Original Research Article

Text mining techniques for Exploring Customer Sentiments towards Packaged Organic Foods in India

Parveen Siwach[1], Shweta Dahiya[2], Prachi Aggarwal[3]*

*1 School of Commerce & Finance, Amity University, Punjab, 201313, India, Email: siwach.parveen23*gmail.com ORCID: https://orcid.org/0000-0002-2770-2251*

2 School of Business, World University of Design, Sonipat, Haryana, 131001, India, Email: shweta.dahiya@yahoo.co.in ORCID ID: https://orcid.org/0000-0001-7056-8716

3 Department of Food Business Management, National Institute of Food Technology Entrepreneurship and Management, Haryana, 131028, India. Email:prachi6196@gmail.com

***Corresponding author**：*shweta.dahiya@yahoo.co.in*

Abstract: Changing customer perception towards organic food is inspiring people to shift to organic counterparts of conventionally available packaged food products. They are willing to purchase the organically manufactured and packaged products in spite of a premium price tag. They thus possess some inherent expectations post purchase of organic food products. The purpose of this study is to explore customer levels of satisfaction and dissatisfaction on various attributes or features of packaged organic food products available in Indian market. For this purpose, 26,905 Online Customer Reviews of 100 unique packaged organic food products were collected and analysed using Word Clouds and Bar Graphs. Text mining techniques of Natural Language Processing (NLP) were applied to transform the unstructured textual review data into a structured form for identification of meaningful patterns and useful insights. Results confirmed that the customers were more dissatisfied than they were satisfied on product features of Price and Smell. On the other hand, Flavor, Delivery and Taste were the factors of highest customer satisfaction as compared to dissatisfaction shown towards each of these product features. Overall, Taste, Quality and Packaging saw maximum number of customer reviews as well as overall customer satisfaction post purchase of organic food products online. The study has implications for common shoppers of organic food products, manufacturers, marketers as well as retailers. The research provides a better understanding of consumer opinion after purchase and consumption of organic food products online.

Keywords: Customer satisfaction, Packaged organic food, Natural language processing, Product attributes, Market insights.

1. Introduction

As people are turning more health conscious, they are leaning towards the purchase of organic food products. Awareness regarding health damaging effects of harmful chemical pesticides and fertilizers used in the production of conventional food is increasing among customers[1]. Other factors such as environmental concern, product quality and farmers' health concern are also driving people towards the regular consumption of organic food products[1,2]. Organic food is produced, processed and stored without the use of chemical fertilizers, pesticides, herbicides and other synthetic chemical substances. It improves environmental sustainability as it is free from any genetically modified material. Organic agricultural practice strictly prohibits the use of growth hormones, antibiotics and other artificial substances in livestock breeding[1].

The Indian organic food industry is at a nascent stage of development. Indian organic food market is growing steadily in the domestic as well as export sectors[4]. India ranks 1st worldwide in terms of total number

Received: 15 June 2023 Accepted: 19 July 2023 Available online: 20 Novermber 2023

of organic food producers[4]. Favourable agro-climatic conditions and an inherited tradition of organic farming attract high number of organic producers and marketers to tap the promising market[4]. Today, many organic brands are emerging in the market as more and more customers are demanding for safe and healthy food choices. It therefore becomes increasingly crucial to understand organic food consumer behaviour as he makes a purchase decision despite a premium price of organic label, thus possessing some inherent expectations.

Advancements in digital technology has made it easier to gather large numbers of detailed customer reviews online. Although the data collection process has been digitized and made more convenient via ecommerce websites such as Amazon, Flipkart and Big Basket, the summarization of online customer reviews to gain useful insights, ultimately makes this data of use to the product manufacturer. It is also useful to common shoppers in making an informed purchase decision[3]. A manufacturer can combine reviews from multiple websites and generate a common summary report for each product[3]. He can further channelize his efforts into resolving customer issues related to the product or service as inferred from the summary reports. A marketer can learn about the product aspect(s) leading to higher customer satisfaction and use this knowledge to attract potential customers towards the brand. The summarized customer reviews are also helpful to the online and offline retailers in product selection and assortment.

