Impact of Data Types on Machine Learning Performance in LLM-Assisted Content Analysis (LACA) of Hospital Online Reviews

Muhammad Hafiz Sulaiman, Nora Muda, Fatimah Abdul Razak

Submitted to: JMIR Formative Research
on: March 15, 2025

Table of Contents

Impact of Data Types on Machine Learning Performance in LLM-Assisted Content Analysis (LACA) of Hospital Online Reviews

Muhammad Hafiz Sulaiman[1] MD, MS; Nora Muda[1] PhD; Fatimah Abdul Razak[1] PhD

[1]Department of Mathematical Sciences Faculty of Science and Technology National University of Malaysia Bangi MY

Corresponding Author:
Muhammad Hafiz Sulaiman MD, MS
Department of Mathematical Sciences
Faculty of Science and Technology
National University of Malaysia
UKM
Bangi
MY

Abstract

Background: LLM-Assisted Content Analysis (LACA) can be utilized to conduct thematic analysis on online reviews. Traditional thematic analysis involves binary thematic coding (0 = the thematic code is not present; 1 = the thematic code is present).

Objective: This paper was intended to investigate the effect of using scales as LACA data type on the ML performance parameters.

Methods: LACA was carried out using a set of thematic codes to produce an independent variable dataset (x) using scale: 0 = not an issue; 1 = a small issue; 2 = a moderate issue; 3 = a serious issue; and 4 = an extremely serious issue. Subsequently x (independent variables) and y (rating) were split to train and test machine learning models. The performances of the models trained with binary and scalar data were compared.

Results: For binary x data type, 8 themes were determined with cumulative explained variance (CEV) of 0.54 and an average Cronbach's alpha of 0.73. LR showed significance on 5 themes with pseudo-R2 of 0.41. ML classification was able to produce f1-score of 0.83 - 0.92 and accuracy of 0.90 for LR and f1-score of 0.74 - 0.88 and accuracy of 0.84 for ANN model. For scalar x data type, 6 themes were determined with CEV of 0.74 and average Cronbach's alpha of 0.86. LR shows significance on 5 themes with pseudo-R2 of 0.55. ML classification was able to produce f1-score of 0.88 - 0.94 and accuracy of 0.92 for LR model; and f1-score of 0.87 - 0.94 and accuracy of 0.92 for ANN model

Conclusions: This study highlights that scalar data encoding in LLM-Assisted Content Analysis (LACA) improves machine learning model performance compared to binary encoding. Scalar data resulted in higher explained variance, Cronbach's alpha, and classification accuracy, suggesting it is a more effective approach for thematic analysis of online reviews. This method can enhance predictive model performance in complex data analysis tasks.

(JMIR Preprints 15/03/2025:73984)
DOI: https://doi.org/10.2196/preprints.73984

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Original Paper

Dr. Muhammad Hafiz Sulaiman, MD, MS[1,2]
Associate Prof. Dr. Nora Muda, PhD[1]
Associate Prof. Dr. Fatimah Abdul Razak, PhD[1]

1. Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
2. Hospital Sultan Idris Shah, Serdang, Selangor, Malaysia

Impact of Data Types on Machine Learning Performance in LLM-Assisted Content Analysis (LACA) of Hospital Online Reviews

Abstract

Background: LLM-Assisted Content Analysis (LACA) can be utilized to conduct thematic analysis on online reviews. Traditional thematic analysis involves binary thematic coding (0 = the thematic code is not present; 1 = the thematic code is present). **Objectives:** This paper was intended to investigate the effect of using scales as LACA data type on the ML performance parameters. **Method:** LACA was carried out using a set of thematic codes to produce an independent variable dataset (x) using scale: 0 = not an issue; 1 = a small issue; 2 = a moderate issue; 3 = a serious issue; and 4 = an extremely serious issue. Subsequently x (independent variables) and y (rating) were split to train and test machine learning models. The performances of the models trained with binary and scalar data were compared. **Results:** For binary x data type, 8 themes were determined with cumulative explained variance (CEV) of 0.54 and an average Cronbach's alpha of 0.73. LR showed significance on 5 themes with pseudo-R^2 of 0.41. ML classification was able to produce f1-score of 0.83 - 0.92 and accuracy of 0.90 for LR and f1-score of 0.74 - 0.88 and accuracy of 0.84 for ANN model. For scalar x data type, 6 themes were determined with CEV of 0.74 and average Cronbach's alpha of 0.86. LR shows significance on 5 themes with pseudo-R^2 of 0.55. ML classification was able to produce f1-score of 0.88 - 0.94 and accuracy of 0.92 for LR model; and f1-score of 0.87 - 0.94 and accuracy of 0.92 for ANN model **Conclusion**: This study highlights that scalar data encoding in LLM-Assisted Content Analysis (LACA) improves machine learning model performance compared to binary encoding. Scalar data resulted in higher explained variance, Cronbach's alpha, and classification accuracy, suggesting it is a more effective approach for thematic analysis of online reviews. This method can enhance predictive model performance in complex data analysis tasks.

Keywords: Large Language Model; Hospital Quality; Patient Satisfaction; Big Data; Online Review, Machine Learning, Prediction.

Introduction

Quality improvement activities in hospitals rely heavily on feedback from patients and families to enhance care and ensure satisfaction, which is essential for patient retention and the financial viability of healthcare providers [1][2]. Traditional feedback methods, such as SERVQUAL surveys, have been used in Malaysian healthcare [3][4][5] but these can be time-consuming and limited in scope, leading to biased results due to small sample sizes and privacy concerns [6][7][8].

In contrast, online reviews provide a spatially unrestricted platform for patients to share their experiences at any stage of their hospital journey, offering a larger and more diverse set of feedback for analysis. [9] noted that online reviews cover a broader range of domains than traditional surveys like HCAHPS, which have fixed questions that may not capture evolving patient needs. [10] emphasized the need for healthcare organizations to adapt to the fourth industrial revolution by leveraging online reviews to gain insights into patient desires and values, recommending the use of supervised machine learning to categorize these reviews into established SERVQUAL domains, thereby improving quality monitoring in hospitals.

LLM-Assisted Content Analysis

Thematic analysis is a qualitative research method that identifies, analyzes, and reports patterns or themes within data. Originally, researchers begin by immersing themselves in the data, reading and reviewing it multiple times to gain a deep understanding. They then generate initial codes by tagging relevant sections with short labels that capture key aspects, which are organized into potential themes that reflect significant features related to the research question. After reviewing and refining these themes to ensure they accurately represent the data and fit cohesively, each theme is clearly defined and named. The final step involves writing up the findings, providing a comprehensive interpretation that weaves together the themes for insightful conclusions. Valued for its flexibility, thematic analysis effectively uncovers patterns within complex qualitative data, with much of the methodology in this study based on [11].

