Applied Artificial Intelligence

An International Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uaai20

Discovering the relationship between attributes of facial masks and review rating in online customer reviews using Explainable Artificial Intelligence (XAI)

Inmaculada Aparicio-Aparicio, Maria-Arantzazu Ruescas-Nicolau, Lirios Dueñas, José Luis Sánchez-Jiménez, M.Luz Sánchez-Sánchez & Enrique Alcántara

To cite this article: Inmaculada Aparicio-Aparicio, Maria-Arantzazu Ruescas-Nicolau, Lirios Dueñas, José Luis Sánchez-Jiménez, M.Luz Sánchez-Sánchez & Enrique Alcántara (2024) Discovering the relationship between attributes of facial masks and review rating in online customer reviews using Explainable Artificial Intelligence (XAI), Applied Artificial Intelligence, 38:1, 2411780, DOI: 10.1080/08839514.2024.2411780

To link to this article: https://doi.org/10.1080/08839514.2024.2411780

Published online: 09 Oct 2024.

APPLIED ARTIFICIAL INTELLIGENCE
2024, VOL. 38, NO. 1, e2411780 (25 pages)
https://doi.org/10.1080/08839514.2024.2411780

Discovering the relationship between attributes of facial masks and review rating in online customer reviews using Explainable Artificial Intelligence (XAI)

Inmaculada Aparicio-Aparicio [a*], Maria-Arantzazu Ruescas-Nicolau [b*], Lirios Dueñas [b], José Luis Sánchez-Jiménez [a], M.Luz Sánchez-Sánchez [b], and Enrique Alcántara [c]

[a]Department of Physical Education and Sports, Research Group of Sports Biomechanics (GIBD), Universitat de València, Valencia, Spain; [b]Physiotherapy in Motion, Multispeciality Research Group (PTinMOTION), Department of Physiotherapy, Universitat de València, Valencia, Spain; [c]Dr. Q. Data Driven Innovation and Association of the Valencian Community for Driving R&D and Impact innovation in Sports (4ivlcesport), Valencia, Spain

ABSTRACT

The COVID-19 pandemic led to a surge in facial mask usage. Analyzing customer feedback is crucial for enhancing this product, given the abundance of options and online reviews available. Techniques like Latent Dirichlet Allocation (LDA) combined with machine learning (ML) models help identify product attributes and their impact on reviews. However, successful models frequently employ black-box techniques that fail to explain how product attributes impact customer satisfaction. To address this, Explainable Artificial Intelligence (XAI) combined with advanced ML methods is suggested to create interpretable models. This paper evaluates the usefulness of combining LDA for topic modeling, Gradient Boost trees (GBT) for ML, and SHapley Additive exPlanations (SHAP) to develop an interpretable model for face mask attribute impact on online ratings. Analyzing 2,047 reviews of 35 mask products revealed seven key attributes. Labeling and nose-strap significantly influenced evaluations, while shopping experience, and packaging, filtering level, and fit were less impactful. This research supports the use of topic modeling combined with advanced ML techniques and XAI in analyzing online customer reviews to offer a time- and cost-effective method. It aids in understanding product attributes influencing satisfaction for product design and improvement, especially in the dynamic context of face mask preferences and purchasing decisions.

ARTICLE HISTORY
Received 14 March 2023
Revised 6 September 2024
Accepted 29 September 2024

Introduction

The increase in demand for facial masks due to the COVID-19 pandemic has accelerated the development and launching of new models into the market,

CONTACT Maria-Arantzazu Ruescas-Nicolau Arancha.Ruescas@uv.es Physiotherapy in Motion, Multispeciality Research Group (PTinMOTION), Department of Physiotherapy, Universitat de València, C/Gascó Oliag, 5, Valencia 46010, Spain

*These authors contributed equally to this work.

reflected in the increase in published papers and research on this product (Al-Rammahi 2022; Bakhit et al. 2021; Li et al. 2021). Published works has focused on protective aspects of masks (e.g., filtration efficiency and materials) (Chua et al. 2020; Tcharkhtchi et al. 2021), durability (Chen et al. 2022; Howard et al. 2021; Varanges et al. 2022), fit (O'Kelly et al. 2021) and life cycle, cleaning, and comfort (Chua et al. 2020).

During the COVID-19 pandemic, millions of people wore different types of facial masks (e.g., ffp2 and kn95) in various contexts. Studies highlight some harms of wearing masks, such as irritation (Bakhit et al. 2021; Scheid et al. 2020; Tasic-Kostov et al. 2022), discomfort (Bakhit et al. 2021), subjective breathing difficulties (Bakhit et al. 2021; Scheid et al. 2020), skin rashes (Bakhit et al. 2021), difficulty in communication (Shack et al. 2020), or headache with prolonged mask use (Bakhit et al. 2021; Scheid et al. 2020). These adverse effects may have reduced the acceptability, adherence and effectiveness of masks (Bakhit et al. 2021) and influenced consumer choices, which are determined by how people think and feel about a product (Imtiaz and Ben Islam 2020). Hence, methods and designs to mitigate the disadvantages of face mask use based on human reasoning and evaluation should be addressed in current research. Identification of user needs, preferences, and requirements could be helpful in this regard.

Traditional methods, such as focus groups, surveys, or interviews, have been used for this purpose. These methods study the user experience by creating different scenarios, such as tasks and usage environments, in which the user interacts with the product. That is one of the reasons they are considered costly and time-consuming (Chua et al. 2020). Therefore, there is a growing interest in using online reviews to lessen the time and cost of this step in the product and service development process (Chua et al. 2020; Dash, Zhang, and Zhou 2021; Fernandes et al. 2022; Fernando and Aw 2023; Imtiaz and Ben Islam 2020; Li et al. 2020; Noori 2021; Rathore and Ilavarasan 2020Nasiri and Shokouhyar 2021; Sharma and Shafiq 2022; Son and Kim 2023; Suryadi and Kim 2018, 2019). In this sense, several authors have stated that customers´ online reviews can be seen as the result of a large-scale experiment involving physical interaction between people and products (Son and Kim 2023; Tuarob and Tucker 2015) in terms of traits related to usage context or product features that will considerably affect the way customers feel about products. Analyzing online reviews will then help improve facial masks and develop better products. Today, there is a general consensus that online reviews contain valuable information about customers' evaluation of the attributes of products, services, and experiences, which can be used to assess customers' requirements, needs, preferences, product defects, and important attributes (Berezina et al. 2016; Joung and Kim 2021; Kumar, Dwivedi, and Anand 2021; Li, Li, and Kim 2023; Neumann et al. 2019; Qiao et al. 2017; Shirdastian, Laroche, and Richard 2019; Trisna and Jie 2022; Xu 2019; Zhang et al. 2021). Some examples include vehicles (Sun et al. 2020), bikes (Chua et al. 2020), smartphones (Joung and Kim 2021), mobile phones (Trappey et al. 2016; Yang et al. 2019), and smartphones and automobiles (Tuarob and Tucker 2015).

To identify product features and attributes, published works employed a combination of text-mining tools of online reviews and statistical techniques. The most common tools are topic analysis methods, such as Latent Semantic Analysis (LSA) (Xu and Li 2016) and Latent Dirichlet Allocation (LDA). LDA is a machine-learning-generative statistical model that outputs a vector representation of the probability of words in a topic (Deshpande and Rokne 2018; Suryadi and Kim 2019).

In parallel, the relative importance of identified features or attributes indicates the weight people give them when making a purchase decision. This is usually assessed by analyzing their influence on customer satisfaction, sales ranking, overall assessment, sentiment, and other measures of product performance (Chua et al. 2020; Noori 2021; Qiao et al. 2017). Machine learning (ML) and deep learning (DL) methods have been used for this purpose. However, the highest accuracy for large datasets, such as those obtained from online reviews, is often achieved by complex black-box like models, such as ensemble gradient boosting trees (GBT) and ensemble or DL models (Joung and Kim 2021; So 2020). These models are difficult to interpret, e.g., when determining the influence of product attributes identified from online customer reviews. Therefore, there is a need to explore alternative methods.

