Does the sum of the parts make up the Whole? Examining the Relationship Between Review Text Clusters and Product Clusters!!!

Madhuri Prabhala
*Indian Institute of Management, Calcutta*, madhurip16@iimcal.ac.in

Indranil Bose
*NEOMA Business School*, indranil.bose@neoma-bs.fr

Follow this and additional works at: [https://aisel.aisnet.org/pacis2022](https://aisel.aisnet.org/pacis2022)

**Recommended Citation**
[https://aisel.aisnet.org/pacis2022/135](https://aisel.aisnet.org/pacis2022/135)
Does the Sum of the Parts Make Up the Whole? Examining the Relationship Between Review Text Clusters and Product Clusters

Short Paper

Madhuri Prabhala
IIM Calcutta, Kolkata
Madhurip16@iimcal.ac.in

Indranil Bose
NEOMA Business School, France
indranil.bose@neoma-bs.fr

Abstract
All product information gathering today begins with the digital world, while product sales could be online or offline. The current study evaluates how online user-generated textual content influences the online search for product information and product sales in the Indian automobile sector. The proposed hypotheses are tested using 23,000 online reviews, Google search index, and offline car sales from the CRISIL database. K-means clustering is adopted to divide the automobile sector into product segments. Two topics are identified from the review text using Latent Dirichlet Allocation. The influence of the review topics is estimated at the overall sector and on identified product segments. The study finds that while textual UGC influences both search and sales, the influence differs for the different product clusters within the category. This study directs managers to suitably segment their markets for better capturing the online UGC influence.

Keywords: Online search, online user-generated content, Latent dirichlet allocation

Introduction
Any product purchase is preceded by a search for information related to the product (Beatty & Smith, 1987), which begins online in today’s digital age. The search for information determines the customer’s consideration set, and the information collected enables better purchase decisions (Wang et al., 2018). The online search for product information reflects the customers’ interest in the product. It precedes product sales in the purchase funnel. While online digital content has been identified to influence product sales, its influence on product interest or online information search is not well studied. The online information search differs depending on the nature of goods (Wang et al., 2018), the utility of the information source (Grant et al., 2007), and other customer and product characteristics (Grant et al., 2007). It is also well known that market segments exist within product categories (Punj & Stewart, 1983; Saunders, 1980), but how the search for information differs within these segments is less known. The factors influencing sales within product clusters are also known to be different. However, the influence of user-generated digital content across product clusters is lesser-known. Further, much of the research about the influence of online digital content is in the context of Western economies. Their generalizability to emerging Asian economies is yet to be well established (Karhade & Kathuria, 2020). Given this background, the following three research questions are proposed.

1. Does the online user-generated textual content influence the online search for product information in an emerging economy?
2. Does the online user-generated textual content influence the offline sales of the product in an emerging economy?
3. How does the influence of online user-generated textual content on the volume of online search and offline product sales differ across market segments within the product category?

As there is lesser research on the influence of digital content in emerging economies (Karhade & Kathuria, 2020), the study is set in the aspirational Indian automobile sector. Answers to the research questions proposed are evaluated by collecting car model sales data, the volume of online search data, and the online
reviews published in a popular web portal in India. The results show that the topics of discussion in the online review text influence both product sales and the volume of online product search. However, the topics that influence both are different, and the influence differs by product clusters within the overall automobile segment. Unsupervised learning techniques of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and the popular k-means clustering (Chau et al., 2006) are used to determine the optimum number of topics being discussed and the optimum product clusters seen in the automobile sector. Answering these research questions will enhance the existing literature on online search and the influence of digital content in an emerging Asian market (Karhade & Kathuria, 2020). The research will provide Indian decision-makers with context-specific directions on utilizing rich digital content proliferating.

The rest of the paper begins with discussing the relevant literature and the hypotheses proposed, followed by the data description, the variable operationalization, and the methodology. The empirical results are discussed next. The last sections summarize the findings and conclude.