Original methods for analyzing text data from surveys and reviews face significant challenges, including time-consuming manual processes and resource-intensive efforts [12]. The vast volume and unstructured nature of textual data available online further complicate the extraction of actionable insights using conventional methodologies, prompting a shift towards automated content analysis [13]. The emergence of Large Language Models (LLMs), such as the gpt-4o-mini model, provides healthcare institutions with powerful tools to navigate and extract valuable insights from this data.

Text mining in big data analytics is gaining recognition for its ability to harness unstructured textual data, revealing significant patterns and correlations [14]. LLMs are sophisticated AI models trained on extensive text corpora, enabling them to understand and generate human-like language with remarkable fluency [15]. Their architecture, which incorporates attention mechanisms and feed-forward neural networks, enhances functionality significantly [16]. With advanced natural language processing techniques, LLMs excel in text summarization, thematic analysis (Xiao et a[17], and sentiment analysis [18], making them well-suited for qualitative data analysis in healthcare.

Recent studies have explored the impact of LLMs on drug discovery [19], medical note extraction [20], diagnosis-related group prediction [21], and diagnostics [22], highlighting opportunities for LLMs in clinical decision making systems. In qualitative researches, GPT's context-awareness also allows it to produce direct textual descriptions, making it advantageous over Latent Dirichlet Allocation topic classification [23]. The advancement of LLMs aligns with global trends toward AI in healthcare, as noted by the World Health Organization [24], which emphasized the importance of AI and big data for improved healthcare delivery. Similarly, the Ministry of Health Malaysia prioritizes digitalization and AI in the country's health reform agenda [25].

LLM-Assisted Content Analysis (LACA), refers to the use of Large Language Models (LLMs) to enhance qualitative content analysis. Researchers input qualitative data, such as text from online reviews, interviews or other documents, into an LLM, which processes and summarizes the content to provide an initial understanding. The LLM assists in coding and categorizing the text by suggesting themes and patterns through advanced natural language processing, thereby facilitating the identification and organization of key topics and underlying themes more efficiently [26]. While discussion about scalar data type was not widely available elsewhere, Awais has demonstrated how to prompt GPT to produce a scalar data type and convert it into a binary data type. His study has shown the advantage of using scalar data type - it enables the researcher to optimize thematic analysis outcome (measured by human - LLM coders agreement) by manipulating threshold for scalar to binary data type conversion.

Our research questions are: (1) What are the outcome profiles of different data types to LLM-assisted content analysis (LACA)? (2) What are the outcome profiles of different ML algorithms to LACA? The purposes of this study are (1) to compare the construct validity and internal consistency of the themes found in LACA using different data types (binary versus ordinal), (2) to determine the precision, recall f-score and accuracy of our predictive models using different data types, (3) to determine the precision, recall, f-score and accuracy of predictive models using different ML algorithms. For these purposes, we developed three hypotheses: (1) the construct validity and internal consistency of the themes found in LACA using ordinal data type shows higher cumulative explained variance for eigenvalues > 1 and higher Cronbach's alpha as compared to binary data type, (2) the precision, recall f-score and accuracy of our predictive models using ordinal data type are higher as compared to our predictive models using binary data, (3) the precision, recall, f-score and accuracy of predictive models using Artificial Neural Network is higher than predictive models using logistic regression.

Methodology

Fake review exclusion

We use the data from online reviews that have hospital quality issues. The data consists of 1,279 data points in text formats and already been filtered to remove fake reviews by using an ML filtering algorithm developed based on studies by [27][28][29][30]. The filtering algorithm involves NLP pre-processing, transforming into TF-IDF vector and training SVM model to classify fake versus real reviews using yelp.com data. Our data also excluded positive comments since [31] showed that bad reviews are the most meaningful to help readers make decisions about a hospital and to limit numbers of variables.

GPT and Human Coding

A random 200 reviews were selected. gpt-4o-mini model API was tasked to use our codebook as a guide to code the 200 reviews. A human researcher was also tasked to code the same 200 reviews. Both GPT and human coders were instructed to code each review. There are two methods in doing this: (1) Binary method: by iterating each item in the codebook and answering 1 if the item present in the review and 0 otherwise (refer **Appendix 1**); (2) Scale method: by iterating each item in the codebook and scale the level of seriousness of the issues inside each online review. Scale 0 = no issue, scale 1 = small issue; scale 2 = moderate issue, scale 3 = serious issue; and scale 4 - extremely serious issue (refer **Appendix 2**). Since there are n_i items or codes in the codebook, the number of API calls will be $200n_i$. The human coder was only required to do the scale method and the data was converted to binary for evaluating inter-rater reliability between human coder and GPT coder across the 200 reviews. Cohen's Kappa, a statistical measure that accounts for chance agreement, is used for this purpose. The process of coding (described above) can then be continued solely by gpt-4o-mini model to the rest of our data.

Training and Testing ML algorithms

Transforming *y* Variable from Scalar to Binary

Original *y* variable is online rating (with lowest rating = 1; and highest rating = 5). The *y* variable was non-normally distributed i.e. 90% of the *y* is distributed in rating 1 and rating 5. This extreme bias in rating was explained by Roh et al. (2021). The polarity suggests that it is wise for such data to be dichotomized to simplify analysis and interpretation. For such a reason, *y* variable was transformed by grouping rating 1 and 2 as low rating and rating 4 and 5 as high rating. Rating 3 and its *x* rows were filtered out from the dataset since the total number is low and the rating is deemed neutral, thus removing it doesn't affect predictability of the outcome.

Splitting Data into Train and Test Datasets

The next phase of our research focuses on preparing the dataset for effective training and testing of our machine learning models. Initially, we split the data into training and testing sets to ensure a robust evaluation of the model's performance. This process involved stratified sampling, which maintains the distribution of classes within both sets. This is essential for achieving reliable results, particularly given the potential for class imbalance.

Balancing Training Data

After the data split, we will address any imbalance in the training dataset to prevent bias toward the majority class. Balancing the training data is crucial for ensuring that the model can accurately identify and classify issues in hospital reviews. Techniques such as oversampling the minority class or undersampling the majority class, as well as synthetic data generation methods, will be employed to create a more equitable training environment.

ML Model Fitting

Once the data is balanced, we will proceed to fit our selected machine learning models. This phase involves tuning hyperparameters and applying cross-validation to optimize model

performance and ensure that it generalizes well to unseen data.

ML Model Performance Measurement

Finally, we will measure the performance of the models using confusion matrix and related measurement such as precision, recall, F1-score, and accuracy [32]. These measurements will provide insights into how well our models can classify review, particularly in identifying negative sentiment related to hospital quality. This comprehensive evaluation will not only validate the effectiveness of our approach but also inform any necessary adjustments to enhance model performance further.