Today, this is a general concern in Artificial Intelligence. As an option, Explainable Artificial Intelligence (XAI) is a field of research that aims to overcome this problem (Joung and Kim 2021; So 2020). Various methods have been developed to interpret the predictions of complex models and explain them to humans in an understandable way (Murdoch et al. 2019). SHapley Additive exPlanations (SHAP) (Joung and Kim 2021; Lundberg and Lee 2017; Lundberg et al. 2018) is one of the most widely used. SHAP issues how feature affects (negatively or positively) the outcomes of a model. It is a method to explain individual predictions based on the game´s theoretically optimal Shapley values. The "game" is the prediction task for a single dataset instance. The "gain" is the actual prediction for that instance minus the average prediction for all instances. The method explains a prediction by assuming that each feature value of the instance is a "player" and that they all combine to make the prediction, which is the profit. Each feature value receives a payoff from this contribution based on Shapley values. Thus, it operates between feature values and outputs only.

In this line, Joung and Kim (2021) combined LDA to detect product attributes from online reviews with aspect-based sentiment analysis (ABSA) (which estimates the polarity [positive, negative or neutral] of comments) to determine the importance of the identified attributes. They evaluated the impact of the sentiments of each product attribute on the overall rating via an explainable, deep neural network applying Shap on a rather large dataset of 33,379 cell phones reviews. Their results outperformed previous sentiment analyses and neural network-based methods. However, no publications have used applications combining XAI and ML classification methods like, for example, XGB, to establish the importance of product attributes in online reviews.

In view of the above, considering the possibilities offered by the SHAP method and the feasibility of online reviews and that, to the best of our knowledge, no previous research has applied topic analysis on face masks' online reviews, the following research questions were formulated:

RQ1: Can SHAP values be used to explain models developed with advanced black-box classification ML techniques to determine the influence of mask attributes on overall customer ratings in online reviews?

Consequently, such a model could be applied to answer the following research question in relation to mask innovation:

RQ2: Which mask attributes influence customer ratings in online reviews?

Therefore, the present study aimed to apply Explainable Artificial Intelligence (SHAP) to identify attributes and features of face masks from online customer reviews and their relevance to purchase decision-making to guide product improvement.

Materials and Methods

This study was conducted in August 2021. To prevent transmission of COVID-19 via droplets, it was recommended to use masks with a filtration efficiency of at least 95% for particles with a diameter of 0.3 microns or larger, as they are the most effective protection against airborne pathogens (Dugdale and Walensky 2020). The specialized filtering masks ffp2 and kn95 were the masks with the highest level of protection available to the general population at that time. Hence, we decided to focus on them.

The present paper follows a semiautomatic process (Figure 1) to chronologically capture online reviews, text-mine them, and statistically analyze the resulting data.

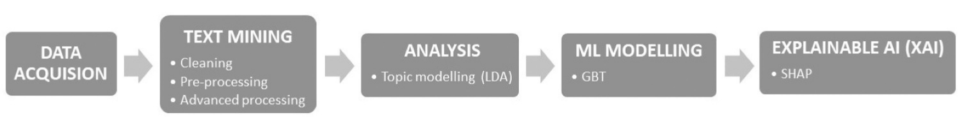

Figure 1. Diagram of the process followed. LDA: Latent Dirichlet Allocation; GBT: Gradient Boost trees; AI: Artificial Intelligence; SHAP: SHapley Additive exPlanations.

To identify customers´ perceptions of the features of the ffp2 and kn95 masks and their relevance to the general rating, online reviews were manually captured from Amazon webpages (Amazon.com, Inc., Seattle, Washington, United States). Only comments written in Spanish were collected. The data registered included the following information: (1) model of the mask, (2) type (kn95 or ffp2), (3) comment, (4) price, (5) number of units included in the price, and (6) customers´ opinions, reflected by the 5-star rating scale (the more the stars selected, the better the opinion).

A text-mining process was then carried out, which is summarized in Table 1. The set of words derived from this process forms the dictionary for the reviews.

Identification of Attributes Using LDA

LDA analysis was applied to identify latent topics. It was carried out using the Gensim module in Python (Python Software Foundation) in a Jupyter notebook within an Anaconda environment. The following criteria were used to determine the number of topics to be extracted:

- *Coherence* was computed for an increasing number of topics from 3 to 25. The influence of the model's hyperparameters, alpha and beta, related to the words' density in a text and the set of texts, respectively, was also assessed. The model's coherence refers to how different the words are in each topic. The more different they are, the more coherent the model is. Coherence was measured by the metric C_V. This metric results from applying a sliding window and a segmentation of the main words in a set, as well as an indirect confirmation based on the normalized pointwise mutual information (NPMI) and cosine similarity criteria. In the literature, this value is around 60%-70% (Lundberg et al. 2018; Qiao et al. 2017).
- *Low overlap*, measured by the distance between topics in the model. Greater distance indicates lower overlap, which is preferable. We utilized the Jacckard distance to quantify this overlap. The Jacckard distance measures dissimilarity between sample sets, representing the ratio of non-overlapping words to the total number of words in both topics. Specifically, for two topics t1 and t2, the Jacckard distance (J) is computed as:
  $d\_J\ (t1, t2) = (|t1 \cup t2| - |t1 \cap t2|)/(t1 \cup t2)$
  Here, $|t1 \cup t2|$ stands for all the words that result from merging both topics, and $|t1 \cap t2|$ refers to words that are in both topics.
  To visually assess overlap, we utilized a distance heatmap, a graphical representation of data where individual values in a matrix are depicted as

Table 1. Text mining process.

Data Preparation and Cleaning	Pre-Processing	Advance Processing
ntlk & spacy tools were used **(1)** Elimination of nulls and non-valid cases **(2)** Spellchecking (module Spellchecker) **(3)** Lower casin comments and deletion of characters, symbols and letters (@, #, urls . . .) **(4)** Elimination of punctuation and non-alphanumeric characters (including emoticons)*	**(1)** Tokenization to split texts into words **(2)** Stemming and lemmatizing. Verbal forms, plurals, and gender were reduced to just one form of each **(3)** Elimination of stopwords.** The present study was done in Spanish with an exploratory purpose, thus, specific stop words dictionaries were created for the different types of words used.	**(1)** Pos-tagging to classify identified words into linguistic categories with the aim of identifying synonyms and leaving one of them. **(2)** Frequency filtering was conducted, leaving only words that fulfilled the following appearance thresholds***: • Lower threshold = at least 1.5% of reviews (>33 comments) • Upper threshold: <30% of reviews.

ntlk: natural language toolkit
*For content analysis this step joins all sentences in a review in a same text and substitutes accented letters with non-accented letters and eliminates the letter *ñ*.
**Stop words do not contribute to the meaning of a sentence and, usually, are eliminated during text mining.
***The lower threshold defines a minimum for further statistical analysis reducing sparseness. The upper threshold eliminates words that are too frequent because they refer to "structural" aspects of the set of comments.

colors. In our case, the heatmap illustrates the distance between topics matrix, generated using the heatmap function from the seaborn Python library.

- *Interpretability* of topics. LDA assigns a weight to each word for each topic and determines the contribution of each topic to each definition. Topics are interpreted by analyzing words with higher contributions to them, as well as definitions in which a given topic has a high weight (Kapadia et al. 2020).

  Finally, the texts were represented as a vector, noting each topic contribution.

Relevance of Product Attributes in Customer Ratings

The relevance of the identified topics in the stars ranking was assessed using a GBT model. For this purpose, each rating was categorized as follows: 1 for scores higher than 3 stars (4 or 5) and 0 for the rest. This approach was used since this research aimed to investigate the impact of various attributes on the likelihood of the masks receiving less than 4 stars rather than achieving a specific value. The analysis was performed via Python software (Python Software Foundation) using the Gradient Boosting Classifier from scikit learn (Bentéjac, Csörgő, and Martínez-Muñoz 2021), with the identified topics as predictors and the categorical ranking as the target.

GBT is an assembling method that performs classification by combining the outputs of individual trees. It combines many decision trees to reduce the risk of overfitting each individual tree. This process applies boosting, combining simple decision trees (weak learners) to improve against a loss function (Hastie, Tibshirani, and Friedman 2009). In this case, *deviance* (= logistic regression) is used with probabilistic outputs for classification.

