Application of natural language processing to enhance qualitative research used for marketing

Poj Netsiri[1] – Marketa Lhotáková[2]
ORCID iD: 0000-0002-3309-2779[1]
netp03@vse.cz, marketa.lhotakova@vse.cz
Prague University of Economics and Business, Faculty of International Relations, Department of International Business, Prague, Czech Republic

DOI 10.18267/pr.2023.kre.2490.13

Abstract: Understanding consumer behavior can help improve marketing of the products. Market research normally applies conventional qualitative analysis to discover the reasons why consumers act and purchase products in a certain way. However, qualitative analysis with small samples is insufficient to make population-level summaries. On the other hand, qualitative analysis with large samples is time consuming. In addition, poor quality of qualitative research due to human error and bias from the researcher can lead to misleading findings. Therefore, to overcome these problems, Natural Language Processing is applied to extract consumer behaviors from large-scale samples of product reviews. The result from 809 product reviews (source: CarMax in US) of preowned luxury cars (Mercedes-Benz) indicates top 10 relevant keywords "ride", "smooth", "luxury", "nice", "feature", "excellent", "beautiful", "comfort", "style", and "expect". These terms correlate with consumer perceived emotional, social, and quality values that could positively influence customer purchase intention toward preowned luxury cars.

Keywords: consumer behaviour, consumer perceived value, qualitative analysis, natural language processing, topic modelling

JEL Classification codes: C55

INTRODUCTION

Understanding consumer behavior is one of the most important tasks for marketing. Knowledge derived from consumer behavior helps marketers understand what customers want and need and enables them to appropriately offer products and services that match their target audience (Radu, 2022). It also helps marketers to market and position their products or services successfully (Hampasagar, 2021). In recent years, consumer purchase intention toward luxury brands is one of the most prominent areas in marketing research. Many researchers have investigated components that can influence consumer purchase intention toward luxury products. It was found that attitude, past purchase experience, perceived behavioral control, perceived emotional value, perceived social value, perceived quality value, and perceived green value can influence the intention to purchase toward second-hand luxury products (Lou et.al., 2022) (Stolz, 2022).

To understand why a consumer has acted and purchased in a certain way, the marketers frequently use qualitative research to analyze and interpret their data. The typical data used in qualitative research is non-numerical, contextualized, and unstructured such as open-ended questionnaires. Conventional qualitative research is criticized as an overly subjective research method partially due to the use of unstructured or semi structured data (Abram et al., 2020). This method is time and cost consuming, especially for large-scale data. Poor quality of

qualitative research can lead to misleading findings (Abram et al., 2020). Furthermore, qualitative research with small samples is insufficient to make population-level summaries. Other than that, qualitative research has many limitations including potential bias in answers, self-selection bias, and potentially poor question from researcher (Abram et al., 2020).

Natural language processing (NLP) is a branch of Artificial Intelligence (AI) concerned with giving computers the ability to understand human language. NLP utilizes computational linguistics and statistical analysis and machine learning. These technologies enable computers to process human language in the form of textual data and to understand its meaning, intent, and sentiment. Recently, NLP has been used for qualitative research with many types of data sources such as open-ended feedback from a customer satisfaction survey, notes in an electronic medical record (EMR) (Abram, 2018, 2020). These studies suggest that NLP could capture the over-all thematic descriptions and analyze content from large-scale unstructured data. NLP could save time and cost of analytical processes as well.

Therefore, NLP can be considered a potential AI technology that can be applied to marketing research (Gkikas, 2022) (Jarek et.al, 2019). As a pilot study of application of NLP to marketing research, eight hundred and nine product reviews of preowned luxury cars (Mercedes Benz) were downloaded from a popular used car dealer website (carmax.com) in U.S. Then, NLP was applied to enhance qualitative analysis and extract terms related to consumer behaviors from this large-scale unstructured data. Finally, the consumer perceived values toward luxury cars derived from this method were examined and compared to the previous findings from other groups.