Explaining Factors Influencing Rating

Themes Formation by Factor Analysis

To make the mechanisms influencing rating more explainable, variables were reduced into a single-digit number using factor analysis. The use of factor analysis to reduce the number of thematic codes into factors (themes) is documented by [33]. We do a normal factor analysis for our scale data and tetrachoric correlation factor analysis for our binary data as suggested by [34][35]. We decided to include / exclude factors based on cumulative explained variance, Cronbach alpha (George & Mallery, 2003)[36] and qualitative assessment - content validity [37]

Calculating Factor Score

Both binary and scalar *x* dataset underwent transformation into factor score using the transformer that has been trained using the train set before. The transformer calculates factor scores based on factor loadings of the 41 variables to reduce it to single-digit latent factors. Despite the factor scores not being directly interpretable, the factors act as secondary variables that can explain the relationship between these latent factors and rating.

Conducting Statistical Analysis

The factor scores were then fit into the statsmodel logit regression python module to determine which factors influencing online rating and the magnitude of their influence. Statistical models were then developed to summarize the mechanisms of determining online rating based on factors. Robustness (in terms of pseudo-R^2) of the models developed using binary and scalar x were compared.

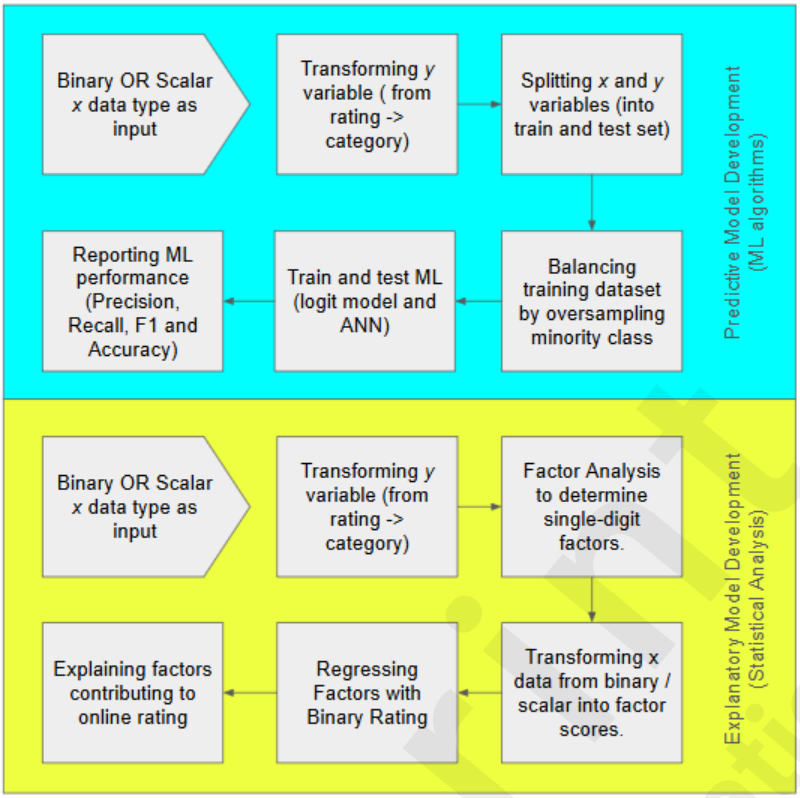


Figure 1. Flow of study methodology

Overall, the methodology involves a systematic approach to collect, integrate, analyze, and interpret data from online reviews to understand the factors influencing demand for private hospitals in Selangor. Advanced techniques like machine learning and natural language processing were used to filter out fake reviews and extract meaningful insights from large datasets

Results

Of the 14,938 data collected, a total of 12,035 (81%) evaluations were accompanied by comments while 2,903 (19%) evaluations were not accompanied by comments. Among all reviews with comments, 1,121 (9.3%) reviews were fake and excluded from the data. There are 1,279 evaluations that have issues (negative sentiments) and 9,635 evaluations do not have issues (positive or neutral sentiments). The 1,279 evaluations give us the following analysis:

Table 1. Confusion Matrix for Different Data Types & Machine Learning Algorithms

1. Logistic Regression with Binary Independent Variables				
Rating		**Predicted**		
		Low	**High**	**Total**
Actual	**Low**	184	9	193
	High	21	72	93
	Total	205	81	286
2. Logistic Regression with Scalar Independent Variables				
Rating		**Predicted**		
		Low	**High**	**Total**
Actual	**Low**	183	10	193
	High	12	81	93
	Total	195	91	286
3. Artificial Neural Network with Binary Independent Variables				
Rating		**Predicted**		
		Low	**High**	**Total**
Actual	**Low**	173	20	193
	High	26	67	93
	Total	199	87	286
4. Artificial Neural Network with Scalar Independent Variables				
Rating		**Predicted**		
		Low	**High**	**Total**
Actual	**Low**	185	8	193

	High	15	78	93
	Total	200	86	286

Table 2. Classification Report for Different Data Types & Machine Learning Algorithms

Logistic Regression with Binary Independent Variables				
Rating	**Precision**	**Recall**	**F1-score**	**Support**
Low	0.90	0.95	0.92	193
High	0.89	0.77	0.83	93
Accuracy	0.90			286
Logistic Regression with Scalar Independent Variables				
Rating	**Precision**	**Recall**	**F1-score**	**Support**
Low	0.94	0.95	0.94	193
High	0.89	0.87	0.88	93
Accuracy	0.92			286
Artificial Neural Network with Binary Independent Variables				
Rating	**Precision**	**Recall**	**F1-score**	**Support**
Low	0.87	0.90	0.88	193
High	0.77	0.72	0.74	93
Accuracy	0.84			286
Artificial Neural Network with Scalar Independent Variables				
Rating	**Precision**	**Recall**	**F1-score**	**Support**
Low	0.93	0.96	0.94	193
High	0.91	0.84	0.87	93
Accuracy	0.92			286

Factor Analysis for Theme Formation

Scree plot shows 8 factors with Eigenvalue > 1 for binary data type (Figure 1a) and 6 factors for scalar data type (Figure 1b). Full list of cumulative explained variance can be viewed in **Appendix 3**.

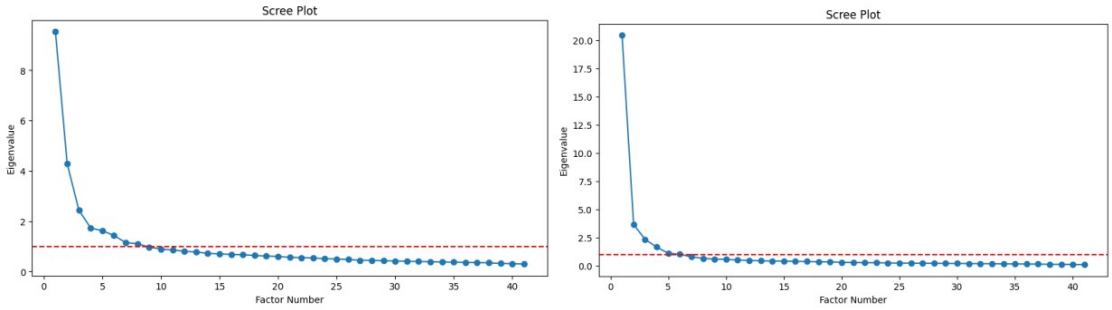


Figure 6. Scree plot for binary (A) and scalar (B) data type.