Mining and summarizing customer reviews has been a challenging task which has been addressed by various researchers over the past few decades. The present study is based on several approaches proposed by researchers in combination with some self-suggested modifications with respect to the type of product(s) under study. In this research, we mine and summarize online customer reviews of various packaged organic food products available in Indian market to explore the factors leading to customer satisfaction and/ or dissatisfaction. As stated earlier, an easy understanding of customer sentiments is essential in applications of new product development, product purchase decision, product selection and assortment, improvements in existing product features as well as marketing of the brand to attract potential customers.

Data extraction and analysis was performed on an open source software R which is widely used for statistical computing and graphics. Over 26000 reviews were collected from Amazon.in for 100 unique packaged organic food products. Text mining algorithms were built to convert unstructured data into structured form for easy interpretation.

The textual reviews were pre-processed prior to “review mining” and extraction of aspect-opinion pairs[7]. Text was cleaned: Sentence fragmentation; removal of stop words, punctuations, digits, single letters and white spaces; followed by lemmatization and tokenisation. Lemmatization is the conversion of words into their base forms, e.g., words “tastes” and “tasted” were converted into “taste”. Tokenisation refers to breaking sentences into individual words.

The very first subtask was to extract top explicit as well as implicit product features that the customers mainly talked about in reviews. Explicit product feature appears in the review, e.g., Taste in “Taste is great”. It is clear by the appearance of the word Taste that the customer is talking on the taste aspect of the product he purchased. Some customer’s mention the product feature of judgement implicitly, e.g., Taste in “It was bitter”. Extracting implicit features poses a challenge as these cannot be identified by simple means[6]. To solve the problem, seed lists of words were prepared which contained Parts-of-Speech other than the Nouns, used by customers for expressing their opinions. The Nouns appear mostly as explicit product features and are therefore easy to extract via Part-of-Speech (POS) tagging. The POS tagging algorithm tags each token or word against its part of speech, e.g., “tasty” is converted to “tasty/JJ” (JJ stands for Adjective). Top frequent POS tagged tokens were studied and seed lists of features along with their exclusive opinion words were prepared manually. Exclusive opinion word refers to the opinion word which can only or mostly be used for a particular product feature as opposed to the generic opinion words. For example, “great”, “amazing”, “pathetic”, “horrible” are generic opinion words which can be used for various product features. On the other hand, “tasty”, “bitter” or

“delicious” imply the subjective opinion of customer on the feature “taste” which cannot be used for expressing opinions on other features such as quality, packaging or price. Each token of a seed list for respective product feature was then used in extracting sentences mentioning that particular product feature. Grouping synonyms together facilitated the procedure and reduced the number of tokens in each seed list. For example, replacing “delicious” and “yummy” with “tasty” allowed for the extraction of review sentences based on a single word “tasty” in one step. It also helps in better generation, representation and easy interpretation of final results as the collective frequency of a group of synonyms is raised. Without grouping similar words together, the synonyms would be scattered wide in Word Clouds as well as would not appear as one of the top frequent words in frequency tables.