1. LITERATURE REVIEW

1.1 Consumer Perceived Value (CPV)

Recently, consumer perceived value has become a popular topic in marketing (Lou et.al., 2022). Consumer perceived value refers to consumers' overall assessment of the utility of a product based on their perceptions. Perceived Value Scale (PERVAL) was developed to assess customers' perceptions of the value of a consumer durable good at a brand level (Sweeney et. al., 2001). PERVAL consists of four value dimensions: emotional value, social value, economic value, and quality value. These perceived values have been recognized as crucial dimensions for luxury consumption. Accordingly, this study adopted these four value dimensions and formulated corresponding hypotheses as follows.

1.2 Economic Value

Financial benefit has been widely cited as a critical driver of purchasing behaviors. The lower price of preowned cars is a frequently mentioned reason why consumers buy preowned cars rather than new ones. New luxury products are considered needlessly expensive. Therefore, economic aspects may influence the intention to purchase preowned luxury products. In general, pre-owned luxury cars have a financially lower price than brand-new luxury cars. The affordability of preowned luxury provides consumers with the opportunity to save their money and obtain financial benefits. Thus, the following hypotheses were proposed:

Hypothesis 1 (H1). *Economic value positively affects consumers' purchase intentions toward luxury cars.*

1.3 Emotional Value

Emotional value refers to emotional perceptions consumers may have while using or shopping for products such as love, empathy, pride, happiness, and nostalgic pleasure. Attitudes toward luxury itself and luxury brands may influence the purchase intentions of luxury cars. Therefore, the following hypothesis was proposed:

Hypothesis 2 (H2). *Emotional value positively affects consumers' purchase intentions toward luxury cars.*

1.4 Social Value

Social value refers to a product's ability to gain favorable evaluation from people in society. It incorporates several aspects such as social image, identification, and status. Consumers with a high need for status tend to spend their money conspicuously on luxury products to display their wealth and purchasing power. In addition, as luxury brands frequently enclose prestigious values, the possession of luxury products serves as a symbolic sign of group membership and as a means of improving individuals' social standing. Thus, the following hypothesis was proposed:

Hypothesis 3 (H3). *Social value positively affects consumers' purchase intentions toward luxury cars.*

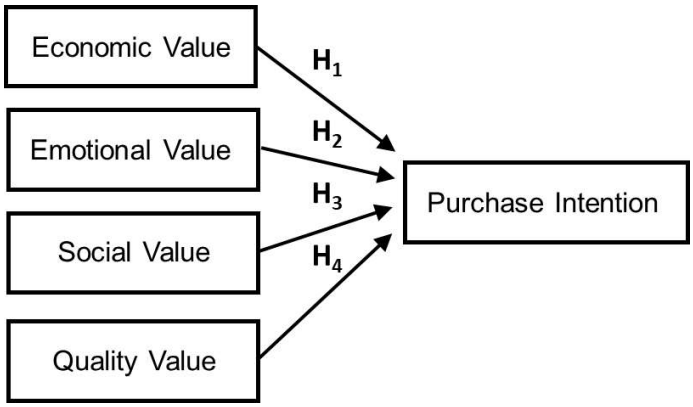
1.5 Quality Value

Perceived quality is defined as the consumer's judgment about the overall excellence of a product. As luxury products are considered to have a top quality and perpetual design, this leads to a perception of a long-life cycle and not being affected by short-term fashion trends. The exceptional quality of luxury products is not only related to the components and materials used but also to the high level of skills and processes involved in the craftmanship. High-quality consciousness is identified as one of the shopping style dimensions in the context of preowned luxury. Therefore, the following hypothesis was proposed:

Hypothesis 4 (H4). *Quality value positively affects consumers' purchase intentions toward luxury cars.*

Based on the literature review above, the following research model was developed. As shown in Fig. 1, consumers' perceived economic value, perceived emotional value, perceived social value, and perceived quality value were proposed to influence their purchase intentions toward preowned luxury cars.