Table 3. Retained factors and proposed labels.

Data Type	Factors	Proposed Factor Labels	Cronbach alpha
Binary	**Factor 1**	Service Efficiency & Staff Management	0.88
	Factor 2	Environment & Facilities Quality	0.81
	Factor 3	Clinical Care & Patient Safety	0.78
	Factor 4	Communication Barriers	0.61
	Factor 5	Financial Concerns & Insurance Issues	0.81
	Factor 6	Organizational Coordination & Efficiency	0.57
	Factor 7	Patient Support	0.48
	Factor 8	Staff Training & Attitude	0.67
Scale	**Factor 1**	Service & Communication Effectiveness	0.97
	Factor 2	Clinical Care & Patient Experience	0.92
	Factor 3	Facilities & Amenities Quality	0.90
	Factor 4	Appointment & Patient Flow	0.89
	Factor 5	Financial & Insurance Management	0.88
	Factor 6	Patient Rights & Accessibility	0.61

Table 4. Logit Regression Result for Binary Independent Variables

Factor	Coefficient	Std. Error	z	p value	CI (Min)	CI (Max)
SESM	-0.8151	0.080	-10.242	< 0.001	-0.971	-0.659
EFQ	0.4513	0.071	6.387	< 0.001	0.313	0.590
CCPS	-0.6345	0.073	-8.651	< 0.001	-0.778	-0.491
CB	-1.4047	0.129	-10.913	< 0.001	-1.657	-1.152
FCII	-0.2873	0.069	-4.185	< 0.001	-0.422	-0.153
Pseudo-R^2 = 0.41						

SESM - Service Efficiency & Staff Management; EFQ - Environment & Facilities Quality; CCPS - Clinical Care & Patient Safety; CB - Communication Barriers; FCII - Financial Concerns & Insurance Issues

Table 5. Logit Regression Result for Scalar Independent Variables

Factor	Coefficient	Std. Error	z	p value	CI (Min)	CI (Max)
SCE	-2.0960	0.106	-19.691	< 0.001	-2.305	-1.887
FAQ	-0.2849	0.078	-3.670	< 0.001	-0.437	-0.133
APF	-0.4409	0.078	-5.686	< 0.001	-0.593	-0.289
FIM	-0.2095	0.083	-2.522	< 0.001	-0.372	-0.047
PRA	-0.5349	0.091	-5.894	< 0.001	-0.713	-0.357
Pseudo-R^2 = 0.55						

SCE - Service & Communication Effectiveness; FAQ - Facilities & Amenities Quality; APF - Appointment & Patient Flow; FIM - Financial & Insurance Management; PRA - Patient Rights & Accessibility.

Discussion

The analysis of review content identified a substantial portion of evaluations (81%) accompanied by comments, with 19% lacking comments. Of the reviews with comments, 9.3% were deemed fake and excluded, leaving 9,594 evaluations without issues and 1,279 with issues. The use of gpt-4o-mini model for coding these reviews showed a high inter-rater reliability, with an average Cohen's Kappa score of 0.81, indicating strong agreement between human and AI coders. This high level of consistency supports the validity of the coding process and the reliability of the insights derived from the data.

Machine Learning Performance

Overall, scalar data type produces higher precision, recall F1 score and accuracy in all ML algorithms as compared to binary data type (Please refer **Table 1 & 2**). Prior to this study, we conducted an experiment using non-categorized ratings i.e rating of 1, 2, 3, 4 and 5 instead of categorizing them into ‘high’ or ‘low’ rating. We found out that maintaining the y variable in its ordinal form brought difficulties in analysis and interpretation. As suggested by literature review, we also found imbalance in distribution of rating where rating 1 and 5, made the majority classes in the data. Our original data is imbalanced with 60% of ratings belonging to rating 1, 30% to rating 5, and 10% to ratings 2, 3, and 4.

Even after oversampling of ratings 2, 3, and 4 (to balance the data), the precision, recall, and F1

score (tested on untouched test data) remain low for ratings 2, 3, and 4 in all ML algorithms. This is due to significant overlap in terms of the variables distribution, making it difficult for the model to distinguish between them. If the differences between these ratings are subtle or not well represented by the features, the model will struggle to accurately classify them, resulting in low performance for these classes.

Despite balancing the class distribution, the model can also develop a bias toward predicting the majority classes (rating 1 and rating 5), especially because the minority classes (ratings 2, 3, and 4) do not have sufficient discriminative power. The model may predict the majority classes more often, which leads to low precision and recall for the minority classes.

Since the ML algorithms show high performances on extreme ends (rating 1 and 5), a reclassification of data made by transforming rating 1 and 2 to ‘low’ and rating 4 and 5 to ‘high’, and abandoning rating 3 (that has very low discriminative power) to produce a dichotomous category fit for logistic regression. The logistic regression shows a robust model with pseudo-R^2 of 0.41 for binary and 0.55 for scalar data type. This shows that the transformation of the original ratings into a dichotomous 'low' and 'high' classification significantly enhances the predictive power of the model, especially when compared to using the full, untransformed data. By eliminating the less informative middle category (rating 3), the model can focus on the more discriminative extremes, leading to improved model performance.

Predictive Model

For binary *x* data type, ML classification was able to produce f1-score of 0.83 - 0.92 and accuracy of 0.90 for LR and f1-score of 0.74 - 0.88 and accuracy of 0.84 for ANN model (refer **Table 2**). For scalar *x* data type, ML classification was able to produce f1-score of 0.88 - 0.94 and accuracy of 0.92 for LR and f1-score of 0.87 - 0.94 and accuracy of 0.92 for ANN model. This is a proof that scalar data produces better performance as compared to binary data in all circumstances of conducting LACA. It is also interesting to note that ANN algorithms produce lower accuracy as compared to logit regressions in both data types, similar to finding with a study by [38]. This suggests that logit regression is a better model for a categorized y variable (since our previous experiment also shows better performance of ANN on non-categorized y variable as compared to logit regression), suggesting that logit regression is better at predicting binary y variable and ANN is better at predicting multi-class y variable.

Table 6. Summary of results for predictive models

y variable	x variables	Model	F1 score	Accuracy
Original Rating (1 - 5) Converted to Binary (Low or High)	Binary	Logit	0.83 - 0.92	0.90
		ANN	0.74 - 0.88	0.84
	Scalar	Logit	0.88 - 0.94	0.92
		ANN	0.87 - 0.94	0.92

Factor Analysis

Factor analysis on binary data identified seven interpretable latent factors: 'Service Efficiency & Staff Management’, 'Environment & Facilities Quality', ‘Clinical Care & Patient Safety’, ‘Communication Barriers’, ‘Financial Concerns & Insurance Issues’, ‘Organizational Coordination & Efficiency’, ‘Patient Support’ and 'Staff Training & Attitude'. The cumulative

explained variance for the seven variables is 0.57, far below a study using similar method for binary data [39] with CEV of 0.71. For scalar data, six latent factors identified are 'Service & Communication Effectiveness', 'Clinical Care & Patient Experience', 'Facilities & Amenities Quality', 'Appointment & Patient Flow', 'Financial & Insurance Management', and 'Patient Rights & Accessibility'. The cumulative explained variance for the six variables is 0.74.