**Literature and Hypotheses**

The relevant literature and the hypotheses developed to answer the research questions are discussed here.

**Review Content**

The decision-making goals of customers move from informational to transactional (Humphreys et al., 2021). During the informational phase, the consumer collects information that directs improved transactional decision-making or what to buy. The consumers' interest in a product lead to the volume of online search. Online user-generated content is an established source of information for customers. The information gained from text mining the online user-generated content contributes to enhanced recommendation systems (Bauman & Tuzhilin, 2021; Xu et al., 2018), identification of emotional distress (Chau et al., 2020), superior identification of fake reviews (Lau et al., 2011), quality enforcement in sharing economy platforms (Meng et al., 2020), prediction of aspect level sentiments (Xia et al., 2021), and information perception of popular products (Bao et al., 2021) apart from predicting sales (Wijnhoven & Plant, 2017) and purchase intention (Yan et al., 2015). Topic mining is one of the important text mining techniques. The topics extracted from online content predict purchase intention (Bulut, 2014), sales (Li et al., 2019), and review helpfulness (Namvar et al., 2021; Vallurupalli & Bose, 2020). While review rating and review volume lead to initial product perceptions, information related to product features is obtained by reading the review text. Suppose the customer is not satisfied by the information in the review text or the aspects being discussed, it would cause the customer to search further. Further, in an emerging Indian economy, the influence of the review text cannot be generalized from the academic research in Western economies (Karhade & Kathuria, 2020). Thus, the below hypotheses are proposed.

H1a: The discussion topics in the online review text influence the volume of online searches for the car models.

H1b: The topics of discussion in the online review text influence the car models’ offline sales.

**Product Clusters**

While understanding the influence of user content at an overall level is essential, determining product clusters within the overall segment and studying the influence within the clusters will enhance the insights gained. Products, namely car models in this study, that are market leaders have good brand recognition. The search behavior of customers intending to buy cars with good brand recognition will differ from where there’s less brand recognition. Similarly, clusters based on price points will also exhibit different behavior. Prior research has identified that the influence of sentiment in user-generated content differs based on the product groups within the overall segment (Wijnhoven & Plant, 2017). Extending this, the importance of topics in online textual reviews would differ according to product groups. Hence the below hypotheses are proposed.

H2a: The influence of discussion topics in online review text on the volume of online search for car models depends on the product cluster the car belongs to.
H2b: The influence of discussion topics in online review text on offline sales for car models depends on the product cluster the car belongs to.

**Control Variables**

Extensive academic research has identified that review volume (Li et al., 2019; Wijnhoven & Plant, 2017) and review rating (Li et al., 2019) influence product outcomes. Hence these two variables are included as controls along with previous period sales and online search volume.

**Methodology**

The details of data and the method adopted to determine the optimal product clusters and topics are discussed here. While the choice of variables to create product clusters is based on prior literature, the decision on optimum number is determined analytically based on numerical measures. The details are provided below.

**Data Collection**

The study is set in the Indian automobile sector, which comprises ten segments such as Compact, Coupe, Mini, and others. The study uses the Compact segment, which makes up around sixty percent of the Indian automobile sector, as the context of the study. The analysis is done in the period January 2016 to December 2018. The monthly car model sales have been collected from the CRISIL database from January 2015 to December 2018. The monthly online search volume has been collected from the Google trends website for the same period. Close to 23,000 online reviews have been scrapped using Python-based Web crawlers from the popular online product discussion forum mouthshut.com. The final analysis retains data for cars with non-zero sales and at least 100 reviews during the study period. The final dataset is panel data of 26 cars and 15,702 reviews.

**Determining Product Clusters**

The average sales and the average volume of online search for each car model for 2018 are taken to identify product clusters within the overall Compact segment. Given that information and transaction are two goals in the decision journey, the volume of online search and product sales are two variables used for creating product clusters (Humphreys et al., 2021).