Fig. 1 Hypotheses of this study

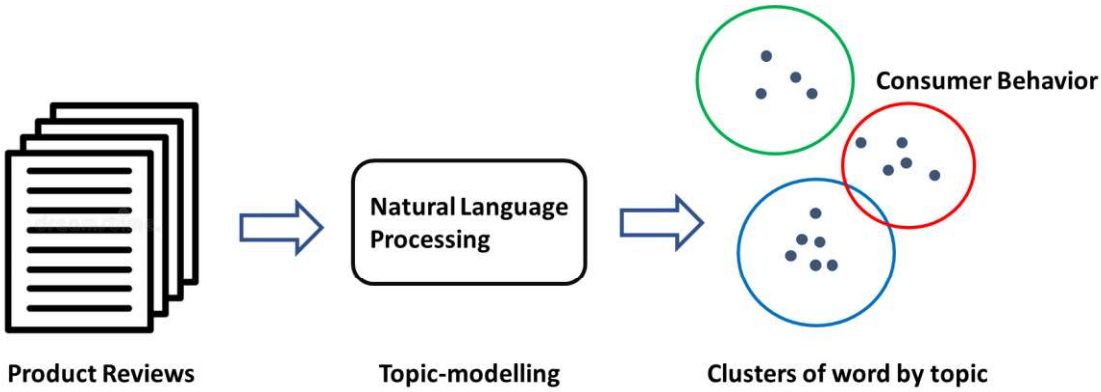


Source: Author, 2022

2. METHODOLOGY

Fig. 2 shows the basic concept of this study. Firstly, product reviews associated with luxury cars are collected from a website of a preowned luxury car dealer. Secondly, these data are processed using topic modelling techniques in Natural language processing. Finally, the topics associated with consumer behavior are identified and their clusters of words are analyzed.

Fig. 2 Basic concept of this study

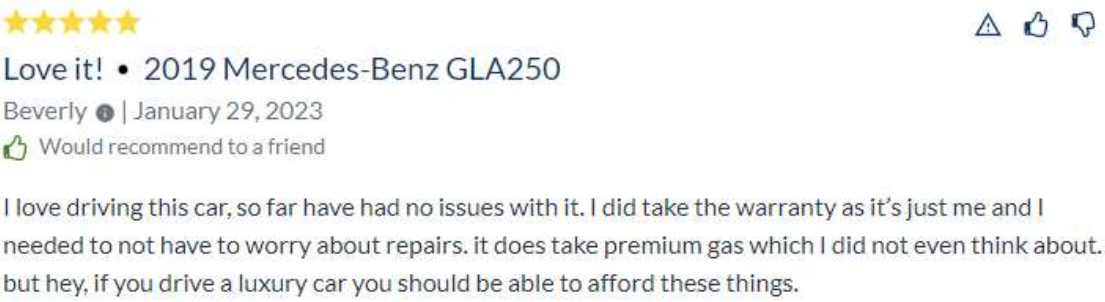


Source: Author, 2022

2.1 Data collection

The raw data set was collected from 809 product reviews of a luxury car (Mercedes-Benz) from a website of Carmax, Inc. in 2022 (https://www.carmax.com/reviews/mercedes-benz). CarMax Inc. is the largest retailer of preowned vehicles in the U.S. Their website provides large-scale product reviews of cars grouped by brand to the public. The number of product reviews collected from this website is sufficient to make a population-level summary for this study. Furthermore, this data set is appropriate for this study because the preowned luxury car consumption could represent a segment in the luxury car market.

Fig.3 Example of raw data



Source: www.carmax.com/reviews/mercedes-benz, 2022

2.2 Topic Modelling

Topic modelling is used to extract useful information related to customer behavior from product reviews data. Topic modelling automatically counts and group similar word patterns from unstructured data to topics. For example, normally, if a marketer wants to know what his customers are saying about particular features of their products, he will have to manually find texts that are related to his topics of interest from product reviews and manually analyze them. However, by detecting patterns such as word frequency and distance between words, a topic modelling can automatically cluster parts of product reviews that are similar, and words and expressions that appear most frequently. With this technology, topic modelling can save processing time, especially with large-scale data (Textrics, 2023).