We were able to get higher (+0.20) cumulative explained variance for the scalar data type as compared to binary data type for the factors with Eigenvalues > 1, showing that the scalar data type captures more variability in the data and explains a greater proportion of the underlying structure. This suggests that scalar features provide more detailed information, allowing the model to capture more meaningful patterns compared to binary features, which may be more limited in their ability to represent complex relationships or variability in the data.

Comparing Latent Factors between Data Types

Both the binary and scalar data analyses identified a factor related to financial issues. In the binary data, this factor is labeled as "Financial Concerns & Insurance Issues" (Cronbach's alpha 0.81) while in the scalar data, it is labeled as "Financial & Insurance Management" (Cronbach's alpha 0.88). The binary data analysis identified a factor related to "Staff Training & Attitude" (Cronbach's alpha 0.67), which likely pertains to perceptions of staff behavior and interpersonal interactions with patients. The scalar data analysis, however, identified a factor called "Service & Communication Effectiveness" (Cronbach's alpha 0.97) which encompass broader aspects of service delivery, including communication between staff and patients. While both factors deal with service quality, the binary factor is more focused on the staff's attitude, while the scalar factor likely captures a broader dimension of service and communication.

The binary data identified "Clinical Care & Patient Safety" (Cronbach's alpha 0.78), while the scalar data identified "Clinical Care & Patient Experience" (Cronbach's alpha 0.92). The scalar data factor explicitly incorporates "patient experience," suggesting a broader, more patient-centered focus. The binary factor "Clinical Care & Patient Safety" is likely more focused on the technical and clinical aspects of care delivery, whereas the scalar factor appears to capture both clinical care and how patients perceive their care experience.

The binary data included a factor called "Environment & Facilities Quality" (Cronbach's alpha 0.81), which likely includes elements related to the maintenance, infrastructure, and engineering of healthcare facilities. The scalar data, on the other hand, identified a factor called "Facilities & Amenities Quality" (Cronbach's alpha 0.90), which focuses more on the quality of the physical environment and amenities from the patient's perspective. While both are related to the physical environment, the binary factor leans more toward operational management, whereas the scalar factor may be more related to patient satisfaction with the physical setting.

The binary analysis revealed a factor related to "Organizational Coordination & Efficiency" (Cronbach's alpha 0.57), which captures the overall effectiveness and productivity of healthcare systems. In contrast, the scalar data identified a factor called "Patient Rights & Accessibility" (Cronbach's alpha 0.61), which focuses more on the accessibility of care and ensuring patients' rights are upheld. These two factors are conceptually distinct, one focuses on operational efficiency within the organization, while the other is concerned with ensuring patients have access to appropriate care and that their rights are respected.

The binary analysis identified a factor called "Communication Barriers", (Cronbach's alpha 0.61) which refers to difficulties in communication, either among healthcare providers or between providers and patients. The scalar analysis did not explicitly identify a similar factor, suggesting that communication barriers may not have emerged as a distinct latent factor in the scalar data, perhaps due to differences in how communication issues were measured or conceptualized.

Scalar data are superior in its internal consistency as compared to the binary data, evidenced by higher Cronbach's alpha in all factors, showing that the scalar data provide more reliable and coherent measurement of the latent constructs, with items within each factor being more closely aligned and consistently measuring the same underlying concept. This indicates that scalar variables, due to their continuous nature, allow for greater variability and finer distinctions in responses, leading to higher consistency in how the factors are represented. In contrast, binary data, with their limited response options, result in less variation and weaker consistency, thus leading to lower Cronbach's alpha values.

Ferreira et al. (2023) [40] in his systematic review, ranked criteria deemed as the most important to evaluate satisfaction in literature (global analysis). The factors generated in our study are all included in the criteria. According to Ferreira et al. (2023), medical care is deemed as most important in literature (34%), communication (31%), doctor's characteristics (28%), accommodations (23%), admission and discharge (13%), nurse characteristics (11%), appointment (8%), environment (8%), medical expenditure (8%), and organization (8%). Ferreira et al. however derive these numbers through literature reviews thus there are overlapping domains within his full list of 56 criteria, almost similar to what we experienced before conducting factor analysis.

Explanatory model

Cumulative explained variance (CEV) of 0.74, average Cronbach's alpha of 0.86 and pseudo-R^2 of 0.55 for the scalar model suggest that the scalar *x* data type performs better at explaining factors that influence online rating as compared to binary *x* data type which give CEV of 0.54, average Cronbach's alpha of 0.73 and pseudo-R^2 values of 0.41. These results also indicate that the reclassification of *y* into binary data, the reduction of variables using factor analysis, transformation of *x* variables into factor scores and logistic regression of the scores against the reclassified *y* are effective in capturing meaningful patterns in the data, making the model a robust tool for explaining the factors influencing online rating. The log-odds formula for our models developed based on binary and scalar *x* data type are written as:

Explanatory model developed based on binary *x* data type

$$\begin{aligned}\text{Log}(P(\text{high})/P(\text{low})) = {} & -0.8X(\text{Service Efficiency \& Staff Management}) \\ & +0.5X(\text{Environment \& Facilities Quality}) \\ & -0.6X(\text{Clinical Care \& Patient Safety}) \\ & -1.4X(\text{Communication Barriers}) \\ & -0.3X(\text{Financial Concerns \& Insurance Issues}) \\ & -0.6X(\text{Staff Training \& Attitude}) \\ & +0.01\end{aligned}$$

Model's pseudo-R^2 = 0.41

Explanatory model developed based on scalar *x* data type

$$\begin{aligned}\text{Log}(P(\text{high})/P(\text{low})) = {} & -2.1X(\text{Service \& Communication Effectiveness}) \\ & -0.3X(\text{Facilities \& Amenities Quality}) \\ & -0.4X(\text{Appointment \& Patient Flow}) \\ & -0.2X(\text{Financial \& Insurance Management}) \\ & -0.5X(\text{Patient Rights \& Accessibility}) \\ & +0.02\end{aligned}$$

$$\text{Model's pseudo-}R^2 = 0.55$$

The first model, which uses binary data types, shows a counterintuitive positive association between "Environment & Facilities Quality" and the outcome (log-odds of being classified as "high" vs. "low"). Specifically, the coefficient for this variable is +0.5, meaning that better perceived environmental and facility quality is associated with higher odds of being classified as "high." This result might seem unusual, especially because we expected problems with facilities and amenities to decrease ratings or outcomes. Counterintuitive findings are not uncommon in observational research [41]. This finding can be explained by several factors related to the nature of the model, the data, and how binary variables are coded and interpreted.