![Figure 1. Choosing the Optimum Number of Clusters](image-url)
One of the most popular clustering algorithms, the \textit{k}-means clustering algorithm (Chau et al., 2006), is applied to identify product clusters in the data. The optimum number of clusters suitable for the data are identified analytically using the "within-cluster sum of squares" (WSS) and the "Gap" statistic (Mehar et al., 2013; Tibshirani et al., 2001). The data is partitioned into clusters ranging from 1 to 10. The WSS and the "Gap statistic" for each cluster are noted. The optimal cluster for the data is where the plot of the “WSS” elbows and the “Gap statistic” is maximized. As figure 1 indicates, the elbow of WSS is at both 2 and 3 clusters. However, given the “Gap statistic” is maximum at 2 clusters, the optimum number of clusters for this data is 2. The R package “cluster” has been used to calculate the statistics and determine the optimum number of clusters (Lemenkova, 2019; Maechler et al., 2021).

Table 1 presents the number of Car Models in each cluster, and the average Car sales and Avg Search Volume in each cluster identified.

<table>
<thead>
<tr>
<th>Cluster Number</th>
<th>Number of Cars</th>
<th>Average Sales</th>
<th>Average Search Volume</th>
<th>Total Sales</th>
<th>Total Search Volume</th>
<th>Cluster Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>20,505</td>
<td>63.84</td>
<td>389,595</td>
<td>1,213</td>
<td>Low Volume (LVC)</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>139,224</td>
<td>270.5</td>
<td>1,113,798</td>
<td>2,164</td>
<td>High Volume (HVC)</td>
</tr>
</tbody>
</table>

Table 1. Optimal Cluster Summary

Table 1 shows that cluster 1 has a higher number of cars but much lower average sales and average search volume and is hence given the name "Low Volume" (LVC). Cluster 2 has fewer cars but has a much higher average sales and search volume and is given the name "High Volume" (HVC).

\textbf{Extracting Review Topics – Latent Dirichlet Allocation}

The Python sci-kit-learn library (Beysolow II, 2018) is used for applying the LDA and determining the key topics of discussion in the published online reviews. Different Topic models with 2, 3, 4, and 5 topics are developed (Vallurupalli & Bose, 2020), and the perplexity scores for each model are noted. The number of topics in the model with the lowest perplexity score is the optimum number of topics for the data (Blei et al., 2003). Figure 2, which plots the perplexity scores for various topics, shows that 2 Topics are suitable for the analysis data.
Table 2 gives the details of the identified Topics, the associated keywords, and the Topic name.

<table>
<thead>
<tr>
<th>Topic Number</th>
<th>Number of Reviews</th>
<th>Percentage of Reviews</th>
<th>Topic Keywords</th>
<th>Topic Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11,842</td>
<td>75%</td>
<td>good, look, buy, space, comfortable, mileage, nice, family, support</td>
<td>Family comfort</td>
</tr>
<tr>
<td>2</td>
<td>3,860</td>
<td>25%</td>
<td>engine, feature, diesel, petrol, rear, power, hatchback, design, speed, interior</td>
<td>Performance</td>
</tr>
</tbody>
</table>

Table 2. Topics Identified from Latent Dirichlet Allocation

**Variable Operationalization**

The topics of discussion in the review text are the variables being studied. The output of the LDA assigns a probability score for each of the 2 topics to each review. The review is classified under the Topic with a higher probability score. The number of reviews for each Topic, in each month, for each Car model is calculated to study the influence of the 2 topics on the outcomes of interest.

The control variable review volume is operationalized as the number of reviews published for the car model each month. The other control variable, review rating is operationalized by taking the average of the daily reviews published for each car model each month.