Second subtask was the identification and extraction of opinion words linked to each shortlisted product feature. It took into consideration the negatively oriented opinions expressed with the help of words “no”, “not” and “never”. For example, in the sentence “Taste was not good”, the orientation of the opinion word “good” changed to “bad” as it was preceded by the word “not”. Therefore, the sentences containing words “no”, “not” or “never” were filtered out and treated separately. The opinion words in close proximity of “no”, “not” or “never” were rendered an opposite meaning by converting them into relevant antonyms. A seed list was prepared for conversion of such words into their antonyms after carefully studying their orientations in the review sentences. Only the words in close proximity of “no”, “not” and “never” were filtered out and rendered an opposite orientation, the rest of the sentence was treated with the rest of the reviews. For example, “taste of the product was not good, it tasted bitter and awful”. The sentence was interpreted as “taste not good”, “taste bitter and awful”. Further, “taste not good” was converted into “Taste bad”. Words “bitter” and “awful” were not rendered an opposite meaning just because they appeared in a sentence containing word “no”, as they did not appear in close proximity of “no”. Therefore, the final opinion words extracted for the product feature taste were “bad”, “bitter” and “awful”. Likewise, the opinion words for top 11 most frequently appearing product features were extracted and graphically represented via Word Clouds.

The final subtask included separating relevant opinion words into positive and negative for satisfaction and dissatisfaction levels respectively. The frequencies for satisfaction and dissatisfaction towards each of the top 11 product features were computed. Finally, the overall levels of satisfaction as well as dissatisfaction of customers on different product features were compared on a Double bar graph.

The three main objectives of this study are:

(1) To identify the most frequently reviewed product attributes of packaged organic food products by customers on Amazon.in.

(2) To extract the opinion words for each of the identified product attributes: Taste, Quality, Packaging, Money value, Price, Flavor, Freshness, Originality, Smell etc.

(3) To compute and compare the overall customers’ sentiments for each of the identified product attributes: Sentiment Analysis.

2. Review of Literature

A positive customer perception of organic food induces customers to switch over to organically grown and marketed food products even if they need to pay a premium price for it[2]. Suggested that four factors which influenced consumer behaviour towards organic food products included health consciousness, knowledge of organic foods, subjective norms and perceived price. The purchase intention was affected by an additional factor of availability. The study also positively concluded the significant influence of three socio-demographic factors viz., age, income and education on actual buying behaviour of customers. Their study followed the approach of collecting survey data through a structured questionnaire and analysing it using various techniques of regression analysis.

Many similar research studies have been conducted in last two decades which have mainly focussed on consumer buying behaviour and suggested factors which affected customers' purchase intention towards organic food. The data in these studies have primarily been collected through questionnaire surveys and face-to-face interviews. These techniques of data collection, however, limit the amount of data gathered and are often restricted by initial setups[5].

In light of the research gap in determining post purchase factors of customer satisfaction/ dissatisfaction, this study focusses on evaluating customer reviews using computer-based text mining and summarizing techniques. The present study, therefore, attempts to extract online customer reviews to increase the scope of data collection as well as the amount of data collected. The customer reviews are further analysed to study the factors of customer satisfaction and/ or dissatisfaction as customer reviews of products represent their satisfaction with a product[5].

Determined the factors influencing customer satisfaction while purchasing organic products online as well as predicted their sales volumes. The research used LDA (Latent Dirichlet Allocation) method along with regression analysis methods to identify important factors for selling organic products online. These factors included: packaging design, nutritional information, food quality, delivery risk, product freshness and source risk. While the packaging design, nutritional information, and food quality had a positive effect on customer satisfaction; the delivery risk, product freshness and source risk had negative effect on satisfaction[5]. Data was collected using questionnaire as well as mining web customer reviews of products.

Our study is based on the methodology proposed by[3] for mining and summarising product reviews based on NLP (Natural Language Processing) methods[3]. suggested the use of data mining and NLP techniques for effectively mining product features as well as their opinion sentences and semantic orientations. They used Part-of-Speech Tagging (POS) to extract product features which usually occurred as nouns and noun phrases in review sentences. These product features are the product attributes that the customers are reviewing and expressing their opinions upon. Similarly, adjectives were used as "opinion words". Opinion words are the words that are used by customers to express their subjective opinions[3]. Their study also identified semantic orientations of each opinion word for predicting the orientation of an opinion sentence, i.e., positive or negative. Finally, a feature-based review summary was generated for each product feature of interest which highlighted related opinion sentences as positive and negative according to the opinion sentences' orientations[3].