With binary data, the model is more likely to show simplified or lumped effects. For example, the "Environment & Facilities Quality" variable could be capturing whether the facility has any issues at all (e.g. whether a problem exists versus no problem), which might not fully capture the severity of the issue. If the dataset includes a substantial number of facilities with minor or non-critical problems, this binary categorization might show an overall positive trend in how people perceive and rate the environment. People could still rate a facility highly if they see that it meets a certain baseline of comfort or aesthetic quality, despite problems with specific amenities.

Binary variables often oversimplify complex relationships. In this case, "Environment & Facilities Quality" might simply be a dichotomous indicator (problem vs. no problem), which fails to account for the degree of problem severity. For example, if a facility has a few minor maintenance issues but is otherwise clean, well-kept, and functional, customers might rate it positively overall. As a result, the binary coding might mask the full complexity of the relationship, leading to a higher log-odds ratio for "better environment and facilities."

The pseudo-R^2 of 0.41 for the binary model indicates that the model explains a moderate amount of the variation in the outcome, but there is still unexplained variance that could be influencing the results. In particular, the binary nature of the predictors in the first model may not fully capture how facility quality influences ratings. The simplicity of binary coding can sometimes lead to counterintuitive findings when there are other confounding factors or interactions the model does not account for.

It's also possible that "Environment & Facilities Quality" is interacting with other variables in ways that aren't immediately apparent in the binary model. For example, Service Efficiency & Staff Management might have a more dominant effect on the outcome, and the model could be highlighting a compensatory relationship where facilities issues are less critical if the service or

staff quality is strong. This interaction might be captured better in a multivariate model with scalar predictors (as seen in the second model), where the more granular data allows for a more precise understanding of how the environment and facilities quality is influencing the outcome.

The use of binary predictors might miss nonlinear effects that are better captured by scalar data. For example, if minor issues with facilities are common and not considered deal-breakers by many customers, the binary categorization may smooth out the differences in how customers respond to these issues. This might result in the positive coefficient for Environment & Facilities Quality, as the model sees it as an overall positive factor. In contrast, the second model, with scalar predictors, may better differentiate between minor and major problems, and thus offer a more nuanced understanding of how facilities quality impacts ratings.

It's also possible that the binary coding of "Environment & Facilities Quality" doesn't fully account for how guests perceive quality. Guests may give a higher rating despite problems because they are more forgiving of minor issues, particularly if other aspects of the experience (e.g., service, comfort) are satisfactory. The binary model may not distinguish between small problems (e.g., an occasional issue with heating or plumbing) and larger, more significant problems, leading to the appearance that environment and facilities quality positively influences the likelihood of a "high" outcome.

Conclusion

Our factor analysis revealed distinct patterns across both data types, with scalar data offering better understanding of the underlying structure. The scalar data not only provided higher cumulative explained variance but also demonstrated superior internal consistency, as evidenced by stronger Cronbach's alpha values. This suggests that scalar features, with their continuous nature, allow for finer distinctions and more reliable measurement of complex constructs compared to the limited binary data.

Machine learning performance further highlighted the advantages of scalar data, which consistently outperformed binary data across key metrics such as precision, recall, F1 score, and accuracy. While challenges related to imbalanced classes and subtle class distinctions remained, the transformation of the original ratings into a dichotomous 'low' and 'high' classification improved the predictive power of the model, particularly in logistic regression. The results from logistic regression, with pseudo-R^2 values of 0.41 for binary and 0.55 for scalar data, underscore the effectiveness of the reclassification approach and validate the superiority of scalar data in capturing meaningful patterns within the review content.

Overall, our findings provide strong evidence that scalar data offers greater utility and predictive power in healthcare review analysis, enabling more precise insights into patient satisfaction and facilitating the development of more robust predictive models. By enhancing both the conceptual understanding and the technical performance of the model, scalar data proves to be a more effective tool for analyzing complex, multidimensional datasets in healthcare settings.

Appendix 1

API call for giving a binary class for each thematic code on each online review.

```
def identify_scale(review, codebook_file):
  list_of_codes = open_txt(codebook_file).split('\n')
  # Initialize the OpenAI API client
  client = OpenAI(
      # This is the default and can be omitted
      api_key= API_KEY
  )
  result_list = []
  for code in list_of_codes:
    chat_completion = client.chat.completions.create(
      messages=[
                {
                  "role": "user",
                  "content": (
    f"Is '{code}' one of the issue(s) inside "
    f"the following statement?\n\nStatement:\n"
    f"{review}\n\nAnswer Yes or No only "
              )
                }
      ],
      model="gpt-4o-mini",
  )
  result_list.append(
    chat_completion.choices[0].message.content.lower().strip()
  )

  return result_list

def reclass(l_):
rec = [1 for x in l_ if ‘yes’ in x.strip() else 0]
```

Appendix 2

API call for giving a scale for each thematic codes on each online review.

```
def identify_scale(review, codebook_file):
  list_of_codes = open_txt(codebook_file).split('\n')
  # Initialize the OpenAI API client
  client = OpenAI(
      # This is the default and can be omitted
      api_key= API_KEY
  )
  result_list = []
  for code in list_of_codes:
    chat_completion = client.chat.completions.create(
      messages=[
                 {
                   "role": "user",
                   "content": (
    f"Given Likert scale 0 = 'Not an issue', "
    f"1 = 'A small issue', "
    f"2 = 'A moderate issue', "
    f"3 = 'A serious issue', "
    f"4 = 'An extremely serious issue', "
    f"is '{code}' one of the issue(s) inside "
    f"the following statement?\n\nStatement:\n"
    f"{review}\n\nAnswer the value of the "
    f"likert scale i.e 0 or 1 or 2 or 3 or 4"
              )
                 }
      ],
      model="gpt-4o-mini",
  )
  result_list.append(
    chat_completion.choices[0].message.content.lower().strip()
  )
  return result_list
```

Appendix 3

Cumulative Explained Values

Data Type	Cumulative Explained Values (F01 - F41)
Binary	F01: 0.23, F02: 0.34, F03: 0.40, F04: 0.44, F05: 0.48, F06: 0.51, F07: 0.54, F08: 0.57, F09: 0.59, F10: 0.61, F11: 0.63, F12: 0.65, F13: 0.67, F14: 0.69, F15: 0.71, F16: 0.72, F17: 0.74, F18: 0.76, F19: 0.77, F20: 0.79, F21: 0.81, F22: 0.81, F23: 0.83, F24: 0.84, F25: 0.85, F26: 0.86, F27: 0.87, F28: 0.88, F29: 0.89, F30: 0.90, F31: 0.90, F32: 0.92, F33: 0.93, F34: 0.94, F35: 0.95, F36: 0.96, F37: 0.97, F38: 0.99, F39: 0.99, F40: 1.00, F41: 1.00.
Scalar	F01: 0.50, F02: 0.59, F03: 0.64, F04: 0.69, F05: 0.71, F06: 0.74, F07: 0.76, F08: 0.77, F09: 0.79, F10: 0.80, F11: 0.81, F12: 0.82, F13: 0.84, F14: 0.85, F15: 0.86, F16: 0.87, F17: 0.88, F18: 0.89, F19: 0.90, F20: 0.90, F21: 0.91, F22: 0.91, F23: 0.92, F24: 0.93, F25: 0.95, F26: 0.96, F27: 0.96, F28: 0.97, F29: 0.98, F30: 0.99, F31: 0.99, F32: 1.00, F33: 1.00, F34: 1.00, F35: 1.00, F36: 1.00, F37: 1.00, F38: 1.00, F39: 1.00, F40: 1.00, F41: 1.00.