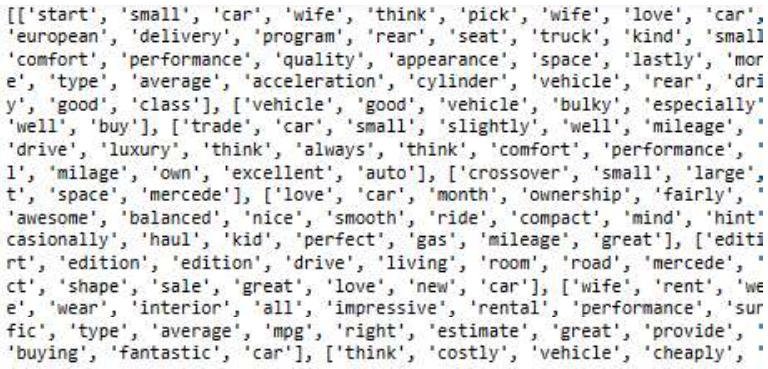
2.3 Latent Dirichlet Allocation

The algorithm used to model topics in this study is Latent Dirichlet Allocation (LDA) (Blei et. al., 2003) (Ma, 2019). The purpose of LDA is to group relevant documents in the corpus into a set of topics which covers the majority of the words in those documents (MonkeyLearn, 2023). LDA identifies latent topics within a corpus by estimating the probability that each document is generated by any specific topic and the probability that any word is generated by a specific topic. Once the LDA model is optimized, researchers can examine the words that are most probabilistically related to each topic to derive topic meaning and understanding of their large-scale textual data (Hagg et. al., 2022).

2.4 Pre-processing and processing

The research methods were conducted in two stages: (1) preprocessing where the data were prepared for analysis using standardization and formatting and (2) processing where the data is actually analyzed. The original data was collected in Excel format and then converted to JSON format. The primary packages and modules used in this study includes: numpy, gensim, nltk, pyLDAvis, Jupyter Notebook and Python. In the first stage, the standard data preprocessing methods for NLP were employed and included: text case standardized to lower, removal of stop words (e.g., "and," "or", "but," "the") punctuation removal and word stemming (removal of word endings allowing a focus on the word root). Excluding words which do not contribute to the identification of relevant themes is a standard process in many approaches to qualitative research. Fig. 4 shows an example of data after preprocessing.