Our study differs from[3] in using top frequent nouns as product features after manual evaluation instead of association mining technique used by[3] to find frequent itemsets which were considered to be the product features. Additionally, parts of speech other than Adjectives were also used in identification of opinion words after manually examining the original dataset as well as the results of POS tagging. These parts of speech included Nouns, Verbs and Adverbs[8].

We also differed from the proposed methodology on selecting Adjectives as well as other parts of speech such as Nouns, Verbs, Adverbs etc as opinion words after manually examining the original dataset and the results of POS tagging.

Various other modifications and suggestions in text mining techniques are made suiting the organic food product industry.

3. Research Methodology

3.1 Product Feature Identification and Selection

3.1.1 Data Collection

Online customer reviews (OCRs) were extracted from Amazon.in using an R script. A total of 26,905 OCRs were extracted covering 38 unique brands and 100 unique packaged organic food products listed on the website. These products fall under 8 broad categories as shown in table 3.1.

Table 1 Customer Review Data (Category Wise)

Product Category	No. of Products	No. of Reviews
Rice, Flour & Pulses	27	3422
Dried Fruits, Nuts and Seeds	21	8199
Spices & Condiments	18	2416
Sugars and Syrups	12	4956
Cereal & Muesli	8	1938
Tea & Coffee	6	2246
Oils & Ghee	4	2825
Ready to Eat	4	903

3.1.2 Text Cleaning & Pre-processing

The reviews were cleaned for the processes of product feature extraction and opinion mining. The data was first broken into individual sentences delimited by the punctuations ".", "?", "!" and ";". Further, sentences containing the conjunctions "but" and "however" were broken into smaller sentences for better interpretation. Words like "isn't", "aren't" etc were replaced with their base forms of "is not", "are not" respectively in order to later facilitate their filtration under a separate Data Frame for opinion word extraction. Next, the text was transformed into lowercase. Stop words, punctuations, digits, single letters and white spaces were removed in the subsequent steps.

3.1.3 Substitution and Removal of Selected Words

Words expressing the intensity of opinions were removed, for example, "very", "extremely", "little". Same words that spelled differently were converted into one, for example "flavour" was substituted with "flavor". Popular brand names were removed as they occurred frequently in reviews. Some other iterations were made to clean and prepare the text for the next step.

3.1.4 Lemmatization

The words were reduced to their base forms. For example, words "loved", "loves" and "loving" were converted to "love". It helps in raising the collective frequencies of similar words to be able to study and evaluate them together.

3.1.5 Grouping of Synonyms

Similar words were manually identified and replaced with a common word, usually having the highest frequency in original dataset. For example, words "yummy" and "delicious" were replaced with "tasty". Also, some of the lemmatized words were replaced with a more meaningful form to better present their meaning under opinion mining. For example, words "defect" and "damage" were replaced with "defective" and "damaged" respectively. Finally, the product names were replaced with the word "product" itself.

3.1.6 Tokenisation and POS Tagging

The continuous strings of characters in each sentence were broken down into individual words or "tokens" delimited by blank spaces. Next, each token was given a Part-of-Speech (POS) tag using the libraries "rJava", "openNLP" and "NLP".

3.1.7 Product Feature Identification and Selection

A frequency table of POS tagged tokens was formed. The top frequent "Nouns" as well as some other Parts-of-Speech expressing exclusive opinions were studied and evaluated as potential product features. The other parts of speech included Verbs, Adverbs, Adjectives etc. For example, words "overpriced/JJ", "expensive/JJ", "affordable/JJ" (JJ stands for adjective) were also included with the product feature price/NN (NN stands for Noun) and their collective frequencies of occurrence were computed to calculate the number of times customers talked about the attribute price. Top 11 product features were extracted along with their frequencies in the original dataset. The results were plotted on a bar graph.