In this study, the following hyperparameters were optimized using cross-validation, with a random sample split into 60% of cases for training and 40% for testing:

(1) N_estimators (trees): This is a critical hyperparameter, as the risk of overfitting increases with the number of trees.
(2) Learning rate: This controls how fast the model learns and the risk of falling into overfitting. The lower it is, the more trees are needed to get good results, but the lower the risk of overfitting becomes.
(3) The max depth of trees in the model marks how much a single tree can learn. It tends to be low.

A common recommendation is to keep the learning rate low and increase the number of trees. A grid search was used to find the best combination of parameters with respect to root mean squared error (rmse) and accuracy with cross-validation.

(1) Initial N_estimators = 20, 100, 500, and 1000 trees.
(2) Learning rate = 0.001, 0.01, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, and 1.
(3) Max depth = None, 1, 3, 5, 10 and 20.

The performance of the resulting model is evaluated in terms of computing accuracy, log_loss, and rmse with 50-fold cross-validation. While accuracy measures the percentage of correct predictions (correct/incorrect), log_loss provides a more nuanced insight into the performance of the model. It takes into account the predicted probability (p, ranging from 0 to 1) of belonging to one class (or the other, 1-p) and measures the difference between the predicted probability and the actual outcomes. The log_loss function is defined as follows: log(p), where p is the probability attributed to the real class. Finally, the precision-recall curve for unbalanced samples is obtained. If the model performs better on the training set than on the test set, the model is likely overfitting.

Precision-recall curves summarize the trade-off between Precision, which is the number of true positives divided by the sum of the true positives and false positives, and Recall, which is the number of true positives divided by the sum of the true positives and the false negatives. They are particularly useful for imbalanced datasets (Saito, Rehmsmeier, and Brock 2015). A no-skill classifier is represented as a horizontal line with a positive-case-ratio value in the dataset. A model with perfect skill is illustrated as a point at (1,1). A proficient model is demonstrated by a curve that bends toward (1,1) above the no-skill flat line.

Explainable Artificial Intelligence (XAI): SHAP

The influence of predictors is measured via several GBT metrics and the Shap Library:

- *Permutation_importance*. This permutation method randomly shuffles each feature and computes the change in model performance. The features with the greatest impact on performance are the most important.
- *Increase in nodes purity*. This method computes the average increase in overall node purity for each tree in the model due to the topic. Purity is measured by the decrease in the metric used to decide node splitting (in this case, the Gini index).
- *SHAP package*. This is model-agnostic and uses the Shapley values from game theory to estimate how each feature contributes to the prediction. Given the current set of feature values, the contribution of a feature value

to the difference between the actual prediction and the mean prediction reflects the estimated Shap value. We obtained shap_values to measure the importance of features and plots (summary and force plots) to understand how each feature influences the predictions. The SHAP values assign an importance value to each feature in a model. Features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is.

Results

A total of 2,047 reviews were collected, from which, after text mining, 97 words with a frequency over 1.5% and under 30% were stored for further analysis. The collected reviews concerned 35 mask products. The distribution of the comments collected per product was imbalanced. The mode was 41 comments, while some models had more than 60 (maximum = 443), and others had fewer than 20 (minimum = 10). On average, each product had 58.5 comments (SD = 73.7).

Figure 2 shows the frequency distribution for the top 20 words. Strap, quality, comfort, and ffp2 appeared in more than 15% of the reviews. Nevertheless, this ranking allowed us to identify the aspects people bring up about face masks: quality, comfort, price, size, color, type, certification, fit, and presentation of the product (parcel and package), and other elements, such as the strap.

Concerning star rankings, Figure 3 shows that the sample was not balanced, as there were far more reviews with 4 or 5 stars (1547/2047) compared to those with less than 4 stars (500/2047).

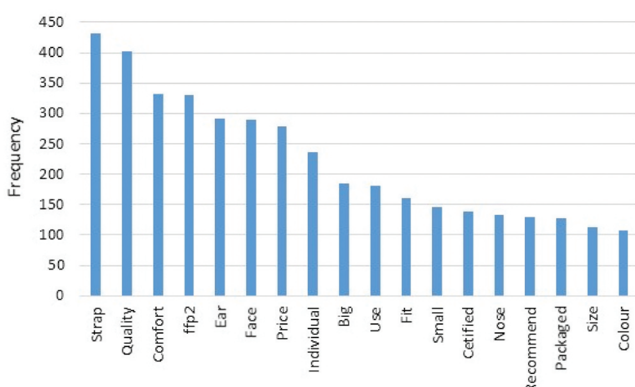

Figure 2. Frequency distribution of top 20 words identified in the reviews.

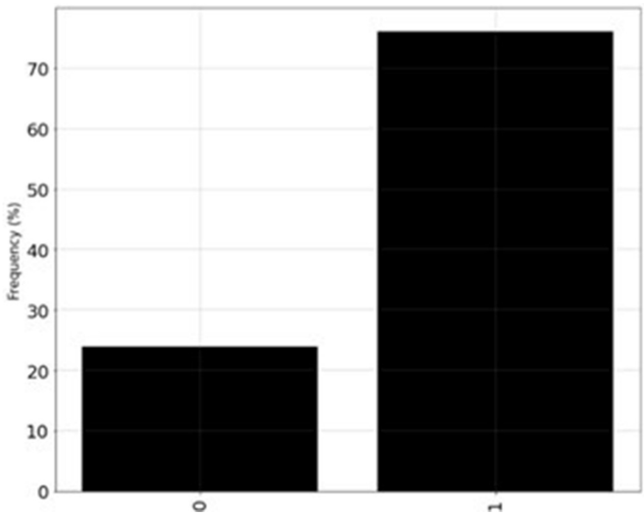

Figure 3. Distribution of stars (1 = 5-4 stars; 0 = 1-3 stars).

Topics

The results of the LDA coherence analysis (Figure 4) showed an elbow at around 65% for seven topics and then a steady increase to a peak of coherence of around 67% for 11 topics. After analyzing the overlap and interpretability, a solution with seven topics was chosen. Optimization allowed the hyperparameter alpha to be fixed at 0.5 and beta at 0.1.

Figure 5 shows the distance matrix for a model with seven topics. The distance between any pair of topics was greater than 0.5, indicating little overlap.

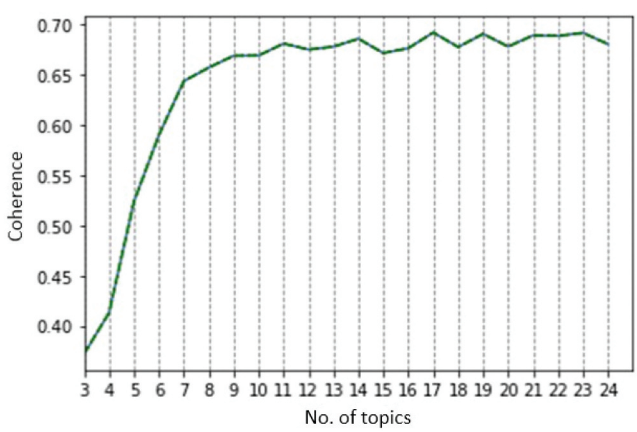

Figure 4. Coherence (C_V parameter) as a function of the number of topics in the LDA model.

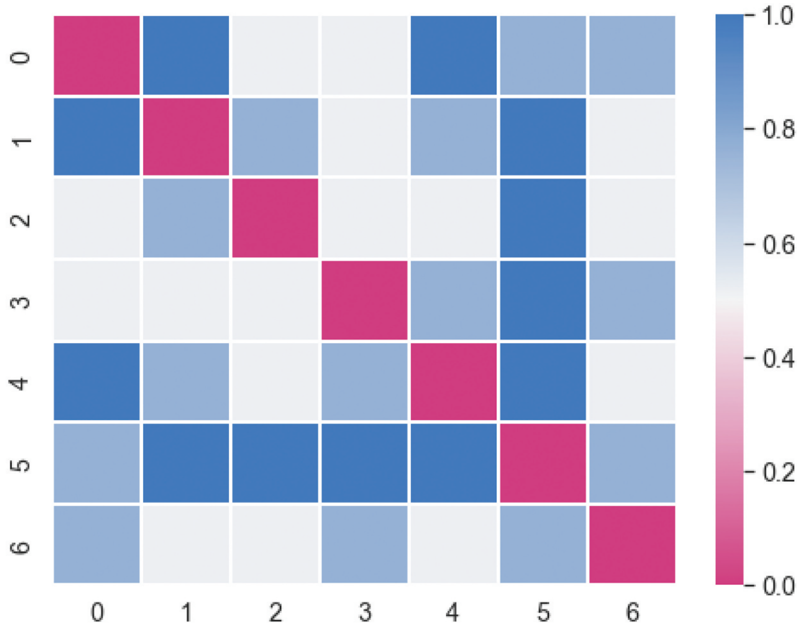

Figure 5. Distance matrix for seven topics of the LDA model.