Fig.4 Example of data after preprocessing

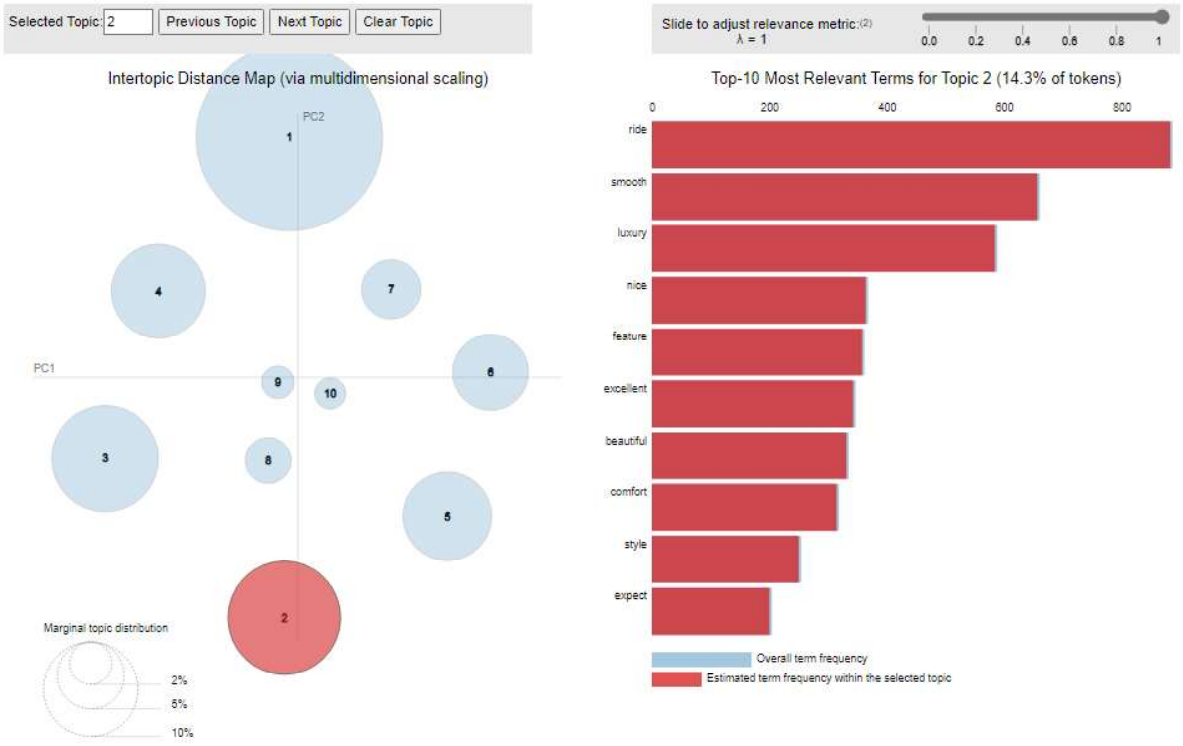


Source: Author, 2022

In the data processing stage, Latent Dirichlet Allocation (LDA) from Genism library was used for topic modeling. To conduct the qualitative assessment, the LDA output was used to develop topic themes and an overall theme for the data. The topic that was the most relevant to consumer behavior was selected for further analysis and interpretation. In addition, LDAvis is used to visualize topics estimated by LDA (Sievert et al. 2014). This visualization tool provides a global view of the topics while at the same time allowing for a deep inspection of the terms associated with each individual topic.

3. RESULTS AND DISCUSSION

Fig. 5 Result from LDA topic modelling



Source: Author, 2022

3.1 Results

Fig. 5 consists of two basic components. Firstly, a global view of the topic model is displayed on the left panel where the bubbles represent the topics. Positions of centers are determined by computing the distance between topics. The size of each circle represents each topic's prevalence, and the topics are sorted in decreasing order of prevalence (Sievert et. al., 2014). Secondly, the right panel of this visualization depicts a horizontal bar chart whose bars represent the individual terms that are the most useful for interpreting the currently selected topic on the left. These terms can indicate the meaning of each topic (Sievert et. al., 2014).

The topics can be viewed as underlying constructs measured by the combination term frequency and distance between terms. By inspecting each of the LDA topics, the pattern can be identified. Since the topic of this study is related to consumer perceived value, topic 2 was selected. This topic contains of terms: "ride" (verb), "smooth" (adverb), "luxury" (adjective), "nice" (adjective), "feature" (noun), "excellent" (adjective), "beautiful" (adjective), "comfort"

(verb), "style" (noun), and "expect" (verb) as shown in Fig. 5 - all of which point to the underlying concept of consumer behavior. In addition, terms of "ride", "smooth" and "luxury" are dominant (top 3 terms) according to the estimated term frequency within the selected topics.

3.2 Discussion

Table. 1 Summary of top 10 keywords and their corresponding category (H_1 – H_4)

Category	Keyword and Rank
Economic value (H_1)	
Emotional value (H_2)	nice (#4), beautiful (#7), comfort (#8), style (#9)
Social value (H_3)	luxury (#3), excellent (#6),
Quality value (H_4)	ride (#1), smooth (#2), feature (#5), excellent (#6), beautiful (#7), expect (#10)

Source: Author, 2022

The current study investigated top 10 keywords extracted using NLP and answered a question of whether customers' perceived values including economic, emotional, social, and quality values can positively influence their purchase intentions. Based on a complex purchasing behavior of expensive product, we could also presume that the customers who created these product reviews were highly involved in the decision making process and considering factors such as cost, return on investment, durability, usability, feedback, recommendations while comparing among different options before finally making a purchase. Table 1 summarizes keywords and rank by their corresponding category (hypothesis) as follows.