3.2 Opinion Mining

3.2.1 Filtering Sentences Mentioning a Product Feature

After text cleaning & pre-processing, strings mentioning a particular product feature explicitly or implicitly, were filtered out. Explicit feature appeared directly in the sentence, for example price in "product is not worth the price". While implicit features appeared hidden, for example price in "product is very expensive hence unaffordable".

3.2.2 Reduction of String Length to Contain Only Nearby Tokens

The strings were reduced to contain a maximum of 7 characters including the product feature itself which lied on the 3rd position. That is, two words prior to the product feature, the product feature and four words after the product feature were extracted in the same order as mentioned. The dataset thus had a maximum of 7 words per sentence or row. There was no minimum limit. This allowed for the extraction of only most relevant and correct opinion words used for a product feature as the adjective in close proximity of a noun has maximum chance of being a correct opinion for that noun.

3.2.3 Filtering out Strings with Negative Orientation

Strings containing words "no", "not" or "never" were filtered out from the original dataset to separately treat them as they rendered an opposite meaning to the opinion word used in close proximity. The string size of these sentences was further reduced to contain a maximum of 3 characters, starting with the words "no", "not" or "never". That is, only two characters following the words "no", "not" or "never" were kept in the new dataset and the rest of the sentence was combined with the original dataset from which negatively oriented sentences were filtered out.

3.2.4 Tokenisation and POS Tagging

The original dataset and the dataset containing negatively oriented strings were separately broken down into individual tokens and each token was assigned a POS tag. Next, all the Adjective tokens were extracted along with some other parts of speech such as Noun, Adverb, Verb etc which acted as opinion words. A seed list of these non-adjective opinion words was manually prepared after evaluation of POS tagged tokens listed in the frequency table prepared in step no. 7. This seed list was used in extraction of opinion words other than the Adjectives.

3.2.5 Conversion of Negatively Oriented Opinion Words Into Respective Antonyms

The opinion words extracted from the negatively oriented dataset were converted into their most relevant antonyms. A seed list for the opinion word-antonym pair was prepared to facilitate the process. For example, word "worthy" was converted into "worthless" as it was preceded by the word "not".

3.2.6 Word Cloud for Graphical Representation of Opinion Words for Each Product Feature

Finally, the opinion words from both datasets were combined and a frequency table was prepared. Algorithm for the generation of Word Clouds was used to graphically represent the overall opinion of customers for each of the 11 product features.

3.3 Sentiment Analysis

3.3.1 Assigning Positive and Negative Sentiments to Opinion Words

The relevant opinion words from the frequency table were extracted and assigned positive or negative orientations with respect to the product feature. For example, "good" was considered as a positive sentiment which led to higher customer satisfaction. Likewise, "bad" was given a negative sentiment which led to customer dissatisfaction. For each product feature, lists of positive and negative opinion words were created which led to customer satisfaction and dissatisfaction respectively, along with their respective frequencies of occurrence.

3.3.2 Comparing Levels of Customer Satisfaction and Dissatisfaction Between Different Product Features

The final comparison was plotted on a double bar graph. The X axis indicated different product features, the Y axis computed frequencies or levels of customer satisfaction as well as dissatisfaction towards each of the product feature listed on X axis.

