors and they show dust. I can run the vacuum and look where I've just cleaned and there's still dust. If I put it on high suction it seems to pick up better, but the battery only lasts about 10-12 minutes. But the worst part is after using about 4 months I noticed scratches on my floor, and they were caused by the vacuum. I've contacted Dyson and I've answered all their questions, done everything they have suggested, taken pictures of the brush head to show them I haven't damaged it in any way, and I still have not received any satisfactory help from Dyson customer service. I will never buy another Dyson. Their warranty means nothing.

ID #815	
GENDER	FEMALE
AGE	69
MARITAL STATUS	MARRIED
OCCUPATION	OFFICE WORKER (R)
RESIDENCE	VIRGINIA BEACH, VA
# PETS	YES; # 4
# DOGS	4
DOG TYPE	PARSON RUSSEL TERRIER YORKER ; (# 2) MORTISE
CHILD OWNERSHIP	YES
# CHILD	1
# DAUGHTERS	1; (A)
SATISFACTION	↓
USAGE TIME	↓
DUST CLEANING	↓
SUCTION MODE(M)	↑
FLOOR SCRATCH	↓
WARRANTY	↓

Spec:
5'3"; 120 lb (@Size); Arthritic (@Med.);
Degeneretive Muscle Disease (@Med.)

Figure 1. *User clusters of static data*

3. METHOD

User content within the digital world presents a passive or active digital footprint. Active footprints contain data that is intentionally generated. With every piece of content produced, this footprint grows and is recorded on the servers. Voluntary and consensual OPRs are other common ways to expand the digital footprint. The online exchange of information is subject to the principles of informed consent. Although the desire to protect private information is a natural human trait, many consumers can consciously share in exchange for visibility, reputation, promotion offers, and special deals (Punjcası, 2007). Therefore, any user may state, for example, that he lives in a big house or as a 78 old guy, I can easily carry the product between the floors to support his conclusion about product performance. Product reviews provide more than just insight into individual product perceptions because of the active footprints describing the consumer. These footprints contribute positively to understanding the consumer and a product's performance and design. By using Amazon Services, reviewers agree to the conditions of use. The data used within the scope of this study are OPR-based active footprints that are deliberately shared by users and presented to the public, even though they can hide.

The study included the Dyson V7 Animal Cordless Vacuum Cleaner as a product and its OPRs through amazon.com. OPRs with verified purchase marks (n=2844) were obtained by crawling software on 05.13.2022 and converted into csv format. This dataset includes the username, the star-rating, the date, the summary, the full review, and the helpfulness scores. In further analysis, the user name was used to extract gender partially, and the date of the comment was used to determine the user's current age. The summary and full comment text created the user's profile and product views.

The star score was used to calculate the correlation between the given rating and the sentiment polarity of the text. In this context, a high correlation shows the quality of OPR content in fraud detection, spam, and credibility and produces beneficial results for analysis. For each reviewer, an ID number and folder (public profile and all profile comments) were assigned to hide the profile. These profiles may contain ID's real name/nickname, residence, self-created About content, and all OPRs made by ID on amazon.com. To determine the user profile in the future, all these ID OPRs were obtained by crawling software, and stored in the relevant ID folder. The first analysis determined that there were 146,437 product OPRs of 2844 IDs and 51 on average for each.

Comprehensive analysis data consisted of 18 static user criteria: the # of comments, gender, age, occupation, marital status, residence, pet ownership, # of dogs, breed, # of cats, house-type, child ownership, # of children, daughters, boys, and grandchildren, with the leading product and general features. The # of comments is direct numerical data and constitutes the # of lines for that ID. Age data may vary in expression as data to be extracted from reviews. As a challenging process, there are different methods to distinguish age from other number expressions (Zhu et al., 2012). In the age extraction, the number indicating age does not differ from other numerical expressions (It weighs about 8 lbs., been using an Electrolux for 50 years), the variety of age expressions (Getting a 32-year-old guy, I'm a single 40-year-old woman), general age expressions (Toddler, elderly) or the term of age referring to another individual or thing (My 86-year-old father, I have five kids under 8 yrs old) may lead to misclassification. While most comments include the exact age statement, some have a wide range (I am in my 50s). Therefore, the US/The Census Bureau age groups were used in the analysis (15–24, 25–34, 35–44, 45–54, 55–64, 65+, 75+, 85+, 100+). A text segment for the age expression generally includes the first singular pronoun (I, im, lam) or possessive pronouns (my son, our dog) and the numerical age expression followed by the year expression (Year, years, yo, yrs, y/o, s) and name (dad, grandchild) if any. The Allen NLP Named Entity Recognition (NER) model was used over Python to determine the age. NER is a machine learning technique that semantically divides the text, identifying parts of a sentence that fits into predefined categories. It is relatively easy to extract gender data by age. For example, the expression my wife indicates that the ID is male and does not require extra processing. However, gender nouns used with the first singular pronoun (woman, man, guy, lady, father, mother) convenience the gender of the profile. The actual names of the reviewers (Mary, George) provided great opportunities to determine their gender. Damegender software was used over Python for gender determination from the name. For the marital status extraction, with regular expressions primarily expressing marital status (single, girlfriend, boyfriend, married, husband, wife, divorced, ex-husband/wife, divorcee, widowed), an algorithm that uses a keyword-based pattern (The US/The Census Bureau scale) matching is used. Data for occupation and residence were analyzed over the Meaning Cloud Topic Extraction API. The Standard Occupational Classification System was used for classification, and state dictionaries were used for residence information. Residence information is entered into the ID directory as an entity and occupation information as a concept. Topic Extraction is accomplished by combining a set of complex natural language processing techniques that allow obtaining a text's morphological, syntactic, and semantic analyses and using them to identify essential elements of different types. If the residence information is on the profile page of the ID, it is processed directly. Even at the county level, data are available for residence determination (e.g., Spring Lake, NJ), a state-level classification has been made for privacy, and state codes have been set up numerically (e.g., [NJ] = 33). Topic Extraction was used again for pet ownership. ID's pet ownership is processed as a 1–0 nominal scale, and the number of pets is included as scale data. The expressions of dog and cat in the texts were included in the classification of mammals as concepts. Pet numbers are numerically added to the ID folder as Quantity Expression. A keyword-based pattern matching algorithm was used over a keyword list that included a dog breed list. A similar method was applied for children and grandchildren's ownership, their number, and gender. This dictionary was created for the types of houses regarding the US/The Census Bureau classification (e.g., single-family, condo, apartment) and was matched with the expressions in the texts.