Table 2 shows the resulting topics named after the most relevant words, with Topics 1 and 7 presenting the largest (17.8%) and smallest (11.2%) percentages of tokens, respectively.

Relationship Between Topics and Star Ranking

The results of the optimization for GBT yielded the following best hyperparameters: 500 estimators, learning_rate of 0.01, max_depth: None, max_features: auto for a rmse of 0.4803. This is less than 0.5, which is acceptable. Utilizing 20 splits, cross-validation for the best model obtained an average accuracy score of 0.71 (SD = 0.05) and an average log_loss of 0.58 (0.03) (which indicates an average assigned probability of 56%, 0.58=Ln(0.56), for the correct class).

Figure 6 shows the precision-recall curve, which is recommended for unbalanced datasets. It revealed an acceptable performance of the model.

In relation to the importance of the predictors, Table 3 shows the results for the different metrics computed. They exhibited that there is no agreement. According to the permutation, the three most relevant topics were shopping experience (0.17), labeling doubts (0.15), and safety level (0.12). The least relevant ones were certification (0.06) and nose problems (0.07). However, taking purity as a criterion, the three most relevant were face coverage (0.18), labeling doubts (0.17), and shopping experience (0.16), while the least relevant were safety level (0.09) and nose problems (0.11).

Table 2. Topics obtained from the analysis and the 10 most relevant words identified in each topic.

Topic	Words	Labelling	% Tokens
Topic 1: relationship between fit and protection (coverage)	″0.191*'quality' + 0.081*'protection' + 0.078*'face' + 0.046*'head' + ″ ″0.045*'nose' + 0.039*'size' + 0.037*'price' + 0.034*'strap' + ″ ″0.028*'tight-fitting' + 0.028*'one-size'″)	Protection_fit: face_nose_head_unique_size (tight)	17.8
Topic 2: doubts about the labeling of the mask	1, ″0.245*'kn95' + 0.107*'quality' + 0.063*'type' + 0.058*'ear' + ″ ″0.049*'trade mark' + 0.048*'face' + 0.047*'price' + 0.038*'strap' + ″ ″0.026*'china' + 0.022*'doubt'″)	Doubt_kn95_type_brand_china	16.5
Topic 3: level of safety related to the type of mask (filtering level) fit and nose clip	″0.119*'quality' + 0.102*'price' + 0.071*'nose' + 0.045*'ear' + ″ ″0.042*'face' + 0.037*'type' + 0.036*'to fit' + 0.030*'security' + ″ ″0.030*'closure' + 0.028*'kn95'″)	Safety_type_fit_clip_nose_ear_face	14.8
Topic 4: talks about the certification (standard, EC label)	″0.111*'quality' + 0.106*'type' + 0.066*'face' + 0.058*'one-size' + ″ ″0.057*'nose' + 0.053*'hours' + 0.050*'rule' + 0.042*'European Community' + 0.036*'price' ″ ″+ 0.034*'comfort'″)	Standard_EC	14.1
Topic 5: comments about the relationship between quality and price	″0.119*'price' + 0.112*'purchase' + 0.098*'quality' + 0.055*'face' + ″ ″0.051*'relationship' + 0.041*'nose' + 0.034*'closed' + 0.030*'chemistry' + ″ ″0.023*'ear' + 0.021*'trade mark'″),	Q-P ratio	13.6
Topic 6: problems in the nose due to the strap	″0.107*'nose' + 0.090*'strap' + 0.078*'fpp2' + 0.077*'quality' + ″ ″0.063*'problem' + 0.054*'type' + 0.044*'to use' + 0.039*'to adapt' + ″ ″0.036*'air' + 0.036*'price'″)	Nose_strap_fpp2_problem_adapt	11.9
Topic 7: shopping experience and the packaging	″0.091*'quality' + 0.084*'delivery' + 0.072*'nose' + 0.070*'time' + ″ ″0.060*'face' + 0.041*'air' + 0.039*'strap' + 0.039*'comfort' + ″ ″0.038*'photography' + 0.031*'olour'″)	Shopping_experinece_delivery_time_picture_smell	11.2

Shap values (Figure 7) showed doubts about labeling (0.25) to be the most relevant topic, followed by quality-price ratio (0.16) and certification (0.14). Problems in the nose due to the strap (0.06) and the level of safety related to the mask type (filtering level), the fit and the nose clip (0.09) were the least important.

In any case, only SHAP allowed an understanding of the type of influence. Figure 8 provides an understanding of the type of influence by combining the value of the feature (red for high and blue for low) and the impact on the model output (horizontal axis). In this way, a topic has a negative influence when high values in comments (red dots) are associated with negative Shap values (x axis) on the model output. In descending order of impact, Topics 7, 6, and 1 showed a similar distribution, indicating a negative influence to a lesser extent than other ones.

The force plot (Figure 9) summarized this for a given output. It obtains an actual value of 1.75 (Shap value) with respect to the base value of 1.23. In this graph, the base value is the value that would be predicted if we did not know

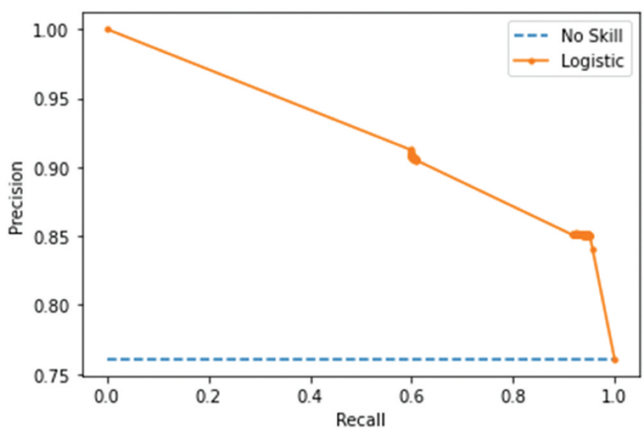

Figure 6. Precision-recall curve of the model obtained in the study. It represents the number of true positives divided by the sum of the true positives and false positives (Precision), and the number of true positives divided by the sum of the true positives and the false negatives (Recall). A no-skill classifier is represented as a horizontal line with the value of the ratio of positive cases in the dataset. A model with perfect skill is depicted as a point at (1,1). A skillful model is illustrated as a curve that bows toward (1,1) above the flat line of no skill.

Table 3. Importance of topics in the model.

Feature	Mean permutation	Node purity	Shap values
Topic 1: relationship between fit and protection (coverage)	0.09	0.18	0.11
Topic 2: doubts about the labeling of the mask	0.15	0.17	0.25
Topic 3: level of safety related to the type of mask (filtering level) fit and nose clip	0.12	0.09	0.09
Topic 4: talks about the certification (standard, EC label)	0.06	0.14	0.14
Topic 5: comments about the relationship between quality and price	0.09	0.14	0.16
Topic 6: problems in the nose due to the strap	0.07	0.11	0.06
Topic 7: shopping experience and the packaging	0.17	0.16	0.12