4. Results and Conclusions

4.1. Product Feature Identification

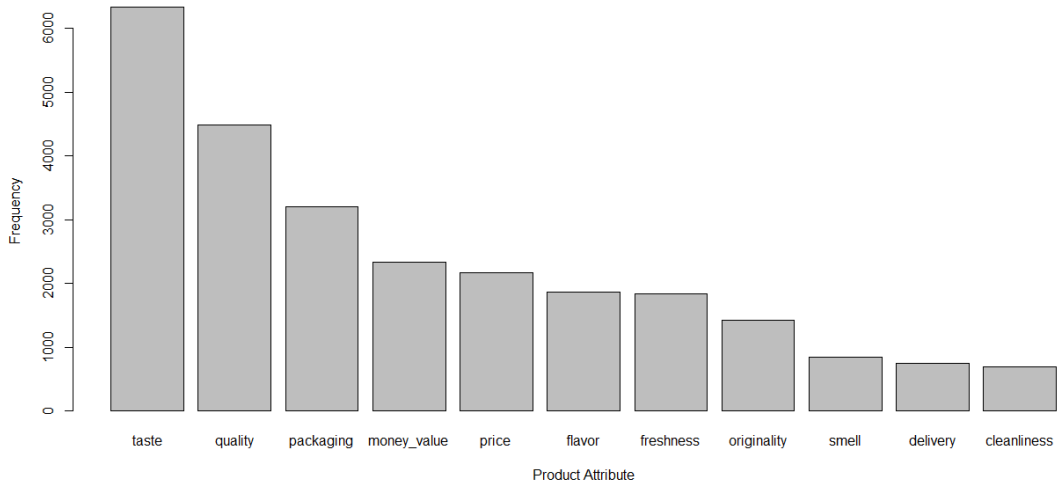


Figure 1 Taste, Quality and Packaging were top 3 product features customer mostly talked about in their reviews online.

4.2. Opinion Words' Extraction

4.2.1 Product Feature: Taste

4.2.2. Product Feature: Quality

4.2.3. Product Feature: Packaging

4.2.4. Product Feature: Money Value

4.2.5. Product Feature: Price

4.2.6. Product Feature: Flavor

4.2.7. Product Feature: Freshness

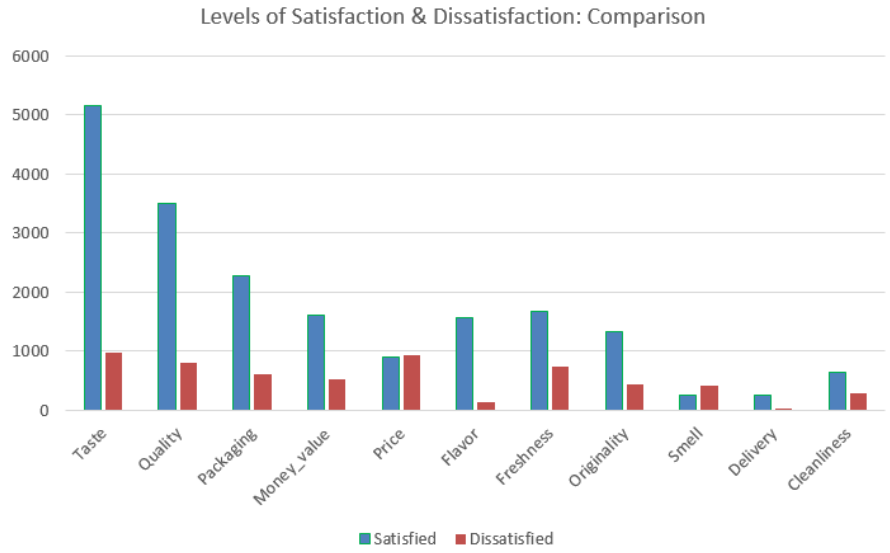
4.2.8. Product Feature: Originality

4.2.9. Product Feature: Smell

4.2.10. Product Feature: Delivery

4.2.11. Product Feature: Cleanliness

4.3. Sentiment Analysis & Comparison



From the above comparison, it can be concluded that the product features of Price and Smell saw higher customer dissatisfaction than satisfaction. Customers showed maximum satisfaction on the attributes of Flavor, Delivery and Taste in comparison to the levels of dissatisfaction on each of these attributes.

5. Implications of the Study

The results of this study have implications for the shoppers, manufacturers, marketers as well as online and offline retailers of packaged organic food products. Common shoppers can refer to the results in making an informed purchase decision which is based on their considerations towards each factor independently. The customer to whom Quality and Taste of the organic food products matter more than other attributes, might find it worthy to invest his or her money into the product. On the other hand, if Price as well as the Freshness of the organic food hold more importance, the customer might not be motivated enough to try the product.