Of all these analyses, with fewer occupation data (16%) out of IDs obtained, occupation data were retained but not used in further research. Thirty dog breeds based on n-gram were determined on a scale; Labrador (12.8%), German Shepherds (11.6%), Golden Retrievers (7.3%), and Terrier (7.3%) were found. IDs express dog breeds as a dominant indicator of express contaminants for the fur type. However, as not all IDs have breed data, the dataset was considered missing. The leading product criterion is the declared vacuum cleaners owned by IDs and used for in-brand versions and inter-brand comparisons. This criterion did not provide enough data and was accepted as missing data and excluded from the assessment. General features criteria are the features that the ID associates with the product. These features include health data (such as arthritis, chronic low back pain), body measurements (such as weight, height, size), and n-gram sets, which the user indicates explicitly for himself (as a Dyson fan, neat freak, meticulous, directly related car brand for cleaning). These data were excluded from the analysis but were stored in folders for further research, such as the arthritis-product weight relationship. After the pre-analysis, it was seen that 944 (33%) of 2844 IDs did not have missing data. Thus, a new analysis dataset was created with 944 ID. This dataset was formed from gender, age, marital status, residence, # of pets, dogs, cats, house-type, child ownership, # of children, daughters, boys, and grandchildren.

4. GENERAL OUTCOMES

The dataset evaluation determined that 56% of the IDs were women. The distribution of age groups is 35–44 (19%), 55–64 (18%), 25–34 (18%), 65+ (17%), 45–54 (17%), 75+ (11%), 15 -24 (0.03%), 85+ (0.03%). 53.7% of IDs are married, 23.9% are divorced, 22.1% are single, and 0.3% are engaged or widowed. 56% are not pet owners. 33% are dog owners with an average of 1.7. 13.8% are cat owners, with an average of 1.8. 5.2% are both dog and cat owners. The declaration shows that 68.5% have children, an average of 1.67. The average for daughters is 1.19 and 1.18 for boys. 66 IDs have different numbers of both daughters and boys. The count of grandchildren ownership is 77 and an average of 1.6. The highest ID residence rate is the single-family house (28.5%). Apartment and condo rates are 21.8%. Single-family house with multi-floor rate is 14%. The farmhouse rate is 15.9%, and the single-family house with a pool rate is 14%. The rate of moveable dwellings such as a home, boat, or RV is 0.03%. The three states with the highest number of states of residence are California (6.2%), New York (4.5%), and Florida (4.2%), and the lowest one is Iowa (0.5%). All these percentage evaluations are static and do not create ID sets as clusters. Therefore, these 13 criteria were subjected to cluster analysis.

5. DEMOGRAPHIC CLUSTERS

After the pre-processing, 944 IDs were subjected to cluster analysis via the SPSS/Direct Marketing/Cluster Tool module, which allows for creating contacts and assigning specific people to each cluster. This exploration tool reveals natural groupings (or clusters) within your data to identify customer groups based on demographics and purchasing characteristics. The number of clusters was determined automatically. SPSS used a two-step algorithm, and five distinctive clusters were obtained (Table 1). The silhouette value was 0.5 (Fair) as an acceptable value.

The most determining factors in the predictive importance ranking over 13 entries were the # children and pets. The importance of the criteria is shown in Figure 2 with color gradation. The age group, gender, and marital status were of medium importance, whereas house-type and residence location were low. This situation is not surprising, yet the product is a vacuum cleaner. The density of the house carries children and animals to the top as a determinant of pollution. The increase in household pollutants increases pollution. Regarding gender, this product was gender-neutral. There was no dominant gender, and a balanced distribution was observed. Regarding the age group, there was a balanced percentage distribution among IDs. This product is seen as a product independent of the age group. Interestingly, the house type received a low value in the importance evaluation.

Table 1. *OPR information contents as static and dynamic data*

CLUS. ID	SIZE	GENDER	AGE GROUP	MARITAL STATUS	RESIDENCE	HOME TYPE	PET	#DOG	#CAT	CHILD	#CHILD	#DAUG	#SON	#GRAND CHILD	BEST REAL SAMPLE
1	38,5% (363)	MALE (65,9%)	35-44 (45,9%)	MARRIED (68,1%)	CALIFORNIA (8,4%)	SINGLE-FAMILY HOUSE (34,3%)	YES (85,6%)	1,3	0,55	YES (100,0%)	1,66	0,73	0,52	0,02	ID#216
2	24,9% (235)	FEMALE (76%)	35-44 (29,9%)	MARRIED (53,1%)	TEXAS (5,5%)	SINGLE-FAMILY HOUSE (39,4%)	NO (100,0%)	0	0	YES (100,0%)	1,59	0,74	0,64	0,03	ID#273
3	17,6% (166)	MALE (55,1%)	65+ (17,6%)	MARRIED (41,0%)	CALIFORNIA (6,1%)	SINGLE-FAMILY HOUSE (23,7%)	NO (100,0%)	0	0,01	NO (100,0%)	0	0	0	0,01	ID#709
4	13,5% (127)	FEMALE (77,4%)	25-34 (21,7%)	SINGLE (53,2%)	NEW JERSEY (7,2%)	FLAT (26,4%)	YES (81,1%)	1,13	0,57	NO (100,0%)	0	0	0	0,01	ID#343
5	5,6% (53)	FEMALE (86,0%)	65+ (62,3%)	MARRIED (75,5%)	FLORIDA (7,5%)	SINGLE-FAMILY HOUSE MULTI F, (26,4%)	NO (93,2%)	0,23	0,04	YES (96,2%)	0,83	0,32	0,25	1,55	ID#302

However, it was expected that the type of house would be decisive for a household appliance. This may be due to specific reasons. First, this product is generally used with other high-capacity home products because of its unique features. Cleaning the whole house in the 20 min, promised by the manufacturer does not seem possible. Second, product performance is generally evaluated by the process in which it runs and is not about size. However, the house's features, such as stairs, have been the primary determinant in the attribute of product features by IDs.