3.2.1 Economic value (H1)

Surprisingly, this topic has no keyword related to economic value. Thus, perceived economic value was found to have no positive influence on consumers' purchase intentions toward luxury cars. This indicates that the luxury car market is well-established. The price is no longer the primary reason for consumers to buy luxury cars. This finding is also in agreement with a previous study from a different group (Lou et.al., 2022).

3.2.2 Emotional value (H2)

According to the number and rank of keywords in this topic, emotional value has the 3rd strongest influence on purchase intention. Therefore, perceived emotional value can positively influence consumers' intentions to purchase preowned luxury cars. This result is consistent with a previous study (Lou et.al., 2022) suggesting that secondhand luxury shopping provides consumers with joy. Feeling joy and pleasure is an important stimulus for an individual to perform a certain behavior. The terms "nice", "beautiful", "comfort", "style" could express consumers' experience of using their luxury cars.

3.2.3 Social value (H3)

According to the number and rank of keywords in this topic, social value has the 2nd strongest influence on purchase intention. Thus, consumers may strongly believe that ownership of

luxury cars can show their social status, which supports the prior studies (Lou et.al., 2022). This finding reinforces the understanding that individuals are concerned about their social identities and tend to choose luxury products to create and maintain a desirable self-image that they can display to the public. The terms "luxury" and "excellent" could express consumers' social status of owning their luxury cars.

3.2.4 Quality value (H4)

According to the number and rank of keywords in this topic, quality value has the 1st strongest influence on purchase intention. The symbolic meanings associated with luxury products includes exclusiveness, high quality, aesthetics, prestige, and craftsmanship. This finding correlates with previous studies (Lou et.al., 2022). The terms "ride", "smooth", "feature", "excellent", "beautiful" and "expect" could express quality value of the luxury cars.

We concluded that these keywords have positive influences on the consumer's purchase intention. Moreover, attributes associated with each keyword could provide further insight into its corresponding perceived value. For example, "smooth" could provide a specific feeling of quality value. According to our findings, the quality value is the 1st strongest influencer on purchase intention toward luxury cars. Thus, providing the best quality product is the most important in marketing them. After the quality value, development of a positive social status associated with luxury cars is the 2nd most important. Finally, development of positive emotion associated with luxury cars is the 3rd most important. There are numerous applications emerging from this finding. For example, marketers could use these keywords as a guideline to create their advertisements. Designers could design and optimize their products based on these keywords and their ranks.

3.2.5 Advantages and limitations

The common sources of large-scale data used in topic modelling research are social media such as forums, Twitter, Facebook, Instagram and databases such as scientific literatures, and formal documentations such as reports, clinical notes, health records, summary statements, letters of recommendation (Hagg et al. 2022). Past studies demonstrate that topic modelling provides researchers with unique flexibility in selecting the type of textual data that can best answer their research questions. Thus, the selection of textual data for analysis plays an influential role in analysis outcomes, as such it is imperative that researchers clearly specify their data inclusion and exclusion criteria to ensure reproducibility. For instance, researchers can utilize either original posts obtained from social media alone to discover a broad overview of topics within a forum/group, or original posts with their subsequent comments (Hagg et al. 2022).

By integrating an NLP into the same qualitative research analysis scheme used in a previous study, and then comparing the original findings using a conventional qualitative analysis (Abram 2018) to new findings using NLP (Abram et al. 2020), a consistent replication of the same results was demonstrated. There are many positive implications of integrating NLP approaches into qualitative research including cost and time savings. For this study, the primary topics were identified within 5 minutes of analysis. The findings from this study were in good agreement with previous studies at hypothesis level (Lou et.al., 2022). While previous studies could not provide insights into the perceived values because limitations of their methods, our method could provide relevant keywords associated with each perceived value.

Although LDA could model the topics from the raw data, it could not provide any descriptions that described the meaning of each topic. The researchers will have to identify the topic and describe the meaning based on their knowledge. Other than that, there is a limit to the number of topics LDA can generate. LDA assumes that words are exchangeable and sentence structure

is not modeled. Moreover, correlation among topics and temporal information are ignored in the topic modelling.