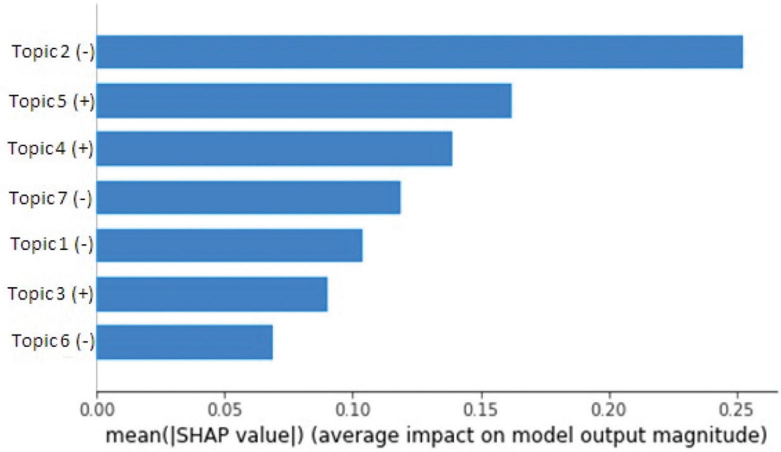

Figure 7. Shap values summary plot. (+): Topics that influence positively; (-): Topics that influence negatively. Topic 2: Doubts about the labeling of the mask; Topic 5: Quality-price ratio; Topic 4: Certification (i.e., standard, EC label); Topic 7: Shopping experience; Topic 1: Unique size coverage; Topic 3: Type of filtering and fit; Topic 6: Problems in the nose due to the strap.

any features (weight of each topic) for the case studied. It is the average of the model output over the training dataset. The numbers on the plot arrows are the value of the feature for this instance: topic 4 = 0.01349, topic 6 = 0.1198, topic 5 = 0.2495, topic 7 = 0.1198, topic 3 = 0.1168 and topic 2 = 0.1379. Red color represents features that pushed the model score higher, and blue color represents features that pushed the score lower. The bigger the arrow, the bigger the impact of the feature on the output. The additive result of these effects move the base value to the actual value of 1.75. It is possible to appreciate that topics 2, 3 and 7 contributed negatively whereas topics 4, 5 and 6 contributed positively.

Discussion

There was a massive increase in the use of face masks during the pandemic. Although there is currently no requirement to wear them, the World Health Organization recommends the use of masks after recent exposure to COVID-19, if someone has COVID-19 or is suspected of having COVID-19, if someone is at a high risk of severe COVID-19, and for anyone in a crowded, enclosed or poorly ventilated space (World Health Organization 2023). Furthermore, it is suggested that in the foreseeable future, people may need to wear masks to combat any airborne disease in the post-pandemic world (Lee et al. 2022).

In this scenario, to rapidly evolve and improve products, developing new strategies for gathering costumers´ opinion in order to implement product updates according to customer demands is relevant. The aim of this study was

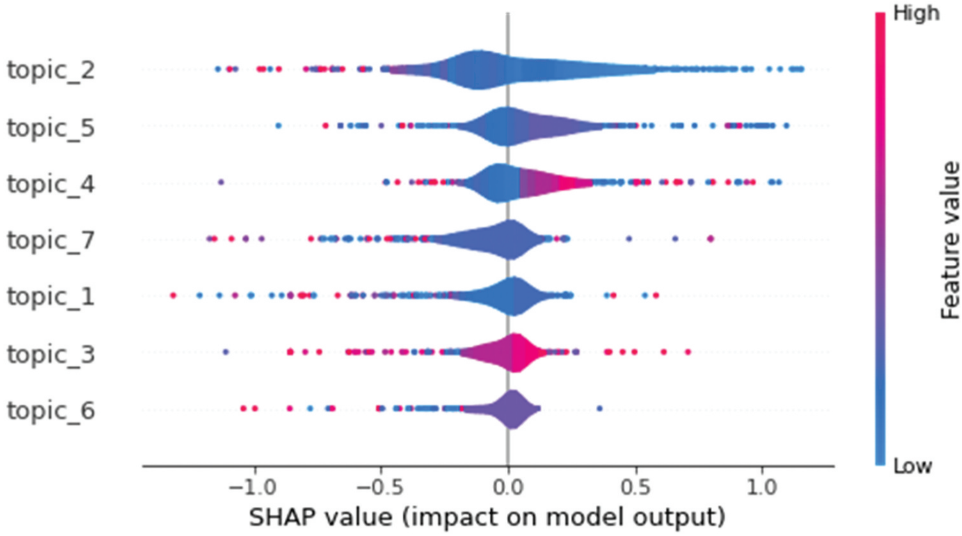

Figure 8. Influence of features in the model output. Each dot represents a row of the dataset. Each point represents a row of data from the original dataset. Topic 2: Doubts about the labeling of the mask; Topic 5: Quality-price ratio; Topic 4: Certification (i.e., standard, EC label); Topic 7: Shopping experience; Topic 1: Unique size coverage; Topic 3: Type of filtering and fit; Topic 6: Problems in the nose due to the strap.

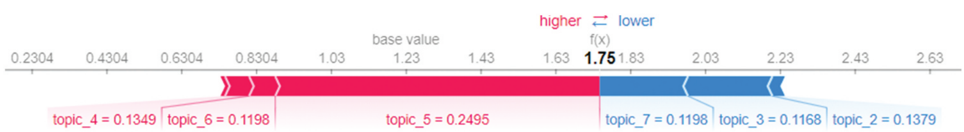

Figure 9. Force plot from SHAP showing the effect of each topic for a particular review. Topic 1 has no influence

to apply Explainable Artificial Intelligence to identify the relevance of attributes and features of face masks from online customers' reviews to guide product improvement. To this end, we captured online customers´ reviews from the Amazon website, text-mined them, identified topics using LDA, and examined the influence of topics using GBT and Shap values.

The main finding of this study is that combining LDA with SHAP to explain complex ML model can identify face mask attributes and their influence on the overall ranking. This had previously been applied for cell phone (Joung and Kim 2021) but applying DL neural network-based methods over a large dataset of reviews (>30k) and not to ML over smaller datasets as in the present research.

This contributes to addressing both RQ1, which examines whether the influence of mask attributes on overall customer ratings in online reviews can be determined by employing SHAP methods to explicate models developed with advanced black-box classification ML techniques, and RQ2, which focuses on identifying the specific mask attributes that influence customer ratings in online reviews.

In descending order of importance, the identified topics were as follows:

(1) **Topic 2**. Doubts about the labeling (e.g., type, precedence, and brand) were the most relevant and showed a negative influence. Comments with high weight of Topic 2 tended to receive less than 4 stars.
(2) **Topic 4**. Certification (e.g., standard, and EC label) was the second most relevant. Comments with high values on this topic had a positive impact on scoring more than 4 stars.
(3) **Topic 5**. Relationship between quality and price. Comments with high values on this topic had a positive impact on having more than 4 stars.
(4) **Topic 7**. Shopping experience and packaging. Comments with high values on this topic tended to have less than 4 stars.
(5) **Topic 1**. Unique-size face coverage. Comments on this topic tended to associate a negative impact with having more than 4 stars.
(6) **Topic 3**. Level of filtering and fit. Highly loaded comments on this topic had a positive impact on having more than 4 stars.
(7) **Topic 6**. Problems in the nose due to the strap. It had a minor negative impact.

To our knowledge, no published works have applied topic analysis on face masks, which makes comparison with the literature difficult. Bakhit et al. (2021), after reviewing the existing literature, revealed discomfort and irritation to be the most commonly reported problems when wearing face masks. The discomfort could be related to Topic 6 in this study, while irritation was not one of the topics identified. However, the same authors concluded that more research is needed to address the harms and disadvantages of face masks. Furthermore, the results of this study showed topics that appear in studies of online reviews of other products, such as the relationship between quality and price (Topic 5) (Jang, Chung, and Rao 2021) and those related to the shopping experience (Topic 7) (Bae and Lee 2011).

At this point, the main purpose of these results is to guide the agenda of face mask designers and manufacturers. To what extent should they focus on the identified topics?

LDA is a probabilistic model; therefore, topics may vary across runs. This is overcome by running the model several times. In addition, coherence and distance metrics facilitate obtaining an approximate structure of topics. Thus, there is a long-standing assumption that the latent space discovered by LDA models is generally meaningful and useful, and that evaluating such assumptions is challenging due to the unsupervised training process. The meaningfulness of LDA may only be validated through comparison with results from traditional methods, which (as noted in the introduction) require substantially more time, cost more, and offer smaller data samples than the methodology employed in this study, which is cheaper and faster. Furthermore, there is no gold standard list of topics that can be compared to each corpus. In any case, the list of topics reflects the most frequent words identified in the comments (quality, comfort, price, size, color, type, certification, fit and presentation of the product [parcel and package] and other elements, such as the strap).