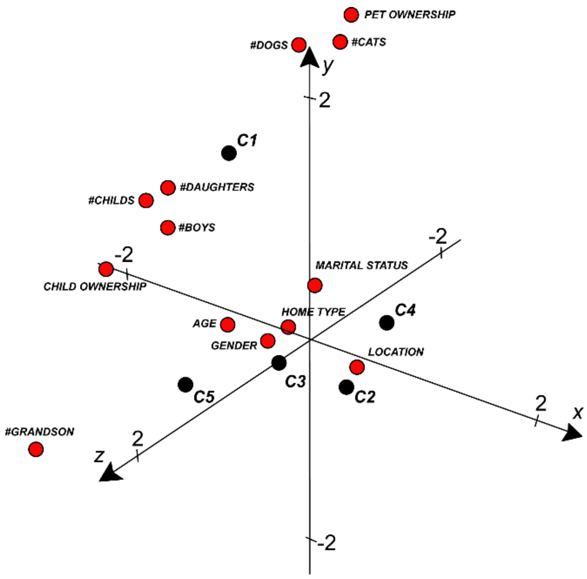


Figure 2. *Bi-Plot of the relationship between 13 criteria and clusters*

In Figure 2, the relationship between 13 criteria and clusters is seen in the Bi-Plot of the Correspondence Analysis. The proportion of inertia for the analysis was 0.896 cumulatively, with a 10% data loss in 3D. In the investigation, it was observed that the most similar clusters were C2 and C4. C1 differs from the others in terms of child and animal ownership. C3 and C5 differ from the others in terms of age groups; C5 is remarkably different regarding grandchild ownership. Marital status and household type were weakly effective in distinguishing between C2 and C4. The most influential factors in forming clusters are the ownership and number of children, pets, and grandchildren far from the origin. Other criteria close to the origin have low Point to Inertia of Dimension values.

The C1 is the most significant (38.5%) cluster, with 363 IDs (Table 1). The members of the cluster are primarily married men aged 35–44. Almost all of the cluster members have pets, and all have children. Their type of residence is a single-family house, and they live in California above average. This group has the highest rate of pets, and the number of children is high. Exceptionally, very few of them were found to have grandchildren. The C2 had an ID rate of 24.9%. Although the proportion of women seems to be higher, the gender distribution is balanced. The age group was between 35 and 44. The marriage rates were close to average. They do not have a pet, but all of them have children. 25% of 48 IDs who declare to live in Texas are in this group. The C3 consisted of 166 IDs. The 65+ age group was predominantly males. No members of this cluster have children or pets. It consists of married people living in the Single Family House. C4, a relatively small group, consists of 127 IDs. This group, between the ages of 24–34, predominantly female, is single and generally lives in a co-op, apartment, or condo. They did not have children, but all had pets. The Northeast region's predominantly residential area was determined. The C5 consists of 65+ married women. Pet ownership was low, but all had adult children. It is a primarily residential area in the southeast region. Grandchild ownership is at its highest. Because of the cluster analysis, each ID was assigned to its determined cluster folder. Then, the ID that best represents that cluster was searched for each. The most representative IDs are presented below by hiding and generalizing the information that indicates their identity.

C1/ID#216(111OPRs): A 43-year-old married man graduated from *** University *** Department. He is the business development director and his wife ***. He lives in the city of *** in the western US. They have one puppy and two boys and live in a house with a private garden. He is interested in sports and considers the Dyson V7 an excellent product.

C2/ID#273(51OPRs): A 37-year-old married woman from the Midwest region *** county and *** city. She is an art teacher with administrative duties. She has no pets. She lives with her 6-year-old daughter and 8-year-old son in a house with a detached garden. They have a Shark vacuum cleaner at home, but she thinks the Dyson V7 is a great product.

C3/ID#709(23OPRs): A 70-year-old married retired mechanic man living in the *** district and ***city in the Northeastern. He lives in a detached house without children or pets. He describes the Dyson V7 as a very easy product to use and install.

C4/ID#343(14OPRs): A 27-year-old Italian immigrant single woman. She is a graduate of *** University *** Department. She lives in the Northeastern region of the *** county and *** city in an apartment with a cat and an Italian coonhound dog. The Dyson V7 product broke within two months. It always shuts down even though it is charged. She thinks that the product is not worth the money.

C5/ID#302(105OPRs): A 68-year-old married woman living in the Southeast owned a 3-story house. She has two daughters and a son. She has no pets and two or more grandchildren. She does artistic work at home as a hobby with clay. She has been using Dyson for years. She thinks the V7 is an excellent product for entering small spaces and cleaning under beds. According to her, hard floor performance is high, and it is easy to empty the canister.