References

[1][] Lambert, M.J. and Shimokawa, K., 2016. Collecting client feedback. *Americal Psychological Association*.

[2][] Gondek, D., Edbrooke-Childs, J., Fink, E., Deighton, J. and Wolpert, M., 2016. Feedback from outcome measures and treatment effectiveness, treatment efficiency, and collaborative practice: A systematic review. *Administration and Policy in Mental Health and Mental Health Services Research*, *43*, pp.325-343.

[3][] Muhammad Butt, M. and Cyril de Run, E., 2010. Private healthcare quality: applying a SERVQUAL model. *International journal of health care quality assurance*, *23*(7), pp.658-673.

[4][] Aliman, N.K. and Mohamad, W.N., 2016. Linking service quality, patients' satisfaction and behavioral intentions: an investigation on private healthcare in Malaysia. *Procedia-social and behavioral sciences*, *224*, pp.141-148.

[5][] Abd Rashid, M., Mansor, A. and Hamzah, M., 2011. service quality and patients' satisfaction in healthcare service in Malaysia. *International Journal of Customer Service Management*, *1*(1), pp.41-49.

[6][] Greaves, F., Laverty, A.A., Cano, D.R., Moilanen, K., Pulman, S., Darzi, A. and Millett, C., 2014. Tweets about hospital quality: a mixed methods study. *BMJ quality & safety*, *23*(10), pp.838-846.

[7][] Hawkins, J.B., Brownstein, J.S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E.O., McIver, D.J., Rozenblum, R., Wright, A., Bourgeois, F.T. and Greaves, F., 2016. Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ quality & safety*, *25*(6), pp.404-413.

[8][] Lee, Y.U., Chung, S.H. and Park, J.Y., 2024. Online Review Analysis from a Customer Behavior Observation Perspective for Product Development. *Sustainability*, *16*(9), p.3550.

[9][] Ranard, B.L., Werner, R.M., Antanavicius, T., Schwartz, H.A., Smith, R.J., Meisel, Z.F., Asch, D.A., Ungar, L.H. and Merchant, R.M., 2016. What can Yelp teach us about measuring hospital quality?. *Health Affairs (Project Hope)*, *35*(4), p.697.

[10][] Rahim, A.I.A., Ibrahim, M.I., Musa, K.I., Chua, S.L. and Yaacob, N.M., 2021, October. Patient satisfaction and hospital quality of care evaluation in malaysia using servqual and facebook. In *Healthcare* (Vol. 9, No. 10, p. 1369). MDPI.

[11][] Braun, V. and Clarke, V., 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), pp.77-101.

[12][] Merriam, S.B. and Tisdell, E.J., 2015. *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

[13][] Tuan, A. and Grandi, S., 2018. Emerging trends in qualitative research: a focus on Social Media. *Mercati e competitività: 4, 2018*, pp.17-26.

[14][] Hassani, H., Beneki, C., Unger, S., Mazinani, M.T. and Yeganegi, M.R., 2020. Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*(1), p.1.

[15][] Stammbach, D., Antoniak, M. and Ash, E., 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. *arXiv preprint arXiv:2205.07557*.

[16][] Vaswani, A., 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

[17][] Xiao, Z., Yuan, X., Liao, Q.V., Abdelghani, R. and Oudeyer, P.Y., 2023, March. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion proceedings of the 28th international conference on intelligent user interfaces* (pp. 75-78).

[18][] Lubis, A.R., 2023. Balancing the Equation: Investigating AI Advantages, Challenges, and Ethical Considerations in the Context of GPT-3, Natural Language Processing, and Researcher Roles. *SAR Journal-Science and Research*, *6*(4), pp.257-262.

[19][] Oniani, D., Hilsman, J., Zang, C., Wang, J., Cai, L., Zawala, J. and Wang, Y., 2024. Emerging Opportunities of Using Large Language Language Models for Translation Between Drug Molecules and Indications. *arXiv preprint arXiv:2402.09588*.

[20][] Chiang, C.C., Luo, M., Dumkrieger, G., Trivedi, S., Chen, Y.C., Chao, C.J., Schwedt, T.J., Sarker, A. and Banerjee, I., 2024. A large language model–based generative natural language processing framework fine-tuned on clinical notes accurately extracts headache frequency from electronic health records. *Headache: The Journal of Head and Face Pain*, *64*(4), pp.400-409.

[21][] Wang, H., Gao, C., Dantona, C., Hull, B. and Sun, J., 2024. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *npj Digital Medicine*, *7*(1), p.16.

[22][] Gandomi, A., Wu, P., Clement, D.R., Xing, J., Aviv, R., Federbush, M., Yuan, Z., Jing, Y., Wei, G. and Hajizadeh, N., 2024. ARDSFlag: an NLP/machine learning algorithm to visualize and detect high-probability ARDS admissions independent of provider recognition and billing codes. *BMC Medical Informatics and Decision Making*, *24*(1), p.195.

[23][] Namita, N.R., Comparison between Traditional topic Modeling and Generative AI Topic Modeling.

[24][] World Health Organization, 2019. Regional action agenda on harnessing e-health for improved health service delivery in the Western Pacific. Chapter 4, p.22.

[25][] Ministry of Health. 2023. Health White Paper for Malaysia. Chapter 3, p.32-55.

[26][] Khan, A.H., Kegalle, H., D'Silva, R., Watt, N., Whelan-Shamy, D., Ghahremanlou, L. and Magee, L., 2024. Automating Thematic Analysis: How LLMs Analyse Controversial Topics. *arXiv preprint arXiv:2405.06919*.

[27][] Asaad, W.H., Allami, R. and Ali, Y.H., 2023. Fake Review Detection Using Machine Learning. *Revue d'Intelligence Artificielle*, *37*(5).

[28][] Elmogy, A.M., Tariq, U., Ammar, M. and Ibrahim, A., 2021. Fake reviews detection using supervised machine learning. *International Journal of Advanced Computer Science and Applications*, *12*(1).