After comparing the results with published research, it can be said that topics like the filtration level of masks support the current research agenda (Chua et al. 2020; Tcharkhtchi et al. 2021). In addition, the improvement of masks should not only focus on materials but also include other characteristics, such as the influence of the fit (Topic 3). In our study, the mask fit (O'Kelly et al. 2021) was included in Topic 1, and it related to the level of protection provided by face coverage and discomfort (Chua et al. 2020). These features also correlated with the nose fit (clip and strap) (Topic 6).

In a recent work, Rahman et al. (2022) reviewed the requirements and types of face masks included in various standards (National Institute for Occupational Safety and Health [NIOSH], American Society for Testing and Materials [ASTM], and International Organization for Standardization [ISO]). These regulations were developed to define performance requirements for source control, the protective efficacy of face masks or barrier face coverings, and standardized products to aid user selection decisions (Rahman et al. 2022). They describe the following requirements as necessary for the effective use of the masks: (1) excellent fit over the user's nose and mouth to prevent leaking, protect, and be comfortable on the face (Topics 1 and 6) and (2) use of multilayer, non-woven but breathable fabrics to ward off bright light, prevent leakage, and allow normal breathing. In this respect, compliance with a standard or the EC label had a strong and positive influence (Topic 4), but this information about the type of mask, protection level, brand, etc., should be enhanced to avoid confusion among customers, as this was the primary cause of the negative comments.

In regard to Topic 7, it highlighted the necessity of adhering to the commonly accepted approach that considers packaging design as part of product development. Packaging design offers significant information and visual cues that impact customer purchase decisions and establishes expectations, affecting subsequent user evaluations (see Shukla, Singh, and Wang 2022 for a review). When considering the shopping experience, since the current study specifically focuses on online product reviews, it is necessary to keep Topic 7 in mind. This suggests accounting for the shopping experience related to product delivery, as it can affect ratings. As a result of this research, it is recommended that this factor be included in the improvement plan for face masks.

Regarding the analyses of the influence of attributes, the model's performance was moderate. Running the model trained on all the data resulted in an accuracy of 0.83, a recall of 0.82, and a log_loss of 0.43, with a 65% probability assigned to the correct class (log_loss is one of the most commonly used metrics for classification tasks). It (e.g., cross-entropy loss) evaluates performance by comparing actual class labels and predicted probabilities. The comparison is quantified using cross-entropy loss, which is a metric used in ML to measure how well a classification model performs. The loss (or error) is measured as a number between 0 and 1, with 0 being a perfect model. In our model, the cross-entropy loss was 0.105 (close to 0), which means that the model is very good (at a 90% confidence interval). This suggests that there are aspects of online reviews other than product attributes that contribute to the overall ranking. Further research should investigate the influence of other aspects, such as real price, mask type or sentiment analysis.

The approach presented in this study may have significant implications for industry. It is essential that industrial companies comprehend consumer preferences and develop products that meet the market demands, and generate effective marketing strategies. To successfully position their products, companies must understand how consumers perceive them (Alzate, Arce-Urriza, and Cebollada 2022). In a dynamic market, they require fast and cost-efficient methods to analyze customers' needs. Examining how customers perceive product attributes through online reviews and the impact these attributes have on their purchasing decisions can be a productive method that aligns well with industry standards.

Strengths, Limitations and Future Research

The present paper is relevant in demonstrating the usefulness and feasibility of merging analyses of online reviews with XAI to establish the relationship between product attributes and ratings to accelerate innovation in products and services. This study offers the prospect of leveraging the mathematical and computational benefits of complex techniques, such as GBT, which employ a black-box approach to produce high-performing models that can be later elucidated by XAI.

Regarding this, GBT metrics (permutation and node impurity) produced divergent outcomes and evaluated the importance of features in the model's performance rather than the outcome variable (like SHAP did).

Joung and Kim (2021) have made a similar contribution in which LDA was combined with ABSA to estimate the importance of product attributes. They evaluated the impact of sentiments associated with each product attribute on overall ratings by using an explainable, deep neural network that utilizes Shap values. The study showed that negative comments led to lower ratings, while positive comments corresponded to higher rankings. However, a comment may have one polarity despite containing contributions on various. Negative reviews potentially include positive comments related to certain topics and vice versa. This presents a compelling avenue for future research, where ABSA and ML techniques can be combined (Trisna and Jie 2022) to explore the mediating impact of polarity on the product attribute- star rating relationship.

Another point for discussion is the approach used for text representation. We used a bag of words (BOW) for simplicity. However, it should be noted that neural networks-based embedding techniques, such as Word2Vec or Bert (Devlin et al. 2019; Suryadi and Kim 2018) when combined with ML and DL techniques, demonstrated better performance than BOW (representing a text only the words it contents), although the latter is much simpler to apply.

In this line of thought, permutation importance is fast to compute, and widely used and understood compared to most other approaches. The method randomly varies the validation data of a feature and compares new predictions with those of the fitted model. New results would be worse depending on how relevant the feature is. They can only get better if a feature is not relevant at all (negative values of the permutation) because if it is relevant, it contributes to the original model. However, it does not explain how important each feature is. If a feature has a medium permutation importance, it potentially has a large effect for a few predictions, no effect in general, or a medium effect for all predictions. Therefore, we used SHAP.

The main difference between SHAP and other popular methods of XAI, such as LIME, is that the latter does not guarantee that the prediction is fairly distributed across the features. Moreover, SHAP does return a prediction model like LIME. It always considers all features and thus cannot be used to make statements about changes in the predictions (ratings) for changes in the input of a single topic.

Another point to discuss regarding the study's methods is the fact that there were more general ratings above 4 stars than below. We assumed that satisfied customers typically leave ratings of 4 or 5 stars, while dissatisfied customers leave ratings of 3 stars or less. Then, the presence of more comments over 3 stars could be due to dissatisfied customers not leaving comments. However, it could also be attributed to the existence of more highly rated products compared to bad ones, which is a well-documented phenomenon known as the "positivity problem." This refers to the fact that there is a tendency toward 4 and 5 star ratings in online reviews (Rocklage, Rucker, and Nordgren 2021).

Several studies present some limitations of online reviews in terms of reflecting users' general opinions about a product. An outstanding field of research is the verification of online reviews' authenticity. Fake reviews are increasingly affecting data reliability (Lim and Tucker 2017; Salminen et al. 2022; Suryadi and Kim 2018). Some websites provide fake comments detection, but these systems are costly since combining automated and human detection strategies still fails to detect a significant portion of bogus reviews, presenting an ongoing challenge (Mohawesh et al. 2021; Salminen et al. 2022).

In addition, intrinsic biases in individuals, such as what they tend to exaggerate (Kapoor et al. 2021) and the realistic, optimistic, pessimistic, or unreliable nature of reviewers (Lim and Tucker 2017), present compelling and noteworthy avenues of research.

In this case, the results should be interpreted with caution when designing new products. Another possibility is that because people use online reviews to decide whether to buy masks, those that receive negative comments are not bought and disappear from the market. Therefore, there are more positive than negative reviews. As we have previously stated, this is the well-documented phenomenon called "positivity problem" (Rocklage, Rucker, and Nordgren 2021).

This study used only Spanish-language reviews to control for potential cultural and linguistic influences on results, as noted by some authors (e.g., Kong and Lou 2023). Future research should examine how differences in language, culture, and other factors affect product ratings, as well as the relationship between customers' valued attributes and their ratings.

Furthermore, this study was conducted when face masks were mandatory. Now that this requirement has been lifted, future research into differences between consumers' reviews during and after the pandemic is warranted.

The findings of this study indicate the potential benefit of utilizing text-mining and natural language processing techniques in online reviews for expediting the product improvement process. Additionally, utilizing Explainable Artificial Intelligence (XAI) methods, such as SHAP, can lead to the extraction of interpretable models. The field of XAI is currently a prominent area of research in the arena of Artificial Intelligence. The outcomes of this study suggest new channels for future research involving the application of XAI in product development and emotional engineering. In particular, there is a growing interest in developing methods for verifying the authenticity of online reviews.