6. DLSA AND ABSA ANALYSIS

At this study stage, 944 OPRs for the Dyson V7 product were evaluated as dynamic data. To test data security, first, the comments of all IDs were subjected to DLSA analysis, and a general polarity value was reached and compared with the star rating given by the IDs. The DLSA concerns whether a document expresses a negative or positive emotion (Güneş, 2020). Within the survey scope, NLP and machine learning-based AYLIEN Analyze Sentiment operator in RapidMiner were used for each interpretation's polarity value (positive, neutral, or negative). The correlation between the ID star rating and the Text polarity was calculated as 0.813. This value, which represents a high confidence level, indicates a match between the given star value and the polarity of the text. There may be some reasons why this correlation value is not 1. However, one of the most important reasons is that users change their reviews or star ratings over time (such as the deterioration of the product from positive to negative over time or the immediate replacement of the product from negative to positive by the manufacturer), without updating their reviews or star ratings. DLSA analysis is a general polarity value on the document, but it is insufficient to determine the reasons. The general evaluation is the whole of the positive or negative opinions of the product's features. Therefore, Aspect-Based Sentiment Analysis (ABSA) was used in the second step. The basic logic of ABSA is to find the polarity of the commentator's views on aspects of an idea to determine the reasons behind it. ABSA has two methods for this analysis: domain-independent solutions and the use of domain-specific knowledge. Domain-independent solutions are not very efficient as they use general dictionaries. However, the domain-specific entity-concept dictionary can produce effective results since it is more focused. In this study, both domain-independent solution dictionaries and domain-specific knowledge dictionaries were used. All ID comments were analyzed using Topic Extractor software to generate the domain-specific dictionary. The topic extraction discovers keywords in documents or databases that capture the gist of the text. It identifies "keywords" and "concepts" for the given input based on the frequency and linguistic patterns in the text and ranks them according to their relative importance. Simultaneously, the keyword pool was diversified by creating n-grams on Rapidminer. Because of these evaluations, 47 pairs of key attributes, positive and negative, were determined for the Dyson V7 product (Figure 3). This operator analyzes the text attribute of IDs and assigns a sensitivity polarity classification (P+, P, Neu, N, and N+) for each identified attribute. This study accepted P+ and P values as 1, Neu 0, and -1 points for N and N+ values. In other words, if an ID made a positive comment on the suction performance of the product, 1 point is scored as a cell value. -1 point if the ID made a negative comment and 0 points if there was no evaluation. After calculating the contextual sensitivity score for each ID attribute, a customized spreadsheet was created to find the analysis-based cause-effect model of the pros and cons of performing a Correspondence Analysis. This table numerically expresses the positive and negative scores of each ID for each attribute, if any. Some ID comments can be only 2–3 words in a text as a general review (Great product, it sucks). Since any attribute does not support these comments, the polarity value is entered directly into the column corresponding to the independent General Satisfaction (+, -) attribute.

7. CLUSTER-INDEPENDENT PRODUCT PROPERTIES

An excel-based contingency table consisting of 944 ID rows and 94 Attribute columns (47 positive and negatives) resulting from ABSA has been evaluated. Below are the core values of the IDs for 47 attributes. Figure 3 lists the most expressed features at the bottom and the least at the top. In the middle, the attribute is shown as a percentage of the total views of that attribute. The right of the table represents the positive evaluation numbers, and the left represents the negatives.

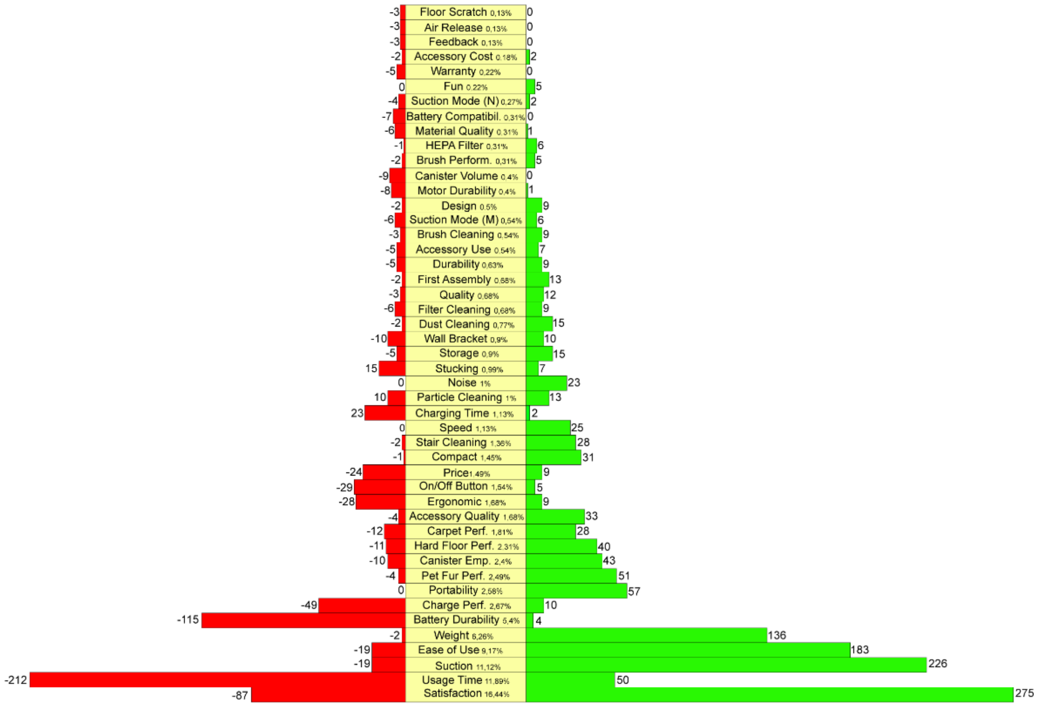
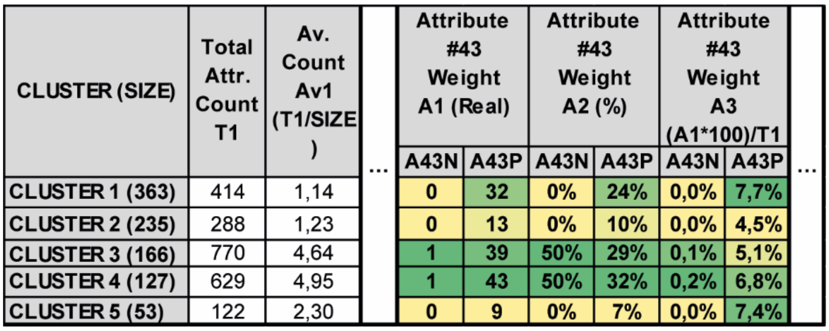