CONCLUSION

Consumers' desire to purchase preowned luxury cars is complex and multifaced. This pilot study applied NLP to extract consumer behavior related to purchasing intentions toward a preowned luxury car. The results showed that consumer's perceived emotional value, perceived social value and perceived quality value positively influence their intentions to purchase preowned luxury cars. Furthermore, major keywords found in this study could provide further insights into their corresponding perceived values. Thus, the ability to efficiently and cost effectively extract meaningful information using NLP represents a promising opportunity for marketing research. Future research should extend this method to extract different information useful for marketing and other areas.

REFERENCES

Abram, M.D. (2018). The role of the registered nurse working in substance use disorder treatment: a hermeneutic study. *Issues in Mental Health Nursing.* 39(6), 490-498. https://doi.org/10.1080/01612840.2017.1413462.

Abram, M.D., Mancini, K.T. and Parker, R.D. (2020). Method to integrate Natural language processing into quantitative research. *International Journal of Qualitative Methods*, 19. 1-6. https://doi.org/10.1177/1609406920984608.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 3, 993-1022. https://dl.acm.org/doi/10.5555/944919.944937.

Gkikas, D.C. and Theodoridis (2022). AI in Consumer Behavior. *Advances in Artificial Intelligence-based Technologies*. 147-176.

Hagg, L.J., Merkouris, S.S., O'Dea, G.A., Francis, L.M., Greenwood, C.J., Fuller-Tyszkiewicz, M., Westrupp E.M. and Macdonald, J.A., Youssef, G.J. (2022). Examining analytic practices in Laten Dirichlet Allocation withing psychological science: Scoping review. *Journal of Medical Internet Research.* 24(11). https://doi.org/10.2196/33166.

Hampasagar, A. (2021). Consumer behavior and its importance in marketing. Retrieve February 19, 2023 from: https://www.linkedin.com/pulse/consumer-behaviour-its-importance-marketing-anant-hampasagar/

Jarek, K. and Mazurek, G. (2019). Marketing and artificial intelligence. *Central European Business Review.* 8(2), 46-55. http://dx.doi.org/10.18267/j.cebr.213.

Lou, X., Chi, T., Janke, J. and Desh, G. (2022). How do perceived value and risk affect purchase intention toward preowned luxury goods? An empirical study of U.S. consumer. *Sustainability.* 14, 11730, 1-16. https://doi.org/10.3390/su141811730.

Ma, X. (2019). Dimensionality-Reduction with Laten Dirichlet Allocation. Retrieved February 19, 2023 from: https://towardsdatascience.com/dimensionality-reduction-with-latent-dirichlet-allocation-8d73c586738c

Monkeylearn (2023). Topic modelling: An introduction. Retrieved February 19, 2023 from: https://monkeylearn.com/blog/introduction-to-topic-modeling/

Radu, V. (2022). Consumer behavior in marketing-patterns, types, segmentation. Retrived February 19, 2023 from: https://www.omniconvert.com/blog/consumer-behavior-in-marketing-patterns-types-segmentation/

Sievert, C. and Shirley, K.E. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceeding of the workshop on interactive language learning, visualization, and interface.* 63-70. http://dx.doi.org/10.3115/v1/W14-3110.

Sweeney, J.C. and Soutar, G.N. (2001). Consumer perceived value: The development of a multiple item scale. *Journal of Retailing*. 77, 203–220. https://doi.org/10.1016/S0022-4359(01)00041-0.

Stolz, K. (2022). Why do(n't) we buy preowned luxury products? *Sustainability.* 14, 8686, 1-24. https://doi.org/10.3390/su14148656.

Textrics (2021). Topic modelling: How can you use it for analyzing unstructured data? Retrieved February 19, 2023 from: https://textrics.medium.com/topic-modelling-how-can-you-use-it-for-analysing-unstructured-data-28045dd502cc