Conclusion

This paper's results imply that analyzing online customer reviews could be useful to speed up the development and evolution of any product or service. The combination of Topic Modeling and Explainable Artificial Intelligence (XAI) allowed to identify and establish the relevance of attributes that people value when buying online and using face masks. Accordingly, the innovation agenda for this product should focus on the following: avoiding customer confusion about mask type (ffp2 or kn95), brand, and origin (Topic 2); compliance with standards and/or EC labeling (Topic 4); balanced quality-price ratio (Topic 5); mask protection (face and nose fit, unique size, and strap) (Topic 1); shopping experience (packaging and smell) (Topic 3); and avoidance of the nose adaptation problems (clip and strap) (Topic 6). Analyzing how customers perceive product features through online reviews and the influence of these attributes on their purchasing decisions can serve as an efficient approach in guiding product improvement in the industry.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ORCID

Inmaculada Aparicio-Aparicio http://orcid.org/0000-0003-3784-8489
Maria-Arantzazu Ruescas-Nicolau http://orcid.org/0000-0001-8875-068X
Lirios Dueñas http://orcid.org/0000-0001-8592-1738
José Luis Sánchez-Jiménez http://orcid.org/0000-0002-5055-1789
M.Luz Sánchez-Sánchez http://orcid.org/0000-0001-5630-8002

Data Availability Statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

Al-Rammahi, A. H. I. 2022. Face mask recognition system using MobileNetV2 with optimization function. *Applied Artificial Intelligence* 36 (1):2145638. doi:10.1080/08839514.2022.2145638.

Alzate, M., M. Arce-Urriza, and J. Cebollada. 2022. Mining the text of online consumer reviews to analyze brand image and brand positioning. *Journal of Retailing & Consumer Services* 67:102989. doi:10.1016/j.jretconser.2022.102989.

Bae, S., and T. Lee. 2011. Gender differences in consumers' perception of online consumer reviews. *Journal of Electronic Commerce Research* 11 (2):201–14. doi:10.1007/s10660-010-9072-y.

Bakhit, M., N. Krzyzaniak, A. M. Scott, J. Clark, P. Glasziou, and C. Del Mar. 2021. Downsides of face masks and possible mitigation strategies: A systematic review and meta-analysis. *BMJ Open* 11 (2):e044364. doi:10.1136/bmjopen-2020-044364.

Bentéjac, C., A. Csörgő, and G. Martínez-Muñoz. 2021. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review* 54 (3):1937–67. doi: 10.1007/s10462-020-09896-5.

Berezina, K., A. Bilgihan, C. Cobanoglu, and F. Okumus. 2016. Understanding satisfied and dissatisfied hotel customers: Text mining of online hotel reviews. *Journal of Hospitality Marketing and Management* 25 (1):1–24. doi:10.1080/19368623.2015.983631.

Chen, H., J. M. Samet, H. Tong, A. Abzhanova, A. G. Rappold, and S. E. Prince. 2022. Can disposable masks be worn more than once? *Ecotoxicology & Environmental Safety* 242:113908. doi:10.1016/j.ecoenv.2022.113908.

Chua, M. H., W. Cheng, S. S. Goh, J. Kong, B. Li, J. Y. C. Lim, L. Mao, S. Wang, K. Xue, L. Yang, et al. 2020. Face masks in the new COVID-19 normal: Materials, testing, and perspectives. *Research (Wash D C)* 2020:7286735. doi: 10.34133/2020/7286735.

Dash, A., D. Zhang, and L. Zhou. 2021. Personalized ranking of online reviews based on consumer preferences in product features. *International Journal of Electronic Commerce* 25 (1):29–50. doi:10.1080/10864415.2021.1846852.

Deshpande, G., and J. Rokne. 2018. User feedback from tweets vs app store reviews: An exploratory study of frequency, timing and content. 2018 5th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE), 15–21, Banff, AB: IEEE. doi: 10.1109/AIRE.2018.00008.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–86, Minneapolis, MN: Association of Computational Linguistics. doi: 10.48550/ARXIV.1810.04805.

Dugdale, C. M., and R. P. Walensky. 2020. Filtration efficiency, effectiveness, and availability of N95 face masks for COVID-19 prevention. *JAMA Internal Medicine* 180 (12):1612–13. doi:10.1001/jamainternmed.2020.4218.

Fernandes, S., R. Panda, V. G. Venkatesh, B. N. Swar, and Y. Shi. 2022. Measuring the impact of online reviews on consumer purchase decisions – a scale development study. *Journal of Retailing & Consumer Services* 68:103066. doi:10.1016/j.jretconser.2022.103066.

Fernando, A. G., and E. C.-X. Aw. 2023. What do consumers want? A methodological framework to identify determinant product attributes from consumers' online questions. *Journal of Retailing & Consumer Services* 73:103335. doi: 10.1016/j.jretconser.2023.103335.

Hastie, T., R. Tibshirani, and J. H. Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. New York, NY: Springer.

Howard, J., A. Huang, Z. Li, Z. Tufekci, V. Zdimal, H.-M. van der Westhuizen, A. von Delft, A. Price, L. Fridman, L.-H. Tang, et al. 2021. An evidence review of face masks against COVID-19. *Proceedings of the National Academy of Sciences USA* 118 (4):e2014564118. doi:10.1073/pnas.2014564118.

Imtiaz, M., and M. Ben Islam. 2020. Identifying significance of product features on customer satisfaction recognizing public sentiment polarity: Analysis of smart phone industry using machine-learning approaches. *Applied Artificial Intelligence* 34 (11):832–48. doi: 10.1080/08839514.2020.1787676.

Jang, S., J. Chung, and V. R. Rao. 2021. The importance of functional and emotional content in online consumer reviews for product sales: Evidence from the mobile gaming market. *Journal of Business Research* 130:583–93. doi:10.1016/j.jbusres.2019.09.027.

Joung, J., and H. M. Kim. 2021. Approach for importance–performance analysis of product attributes from online reviews. *Journal of Mechanical Design* 143 (8):081705. doi:10.1115/1.4049865.

Kapadia, S. 2020. *Evaluate topic models: Latent Dirichlet Allocation (LDA)*. Medium. https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0.

Kapoor, P. S., M. S. B, M. Maity, and N. K. Jain. 2021. Why consumers exaggerate in online reviews? Moral disengagement and dark personality traits. *Journal of Retailing & Consumer Services* 60:102496. doi: 10.1016/j.jretconser.2021.102496.

Kong, J., and C. Lou. 2023. Do cultural orientations moderate the effect of online review features on review helpfulness? A case study of online movie reviews. *Journal of Retailing & Consumer Services* 73:103374. doi:10.1016/j.jretconser.2023.103374.

Kumar, P., Y. K. Dwivedi, and A. Anand. 2021. Responsible artificial intelligence (AI) for value formation and market performance in healthcare: The mediating role of patient's cognitive engagement. *Information Systems Frontiers* 25 (6):2197–220. doi: 10.1007/s10796-021-10136-6.

Lee, P., H. Kim, Y. Kim, W. Choi, M. S. Zitouni, A. Khandoker, H. F. Jelinek, L. Hadjileontiadis, U. Lee, and Y. Jeong. 2022. Beyond pathogen filtration: Possibility of smart masks as wearable devices for personal and group health and safety management. *JMIR mHealth and uHealth* 10 (6):e38614. doi:10.2196/38614.

Li, X., Q. Li, and J. Kim. 2023. A review helpfulness modelling mechanism for online e-commerce: Multi-channel CNN end-to-end approach. *Applied Artificial Intelligence* 37 (1):2166226. doi: 10.1080/08839514.2023.2166226.

Li, Y., M. Liang, L. Gao, M. Ayaz Ahmed, J. P. Uy, C. Cheng, Q. Zhou, and C. Sun. 2021. Face masks to prevent transmission of COVID-19: A systematic review and meta-analysis. *American Journal of Infection Control* 49 (7):900–06. doi:10.1016/j.ajic.2020.12.007.

Li, Z., Z. G. Tian, J. W. Wang, and W. M. Wang. 2020. Extraction of affective responses from customer reviews: An opinion mining and machine learning approach. *International Journal of Computer Integrated Manufacturing* 33 (7):670–85. doi:10.1080/0951192X.2019.1571240 .

Lim, S., and C. S. Tucker. 2017. Mitigating online product rating biases through the discovery of optimistic, pessimistic, and realistic reviewers. *Journal of Mechanical Design* 139 (11):111409. doi:10.1115/1.4037612.