Figure 3. *Attribute scores and percentages according to 944 ID*

Figure 3 shows that the product attributes with the most comments by IDs over unweighted values are Satisfaction, Usage Time, Suction, Ease of Use, Weight, and Battery Durability, respectively. The lowest comments were placed on Floor Scratch, Air Release, Feedback, Accessory Cost, and Warranty. On the positive side, the most intense differentiation was Suction, Satisfaction, Ease of Use, Weight, and Portability, while on the negative side, Usage Time, Battery Durability, Charge Performance, On/Off Button, and Charging Time. When an overall evaluation is made in terms of all IDs, the Dyson V7 is a successful product with high suction performance, easy to use, lightweight and portable. However, it is a product that runs out of batteries quickly, has a low usage time, takes a long time to charge, the battery deteriorates quickly, and the need to constantly press the on-off trigger is a problem. When the product feature scores are examined, the Battery Compatibility, Floor Scratching, Product Feedback, Canister Volume, Air Release, and Warranty aspects have received entirely negative reviews. In contrast, the product's Noise, Portability, Speed , and Fun have positive reviews. Positive and negative comments are equal regarding the Maximum Suction Mode, Wall Bracket, and Accessory Cost. Some of the product features in the table have emerged with very few opinions. For example, only 3 of 944 users stated that the product scratches the floors.

8.CONTINGENCY TABLE AND CORRESPONDENCE ANALYSIS (CA)

A contingency table for Correspondence Analysis was prepared by calculating each cluster separately for each attribute. The sample snapshot of the table below shows the scores for Weight attribute #43 (Figure 4).

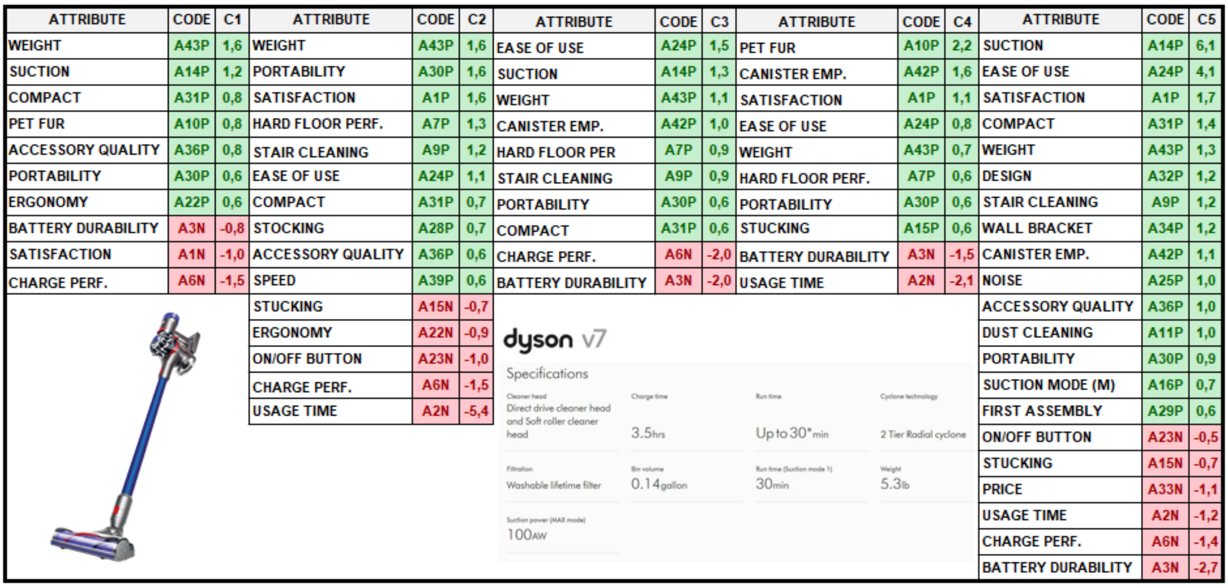


CLUSTER (SIZE)	Total Attr. Count T1	Av. Count Av1 (T1/SIZE)	...	Attribute #43 Weight A1 (Real)		Attribute #43 Weight A2 (%)		Attribute #43 Weight A3 (A1*100)/T1		...
				A43N	A43P	A43N	A43P	A43N	A43P	
CLUSTER 1 (363)	414	1,14		0	32	0%	24%	0,0%	7,7%	
CLUSTER 2 (235)	288	1,23		0	13	0%	10%	0,0%	4,5%	
CLUSTER 3 (166)	770	4,64		1	39	50%	29%	0,1%	5,1%	
CLUSTER 4 (127)	629	4,95		1	43	50%	32%	0,2%	6,8%	
CLUSTER 5 (53)	122	2,30		0	9	0%	7%	0,0%	7,4%	

Figure 4. *Snapshot of sample weighted conversion table Att.#43*

Weighted A3 scores were used for CA. The A3 score was obtained by multiplying the actual scores by the coefficient. The coefficient's logic is related to the recall number of attributes (A1) to the all-mentioned attribute amount of the cluster (T1). The coefficient's effect can be observed in the variation between A2 and A3. The determinant of an attribute is not its number but its frequency in the interpretation of the entire cluster. As in column A3, C1 and C5, which were 65+, are more sensitive to the Weight attribute.

When the preliminary evaluation output is examined (Figure 5), positive and negative opinions of each cluster for Dyson 7 can be evaluated. Regarding the negative values, C1 had the lowest overall satisfaction. Dissatisfaction depends on the product's battery durability and charging performance. Complaints about the product's battery durability and charging performance were present in all clusters. This situation also negatively affects the duration of use. C2 and C5 clusters reported negative feedback for the On/Off Button, where the C2 cluster associated button usage with ergonomics (for Dyson V7 to work, the On/Off button must be constantly pressed). There is variation among clusters in positive views. C1, C3, and C5 clusters highlight the suction power of the product. Ease of use and weight are common positive comments. Pet owner clusters evaluated the pet fur suction performance highly. The light and compact structure of the product positively affects its portability, use on stairs, and storage.