**Model Equations**

The dataset has a panel structure with 26 car model values varying across 36 months. All the variables are log-transformed and standardized by subtracting the average at the car model level from individual month values. The two equations to be estimated for the study are presented below.

\[
\text{search}_{i,t} = \alpha_0 + \alpha_1 (\text{theme}_{i,t-1}) + \beta_2 (\text{vol}_{i,t-1}) + \beta_3 (\text{rat}_{i,t-1}) + \beta_4 (\text{srch}_{i,t-1}) + \varepsilon_{s,t} \quad (1)
\]

\[
\text{sales}_{i,t} = \gamma_0 + \gamma_1 (\text{theme}_{i,t-1}) + \lambda_2 (\text{vol}_{i,t-1}) + \lambda_3 (\text{rat}_{i,t-1}) + \lambda_4 (\text{sales}_{i,t-1}) + \lambda_5 (\text{sales}_{i,t-1}) + \varepsilon_{p,t} \quad (2)
\]

The equations estimate the influence of the review topics \((\alpha_1 / \gamma_1)\) on the monthly volume of online search (eq 1) and the monthly sales (eq 2) for car model \(i\) in month \(t\) when controlling for the influence of the previous month's review volume \((\beta_2 / \lambda_2)\), review rating \((\beta_3 / \lambda_3)\), the volume of online search \((\beta_4 / \lambda_4)\) and car sales \((\lambda_5)\). The equations are estimated for the overall dataset with 26 car models and the 2 clusters HVC and LVC with 19 and 7 cars, respectively.

**Empirical Results**

The two equations are estimated using ordinary least squares analysis. The results are presented in tables 3 and 4 at the overall level \((n = 26\) cars) and for the 2 clusters, LVC \((n = 19\) cars) and HVC \((n = 7\) cars).

**Model Results – Online Search Volume**

Table 3 shows how the review topics influence the volume of online searches. The results for the cluster HVC show the highest adjusted r-square. Given the review volume, rating, and the previous month’s online search volume, more discussion on Family Comfort has a significant negative influence \((\alpha_1 = -0.05, p < 0.05)\) on the overall volume of online search and in the LVC cluster, while having no influence in the HVC. More discussion on Performance has no overall influence. It has a significant negative influence in the LVC \((\alpha_1 = -0.09, p < 0.01)\) and a strong significant positive influence on HVC \((\alpha_1 = 0.12, p < 0.05)\). These results indicate support to hypotheses 1a and 2a.
Model Results – Product Sales

Table 4 presents the results for car sales. The results show that for the overall segment and LVC, the discussion topics in the review text do not influence sales. In contrast, the control variable review volume has a significant positive influence ($\lambda_1 = 0.14$, $p < 0.001$). In the HVC, the review volume has no influence, while both the topics Family Comfort ($\gamma_1 = 0.14$, $p < 0.05$) and Performance ($\gamma_2 = 0.20$, $p < 0.01$) have a strong positive influence. These results indicate partial support to hypothesis 1b and support hypothesis 2b. The adjusted R–square for HVC is the lowest.

Summary of Findings

The results have interesting managerial implications.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Finding</th>
<th>Implication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Conclusion**

The proposed research questions are answered by bringing together the car model sales, car model online reviews, and the volume of online information search in the Indian automobile sector. The LDA is used to cluster the review texts into two discussion topics. As information and transaction are two goals of the decision journey (Humphreys et al., 2021), the online search volume and offline sales of the year 2018 are used to cluster the car models into High volume and Low volume clusters. The popular k-means clustering (Chau et al., 2006) algorithm is used on the average car sales in 2018 for determining the product clusters.

The results indicate that online user-generated textual content explains the online product research and offline sales in the Indian automobile sector. The nature of the influence differs depending on the specific clusters a car model belongs to within the larger Compact segment. A cluster-wise influence of user-generated content as studied here has not been well-studied in the Western economies studying user-generated content. The research contributes to management research in the areas of online search (Grant et al., 2007), user-generated content (Bao et al., 2021; Bauman & Tuzhilin, 2021), the role of digital content in Asian economies (Karhade & Kathuria, 2020) and web mining (Chen & Chau, 2004). The study also directs managers to the importance of utilizing digital content for better decision-making.

As part of immediate future research, this analysis will be expanded to include other segments within automobiles and other product categories across geographies.

**References**