Lundberg, S. M., and S.-I. Lee. 2017. A unified approach to interpreting model predictions. In Advances in neural information processing systems, vol. 30. [place unknown]: Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Lundberg, S. M., B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D.-W. Low, S.-F. Newman, J. Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2 (10):749–60. doi: 10.1038/s41551-018-0304-0.

Mohawesh, R., S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood. 2021. Fake reviews detection: A survey. *Institute of Electrical and Electronics Engineers Access* 9:65771–802. doi:10.1109/ACCESS.2021.3075573.

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences USA* 116 (44):22071–80. doi:10.1073/pnas.1900654116.

Nasiri, M. S., and S. Shokouhyar. 2021. Actual consumers' response to purchase refurbished smartphones: Exploring perceived value from product reviews in online retailing. *Journal of Retailing & Consumer Services* 62:102652. doi:10.1016/j.jretconser.2021.102652.

Neumann, Ł., R. M. Nowak, R. Okuniewski, and P. Wawrzyński. 2019. Machine learning-based predictions of customers' decisions in car insurance. *Applied Artificial Intelligence* 33 (9):817–28. doi:10.1080/08839514.2019.1630151.

Noori, B. 2021. Classification of customer reviews using machine learning algorithms. *Applied Artificial Intelligence* 35 (8):567–88. doi:10.1080/08839514.2021.1922843.

O'Kelly, E., A. Arora, S. Pirog, J. Ward, P. J. Clarkson, and A. Mukherjee. 2021. Comparing the fit of N95, KN95, surgical, and cloth face masks and assessing the accuracy of fit checking. *PLOS ONE* 16 (1):e0245688. doi:10.1371/journal.pone.0245688.

Qiao, Z., X. Zhang, M. Zhou, G. A. Wang, and W. Fan. 2017. A domain oriented LDA model for mining product defects from online customer reviews. Hawaii International Conference on System Sciences 2017, HICSS–50) [place unknown]. https://aisel.aisnet.org/hicss-50/dsm/data_mining/2.

Rahman, M. Z., M. E. Hoque, M. R. Alam, M. A. Rouf, S. I. Khan, H. Xu, and S. Ramakrishna. 2022. Face masks to combat coronavirus (COVID-19)—processing, roles, requirements, efficacy, risk and sustainability. *Polymers* 14 (7):1296. doi:10.3390/polym14071296.

Rathore, A. K., and P. V. Ilavarasan. 2020. Pre- and post-launch emotions in new product development: Insights from twitter analytics of three products. *International Journal of Information Management* 50:111–27. doi:10.1016/j.ijinfomgt.2019.05.015.

Rocklage, M. D., D. D. Rucker, and L. F. Nordgren. 2021. Mass-scale emotionality reveals human behaviour and marketplace success. *Nature Human Behaviour* 5 (10):1323–29. doi:10.1038/s41562-021-01098-5.

Saito, T., M. Rehmsmeier, and G. Brock. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE* 10 (3):e0118432. doi:10.1371/journal.pone.0118432.

Salminen, J., C. Kandpal, A. M. Kamel, S. Jung, and B. J. Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing & Consumer Services* 64:102771. doi:10.1016/j.jretconser.2021.102771.

Scheid, J. L., S. P. Lupien, G. S. Ford, and S. L. West. 2020. Commentary: Physiological and psychological impact of face mask usage during the COVID-19 pandemic. *International Journal of Environmental Research and Public Health* 17 (18):E6655. doi:10.3390/ijerph17186655.

Shack, A. R., L. Arkush, S. Reingold, and G. Weiser. 2020. Masked paediatricians during the COVID -19 pandemic and communication with children. *Journal of Paediatrics and Child Health* 56 (9):1475–76. doi:10.1111/jpc.15087.

Sharma, A., and M. O. Shafiq. 2022. A comprehensive artificial intelligence based user intention assessment model from online reviews and social media. *Applied Artificial Intelligence* 36 (1):2014193. doi:10.1080/08839514.2021.2014193.

Shirdastian, H., M. Laroche, and M.-O. Richard. 2019. Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. *International Journal of Information Management* 48:291–307. doi:10.1016/j.ijinfomgt.2017.09.007.

Shukla, P., J. Singh, and W. Wang. 2022. The influence of creative packaging design on customer motivation to process and purchase decisions. *Journal of Business Research* 147:338–47. doi:10.1016/j.jbusres.2022.04.026.

So, C. 2020. Understanding the prediction mechanism of sentiments by XAI visualization. Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval, Seoul, Republic of Korea: Association for Computing Machinery, 75–80. doi: 10.1145/3443279.3443284.

Son, Y., and W. Kim. 2023. Development of methodology for classification of user experience (UX) in online customer review. *Journal of Retailing & Consumer Services* 71:103210. doi:10.1016/j.jretconser.2022.103210.

Sun, H., W. Guo, H. Shao, and B. Rong. 2020. Dynamical mining of ever-changing user requirements: A product design and improvement perspective. *Advanced Engineering Informatics* 46:101174. doi:10.1016/j.aei.2020.101174.

Suryadi, D., and H. Kim. 2018. A systematic methodology based on word embedding for identifying the relation between online customer reviews and sales rank. *Journal of Mechanical Design* 140 (12). doi: 10.1115/1.4040913.

Suryadi, D., and H. M. Kim. 2019. A data-driven methodology to construct customer choice sets using online data and customer reviews. *Journal of Mechanical Design* 141 (11):111103. doi:10.1115/1.4044198.

Tasic-Kostov, M., M. Martinović, D. Ilic, and M. Cvetkovic. 2022. Cotton versus medical face mask influence on skin characteristics during COVID-19 pandemic: A short-term study. *Skin Research & Technology* 28 (1):66–70. doi:10.1111/srt.13091.

Tcharkhtchi, A., N. Abbasnezhad, M. Zarbini Seydani, N. Zirak, S. Farzaneh, and M. Shirinbayan. 2021. An overview of filtration efficiency through the masks: Mechanisms of the aerosols penetration. *Bioactive Materials* 6 (1):106–22. doi:10.1016/j.bioactmat.2020.08.002.

Trappey, A. J. C., C. V. Trappey, A.-C. Chang, and L. W. L. Chen. 2016. Using web mining and perceptual mapping to support customer-oriented product positions designs. Advances in Transdisciplinary Engineering- Proceedings of the 23rd ISPE Inc. International Conference on Transdisciplinary Engineering, ed. M. Borsato, et al. 533–42, Amterdam, The Netherlands: IOS Press. doi: 10.3233/978-1-61499-703-0-533.

Trisna, K. W., and H. J. Jie. 2022. Deep learning approach for aspect-based sentiment classification: A comparative review. *Applied Artificial Intelligence* 36 (1):2014186. doi:10.1080/08839514.2021.2014186.

Tuarob, S., and C. S. Tucker. 2015. Quantifying product favorability and extracting notable product features using large scale social media data. *Journal of Computing and Information Science in Engineering* 15 (3). doi: 10.1115/1.4029562.

Varanges, V., B. Caglar, Y. Lebaupin, T. Batt, W. He, J. Wang, R. M. Rossi, G. Richner, J.-R. Delaloye, and V. Michaud. 2022. On the durability of surgical masks after simulated handling and wear. *Scientific Reports* 12 (1):4938. doi:10.1038/s41598-022-09068-1.

World Health Organization. 2023. WHO updates COVID-19 guidelines on masks, treatments and patient care. https://www.who.int/news/item/13-01-2023-who-updates-covid-19-guidelines-on-masks–treatments-and-patient-care.

Xu, X. 2019. Examining the relevance of online customer textual reviews on hotels' product and service attributes. *Journal of Hospitality & Tourism Research* 43 (1):141–63. doi:10.1177/1096348018764573.

Xu, X., and Y. Li. 2016. The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. *International Journal of Hospitality Management* 55:57–69. doi:10.1016/j.ijhm.2016.03.003.

Yang, B., Y. Liu, Y. Liang, and M. Tang. 2019. Exploiting user experience from online customer reviews for product design. *International Journal of Information Management* 46:173–86. doi:10.1016/j.ijinfomgt.2018.12.006.

Zhang, M., B. Fan, N. Zhang, W. Wang, and W. Fan. 2021. Mining product innovation ideas from online reviews. *Information Processing & Management* 58 (1):102389. doi:10.1016/j.ipm.2020.102389.