ATTRIBUTE	CODE	C1	ATTRIBUTE	CODE	C2	ATTRIBUTE	CODE	C3	ATTRIBUTE	CODE	C4	ATTRIBUTE	CODE	C5
WEIGHT	A43P	1,6	WEIGHT	A43P	1,6	EASE OF USE	A24P	1,5	PET FUR	A10P	2,2	SUCTION	A14P	6,1
SUCTION	A14P	1,2	PORTABILITY	A30P	1,6	SUCTION	A14P	1,3	CANISTER EMP.	A42P	1,6	EASE OF USE	A24P	4,1
COMPACT	A31P	0,8	SATISFACTION	A1P	1,6	WEIGHT	A43P	1,1	SATISFACTION	A1P	1,1	SATISFACTION	A1P	1,7
PET FUR	A10P	0,8	HARD FLOOR PERF.	A7P	1,3	CANISTER EMP.	A42P	1,0	EASE OF USE	A24P	0,8	COMPACT	A31P	1,4
ACCESSORY QUALITY	A36P	0,8	STAIR CLEANING	A9P	1,2	HARD FLOOR PER	A7P	0,9	WEIGHT	A43P	0,7	WEIGHT	A43P	1,3
PORTABILITY	A30P	0,6	EASE OF USE	A24P	1,1	STAIR CLEANING	A9P	0,9	HARD FLOOR PERF.	A7P	0,6	DESIGN	A32P	1,2
ERGONOMY	A22P	0,6	COMPACT	A31P	0,7	PORTABILITY	A30P	0,6	PORTABILITY	A30P	0,6	STAIR CLEANING	A9P	1,2
BATTERY DURABILITY	A3N	-0,8	STOCKING	A28P	0,7	COMPACT	A31P	0,6	STUCKING	A15P	0,6	WALL BRACKET	A34P	1,2
SATISFACTION	A1N	-1,0	ACCESSORY QUALITY	A36P	0,6	CHARGE PERF.	A6N	-2,0	BATTERY DURABILITY	A3N	-1,5	CANISTER EMP.	A42P	1,1
CHARGE PERF.	A6N	-1,5	SPEED	A39P	0,6	BATTERY DURABILITY	A3N	-2,0	USAGE TIME	A2N	-2,1	NOISE	A25P	1,0
			STUCKING	A15N	-0,7							ACCESSORY QUALITY	A36P	1,0
			ERGONOMY	A22N	-0,9							DUST CLEANING	A11P	1,0
			ON/OFF BUTTON	A23N	-1,0							PORTABILITY	A30P	0,9
			CHARGE PERF.	A6N	-1,5							SUCTION MODE (M)	A16P	0,7
			USAGE TIME	A2N	-5,4							FIRST ASSEMBLY	A29P	0,6
												ON/OFF BUTTON	A23N	-0,5
												STUCKING	A15N	-0,7
												PRICE	A33N	-1,1
												USAGE TIME	A2N	-1,2
												CHARGE PERF.	A6N	-1,4
												BATTERY DURABILITY	A3N	-2,7

Cleaner head	Charge time	Run time	Cyclone technology
Direct drive cleaner head and Soft roller cleaner head	3.5hrs	Up to 30*min	2 Tier Radial cyclone
Filtration	Bin volume	Run time (Suction mode 1)	Weight
Washable lifetime filter	0.14gallon	30min	5.3lb
Suction power (MAX mode)			
100AW			

Figure 5. *Cluster-based positive and negative opinions*

The hard surface suction performance of the product was found to be high. C3, C4, and C5 clusters drew attention to the ease of emptying the canister. Overall, the battery and charging performance and constantly pressing the On/Off button are problematic. However, there is also a clogging problem. C5 cluster thinks the product is expensive. However, the product is useful and light, does its job "when it works" and has high absorption of fur and hard floors.

For the CA, positive and negative A3 values for each attribute were processed as numerical values in columns and clusters as rows. SPSS program was used for Multiple CA (3D), and cluster relationships with attributes were tried to be explained. Multiple CA is an advanced analysis that helps understand and visualize the relationships between different categorical variables where clusters give row and attribute column values. Each cluster and attribute (> 0.5) are located in a 3D space, as in the graph presented below, and the total cumulative Proportion of Inertia value is calculated as 0.889/1.

Attributes with a score value <0.5 were not evaluated due to their percentage level. In the bi-plot (Figure 6), C3 is in the center as it shares similar attributes with others and does not highlight any different attributes. The two closest clusters over the inertia in the X dimension are C2 and C5. The product stair cleaning performance combines these two clusters with overall satisfaction. However, the On/Off Button and clogging problem brought them closer. C2 differs from C5 in terms of battery performance and low usage time.

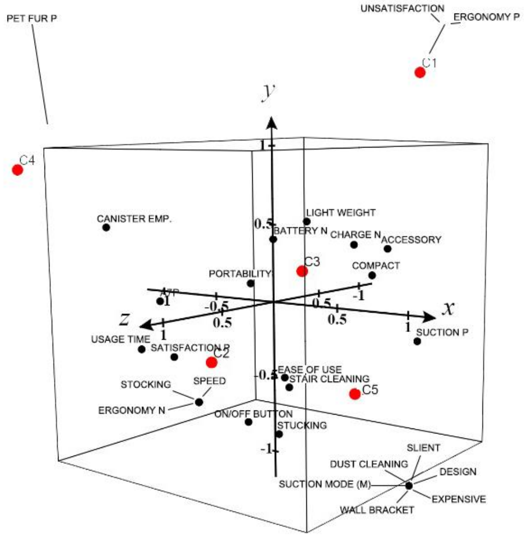


Figure 6. *3D Bi-plot of clusters vs. attributes*

C5, on the other hand, mainly focuses on design, silence, expensiveness, dust extraction, and wall bracket attributes. The most differentiated cluster in the Y dimension was C1. C1 was not satisfied with the product. For the C1, the product is light and has good suction. Nevertheless, battery and charging problems cannot produce the expected performance. In the Z dimension, C1 and C4 clusters converge. The essential attributes of the C4 are the pet hair suction, and the canister, which is immediately filled due to good suction, can be quickly emptied.