The manufacturers can take help of the results and focus on improving the features leading to higher customer dissatisfaction than satisfaction. They can channelize their efforts towards improving product Cleanliness, overall aroma and making their products available at lower market Prices to attract more customers. The marketers can reorient their marketing strategies and highlight the features of Flavor, Taste and Quality to convert more customers for the purchase. They shall also highlight prominently the organic certification(s) possessed by the product as Originality is one of the key features important to the customers when they assess and review the organic product post purchase.

Retailers can also compare the results presented in this study with other products and decide whether to keep organic products in their stores or not as well as how to position them.

6. Limitations and Future Research Directions

Authenticity of online customer reviews poses the biggest challenge in making interpretations of results of this study. While efforts were made in extracting reviews mostly from “verified purchases” only, reports suggest that there is no significant link between the verification of purchaser and the authenticity of posted review. Future research studies may also gather offline reviews directly from customers or manufacturers/ retailers for improving the reliability of data collected for analysis.

The text mining approach used for data analysis could only read and interpret rightly spelled reviews. The informal language could not be interpreted as well as language(s) other than English.

The text mining approach is also based on the usage of different lexicons. The lexicons are pre-defined

dictionaries which are used for various tasks. For example, lexicons were used in finding stop words, lemmatized forms of words as well as part-of-speech tagging. The lexicons might not be updated, or new words might be missing from the dictionaries which may not have existed when the lexicon was first drafted. To avoid higher dependencies on lexicons, majority of the dictionary based tasks were completed with manual intervention. For example, categorising words into positive or negative during the final step of computing levels of satisfaction and dissatisfaction was done after examining each word as well as its orientation in the review sentences.

Finally, with the current method, intensities of opinions could not be studied because of the removal of words like “very”, “extremely” and “pretty”. Although, at some level it was possible to understand the expressions of customers better when they used different words to reflect their intensities of opinion, e.g., “okay”, “good” and “great”.

The present study although was conducted on computer making use of the statistical software R Studio, was not entirely computer based, neither did it make use of Artificial Intelligence in analysing unstructured textual review data. It also involved some manual work which was carried out at various steps for generation of most reliable results. Future researchers can make maximum use of various supervised and/ or unsupervised learning methods in text mining for ensuring minimum manual intervention and ease of processing data.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Basha, M.B.; Mason, C.; Shamsudin, M.F.; Hussain, H.I.; Salem, M.A. Consumers attitude towards organic food. Procedia Econ. Financ. 2015, 31, 444–452.
2. Anupam Singh, Priyanka Verma. Factors influencing Indian consumers' actual buying behaviour towards organic food products. Journal of Cleaner Production 167 (2017) 473-483
3. Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews.
4. APEDA, 2020. National Programme for Organic Production (NPOP). Retrieved from: apeda.gov.in/apedawebsite/organic/Organic_Products.htm
5. Lyu, Fang, and Jaewon Choi. 2020. The Forecasting Sales Volume and Satisfaction of Organic Products through Text Mining on Web Customer Reviews. Sustainability 12, no. 11: 4383
6. T. A. Rana and Yu-N Cheah, Hybrid rule-based approach for aspect extraction and categorization from customer reviews, 2015 9th International Conference on IT in Asia (CITA), 2015, pp. 1-5
7. Naime F. Kayaalp & Gary R. Weckman & William A. Young II & David Millie & Can Celikbilek, 2017. Extracting customer opinions associated with an aspect by using a heuristic based sentence segmentation approach, International Journal of Business Information Systems, Inderscience Enterprises Ltd, vol. 26(2), pages 236-260
8. Exploiting user experience from online customer reviews for product design. Bai Yanga, Ying Liub,□, Yan Liangb, Min Tanga