[29][] Alsubari, S.N., Deshmukh, S.N., Alqarni, A.A., Alsharif, N., Aldhyani, T.H., Alsaade, F.W. and Khalaf, O.I., 2022. Data analytics for the identification of fake reviews using supervised learning. *Computers, Materials & Continua*, *70*(2), pp.3189-3204.

[30][] Zhang, D., Zhou, L., Kehoe, J.L. and Kilic, I.Y., 2016. What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems*, *33*(2), pp.456-481.

[31][] Roh, M. and Yang, S.B., 2021. Exploring extremity and negativity biases in online reviews: Evidence from Yelp. com. *Social Behavior and Personality: an international journal*, *49*(11), pp.1-15.

[32][] Powers, D.M., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

[33][] Sovacool, B.K., 2013. A qualitative factor analysis of renewable energy and Sustainable Energy for All (SE4ALL) in the Asia-Pacific. *Energy Policy*, *59*, pp.393-403.

[34][] Starkweather, J., 2014. Factor analysis with binary items: A quick review with examples. *University of North Texas Benchmarks*.

[35][] STATA, 2024. Tetrachoric correlations for binary variables – Stata

[36][] George, D., & Mallery, P. (2003). SPSS for Windows Step by Step: A Simple Guide and Reference. 11.0 Update (4th ed.). Boston: Allyn & Bacon.

[37][] Yusoff, M.S.B., 2019. ABC of content validation and content validity index calculation. *Education in medicine journal*, *11*(2), pp.49-54.

[38][] Eftekhar, B., Mohammad, K., Ardebili, H.E., Ghodsi, M. and Ketabchi, E., 2005. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. *BMC medical informatics and decision making*, *5*, pp.1-8.

[39][] Mehrolia, S., Alagarsamy, S. and Solaikutty, V.M., 2021. Customers response to online food delivery services during COVID-19 outbreak using binary logistic regression. *International journal of consumer studies*, *45*(3), pp.396-408.

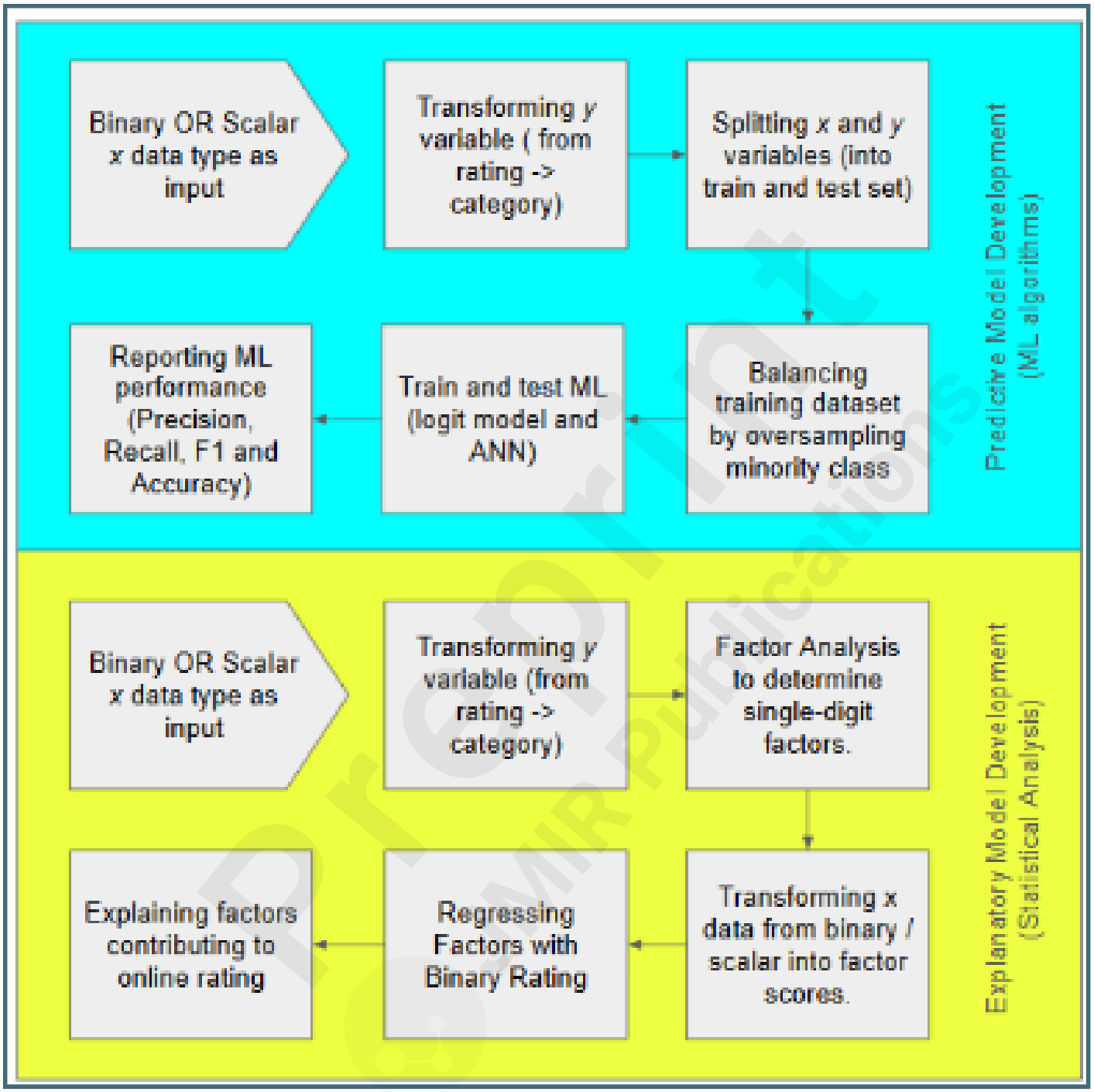
[40][] Ferreira, D.C., Vieira, I., Pedro, M.I., Caldas, P. and Varela, M., 2023, February. Patient satisfaction with healthcare services and the techniques used for its assessment: a systematic literature review and a bibliometric analysis. In *Healthcare* (Vol. 11, No. 5, p. 639). MDPI.

[41][] Doty, E., Stone, D.J., McCague, N. and Celi, L.A., 2019. Counterintuitive results from observational data: a case study and discussion. *BMJ open*, *9*(5), p.e026447.

Supplementary Files

Figures

Flow of study methodology.



Multimedia Appendixes

API call for giving a binary class for each thematic code on each online review.
URL: http://asset.jmir.pub/assets/a605e4d7581168dd850e951270137470.docx

API call for giving a scale for each thematic codes on each online review.
URL: http://asset.jmir.pub/assets/26fcebea5a66c8a4bafb09c09446e7e8.docx

Cumulative Explained Values.
URL: http://asset.jmir.pub/assets/e89b42cfd144bd6117f6153bdf13ddf1.docx