9.GENERAL DISCUSSION

After a series of data mining and statistical analyses, it was seen that OPRs allow obtaining demographic data due to the clues they contain. These clues are static demographic and dynamic personal context data that help understand IDs' needs, experiences, behavior, and goals. Some OPRs may not contain personal information due to the secretive attitude of IDs. However, it was observed that 33% of the total crawled (n=2844) IDs provided the targeted data with the OPRs they produced. It is thought that this ratio can be increased by expanding the ID pool and further mining analysis. The study's static and dynamic data can help designers empathize by understanding their users' data and personal contexts; in a way, it can produce a basis for developing a persona. Designers must internalize fundamental data analysis and statistical methods to create and effectively use this automated and accessible understanding.

944 IDs are real people, so their publicized data are not fictitious guesses, and the actual data they provide reflect fundamental user patterns. Each ID makes statements about the product, their experiences, and how they interact with it. IDs are context-specific; that is, they focus on the behavior and goals related to the specific area of the product with their product/attribute-based comments. The more interesting thing is; that OPRs eliminate the primary research constraints of time and money. It allows passive interviewing and observation with real people and makes it possible to create realistic (Norman, 2018) inferences.

It is possible to represent unique IDs with sufficient data within a specific cluster. Thus, not only individual users but also more controllable, sometimes surprising, homogeneous groups can be created. In the V7 promotional ad (https://www.youtube.com/watch?v=iyikrxxuotw) published by Dyson on May 19, 2000, the pet owner residing in the single-family house is a 3-person family is depicted. The daughter's age seems to be in the range of 10–15, and the parents' age is in the field of 34–44. This description fits perfectly into C1, which has the most members (38.5%) and is defined in Analysis 2. The ad shows children and animals as the primary household pollutants. This coincides with the high importance of child and animal ownership in demographic data-based cluster analysis. However, cluster analysis results also yielded surprising results. Similarly, families with children and pets living in a single-family house in the 34–44 range were used by Dyson (V8, V10, and V15) in the product-release advertisements that followed. However, a mass of 61.5% in cluster analysis does not fully comply with this definition. It has been determined that although the 24–34 age group, single women, and 65+ age group women and men constitute a large mass, they are ignored by the company. In evaluating these groups, particular attention should be paid to cluster-specific complaints because the On/Off button, the product's price, and the clogging problem are not defined by the C1.

Another valuable result came from the comparison of static and dynamic user data. Because of the comparison, clusters based on demographic data shift to the center and become homogeneous after the product experience. The main reason for this experience shift is as follows: Clusters based on static demographic data have 13 criteria that vary ordinally, such as age group, during the formation phase. As a result, each criterion affects the position of the cluster in 3D space, and diversity causes clusters to be relatively homogeneous within themselves but heterogeneous in terms of distribution by sharp lines. However, looking at the comments for the product, there seems to be a severe concentration in some product attributes. In other words, not all product attributes are equally decisive. Figures 5 and 7 show that when common attribute comments increase, these attributes become central and begin gathering clusters around themselves.

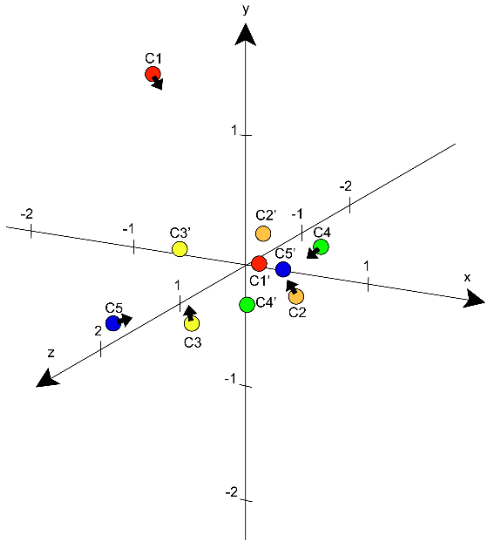


Figure 7. *The Experience Shift*

Thus, for example, the Weight attribute produced data by each cluster (Figure 4) and pulled the clusters toward itself as a central attribute. In contrast, the Air-Release attribute may not yet only three IDs complained that the product blew air to the face during air evacuation. This attribute effect was weak in determining the position of the clusters and could not even be evaluated.

What does it mean for clusters to become increasingly homogenous after a production experience? Although there are very different heterogeneous IDs, homogeneity increases when a single and fixed product and its properties with consensus are the main determinants. This effect detected in this study is defined as Experience Shift. In our example, when we look at the Weight attribute, only two of 138 scores indicate that it is heavy (ID# 155: It is not heavy, but all of its weight is right on your wrist; ID# 937: it was heavy for my arthritic). This product weighs 8lb, agreed as it is light. However, the first criticism (ID#155) is about the usual composition of the product. The other (ID# 937) is due to the unique physiological problem of ID. Thanks to its positive and negative features, the product transforms market divisions after experience and creates similar needs and desires. The main thing is to focus on the clusters that become homogenized after the product experience, to find the points that make the clusters different, and base the design decisions on these.

When Figure 7 is examined, the clusters showing relatively homogenizing tendencies are the C1 and C5. When the C1 was concerned, it was in a particular position since it was the most significant cluster because of the intense ownership of children and animals. Still, it was thoughtfully differentiated in terms of general dissatisfaction in attributes. So why has this cluster slumped toward the center? This was because the issues that created dissatisfaction were not determined and had already been mentioned by other groups. C5 was the smallest cluster and was almost demographically extreme. This cluster also shifted toward the center because of common pros and cons. The relatively small number of attributes specific to this cluster did not prevent this cluster from being dragged to the center due to weighted attributes.

The goal of each firm is to divide the heterogeneous market into inherently homogeneous groups to reduce costs and make focal product decisions so that firms can focus entirely on a set of customer needs and plan their marketing mix accordingly. Homogeneity should not be demographic unity and should be interpreted with experience and product context. Within the scope of this study, it is revealed that clusters based on purely demographic static data are insufficient in design decisions; yet, static data are just recipes; they are a measurable characteristic of a given population (Ramachandran, 2017) and are a valuable starting point for further interpretation. Decisions based on static demographic data can often bias our design decision-making processes. Seeing products from the perspective of real people is only possible by evaluating the needs and requirements in context. For this reason, and to avoid stereotypes, in the study, real persons were first converted to IDs. However, after the targeted data were obtained, the IDs were expressed as clusters, and then the clusters were simplified and reduced to unique discourses to support design decisions. Again, to answer some of the crucial questions, designers should genuinely understand what is hidden in the discourses of the users. Essential data mining and statistical analysis methods make it possible to identify and recognize users through OPRs.

10.IMPLICATIONS FOR PRODUCT DESIGN PRACTICE AND CONCLUSION

OPRs are voluntary production and are available in large numbers, publicly available, and accessible. These features increase the number of samples and save money and time for designers to understand the user. The number of OPRs is simple CSV files that do not require high processing capability. In contrast, the unbiasedness of the sample and its large amount prevent standard errors. Insight from large volumes of data allows designers to understand the context and learn and develop safely and independently.

The analysis and methods used in OPR inferences are not very complex. There is a massive amount of free software, APIs, and data processing platforms to be used in this regard that designers must internalize and put data analytics on their agenda. Data analytics helps facilitate information visibility and process automation in design and transforms data into consumable information assets.

With the information in OPRs, designers can make many inferences, such as user profiles from the desk. OPRs profiles can be represented as clusters, the best representation of each cluster can be determined, the attributes that each ID or cluster considers essential about the product can be extracted, development areas for product attributes can be identified, and new clusters can be created at the end of the product experience. All the inferences obtained can be developed and enriched with more advanced processes for persona generation.

User representations produced by designers are likely to create stereotypes (Turner and Turner, 2011). The representation should be supported with real data to prevent this situation caused by incomplete information and intense homogenization efforts. It is possible to avoid clichés because of working with real people. Within the scope of this study, five different user clusters were identified. Of these clusters, only C1 (38.5%) corresponds to the user group highlighted by the company. However, it has been possible to prevent the designers' remaining users (61.5%) from being neglected.

Imagining the user as a real user strengthens empathy (Pruitt and Adlin, 2006). If the data are based on real users, the representations produced will also be real or realistic. The critical point here is that the inferences in this study are not based solely on demographic data. Because the views of each ID regarding the product remove the IDs from the recognition of the target group of traditional marketing and directly associate them with the product, from this perspective, designers will more easily imagine similar experiences by focusing on the real experiences of real people.

Actual and current product experiences are embedded in the context-based experiences of end users. This study reflects the end user's perspective, including validated purchase experiences and user stories. Inferences based on demographic data are static and inadequate without product experience. It can be used to make visible post-product trends of clusters, revealing hidden patterns and trends. These insights can help designers discover new customers who fit the same molds as existing customers and offer competitive products that are not yet widely available.

Finally, according to the 944 ID OPRs, the Dyson V7 is a successful product that wins the consumer's appreciation, as long as it overcomes the battery usage and charge problem. Regarding form design decisions, the most severe problem to be solved seems to be the On/Off button, which must be pressed all the time to save the battery. This situation causes the product to be subject to low evaluation in terms of ergonomics. Another design problem relates to the flow geometry. A narrow, sharp-angled geometry leads to clogging, especially with dense hair. Design decisions on this issue also need to be reconsidered.

Acknowledgement

The author wishes to thank to Gazi University Design Application and Research Center for their technical support.

REFERENCES

[1] Almaliki, M., Ncube, C., Ali, R. (2015). Adaptive Software-Based Feedback Acquisition: A Persona-Based Design. In *2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS),* 100-111. Doi:10.1109/rcis.2015.7128868

[2] Boush, D., Kahle, L. (2001). Evaluating Negative Information in Online Consumer Discussions: From Qualitative Analysis to Signal Detection. *Journal of Euro-marketing* 11(2): 89-105.

[3] Cooper, R. (2011). Users, Users, Users and the Use of Design. *The Design Journal* 14(4): 387-389.

[4] Güneş, S. (2020). Extracting Online Product Review Patterns and Causes: A New Aspect/Cause Based Heuristic for Designers. *The Design Journal* 23(2): 375-393. Doi:doi.org/10.1080/14606925.2020.1746611

[5] Margolin, V. (1997). Getting to Know the User. *Design Studies* 18(3): 227-236. Doi:10.1016/S0142-694X(97)00001-X

[6] Miaskiewicz, T., Kozarb, K. (2011). Personas and User-Centered Design: How Can Personas Benefit Product Design Processes? *Design Studies* 32(5): 417-430. Doi:https://doi.org/10.1016/j.destud.2011.03.003

[7] Norman, D. (2018). Ad-Hoc Personas and Empathetic Focus. https://jnd.org/ad-hoc_personas_empathetic_focus/ . Last Accessed: 20.08.2022

[8] Oygür, I. (2018). The Machineries of User Knowledge Production. *Design Studies* (54): 23-49. Doi:https://doi.org/10.1016/j.destud.2017.10.002

[9] Park, D., Lee, J., Han, I. (2007). The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement. *International Journal of Electronic Commerce* 11(4): 125-148. Doi:10.2753/JEC1086-4415110405

[10] Pruitt, J., Adlin, T. (2006). *Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Burlington: Morgan Kaufmann.

[11] Ramachandran, G. (2017). Impact of Demographic Variables on the Use Patterns of Electronic Information Resources Among Aerospace Scientists and Engineers of Bangalore. *International Journal on Environmental Sciences* 8(1): 90–104.

[12] Rogers, E. (2003). *Diffusion of Innovations.* New York: Free Press.

[13] Turner, P., Turner, S. (2011). Is Stereotyping Inevitable When Designing with Personas? *Design Studies* 32(1): 30-44. Doi:https://doi.org/10.1016/j.destud.2010.06.002

[14] Zhan, J., Loh, H., Liu, Y. (2009). Gather Customer Concerns from Online Product Reviews – A Text Summarization Approach. *Expert Systems with Applications* 36(2): 2107–2115. Doi:10.1016/j.eswa.2007.12.039