ISSN: 3049-7159

Review Article

Journal of Business and Econometrics Studies

Amazon Phone Reviews Analysis: Using an Econometric and Deeplearning Approach

Sebastián Chamarravi*, Daniel Duque and Bilal Faraj

Pontifical Xavierian University, Colombia

***Corresponding author**

Sebastián Chamarravi, Pontifical Xavierian University, Colombia.

Received: June 13, 2026; **Accepted:** June 22, 2026; **Published:** June 29, 2026

ABSTRACT

In this study, 334,328 comments on unlocked mobile phones from Amazon are analyzed, exploring descriptive statistics and constructing predictive models for rating and price using machine learning techniques, an analysis that provides information still lacking sufficient robustness in the literature. To do this, reviews were processed to generate an approximate sentiment score using the TextBlob library, related to the user's rating. In turn, simple and multiple linear regressions, and regressions including interactions, were performed to predict price, as well as multiple linear regression models to predict rating using sentiment score, price, and Review Votes (support for other users' reviews).

Additionally, Multilayer Perceptron (MLP) neural network models and Transformer based models from the Hugging Face library were trained for both prediction tasks. The results show a moderate correlation between sentiment score and Rating ($r \approx 0.554$) and a significantly higher fit for rating prediction with more complex models ($R^2 \approx 0.727$) than with linear regression ($R^2 \approx 0.308$). In contrast, all approaches were poor at attempting to predict price ($R^2 \leq 0.086$), indicating the limited linear relationship between the available attributes and the product's price.

Keywords: Language Processing, Price Prediction, Review Evaluation, Transformers, Finetuning

Literature Review

It would be complicated and almost illogical to deny the existing relationship between Artificial Intelligence and marketing. In the context of this research, it manifests through ratings in product purchases, which, although not a direct type of promotion, does have a significant psychological effect on consumers. These stimuli act as signals of quality and social consensus, reinforcing consumer confidence and familiarity with the product, psychological mechanisms well documented in the literature [1].

Today, digital transformation has greatly simplified and accelerated access to and analysis of large volumes of unstructured data. These reviews have become an important source for understanding consumer behavior and evaluating products from the customer's perspective [2].

Various studies have analyzed sentiment in reviews, meaning they have tried to classify opinions as positive, negative, or neutral. For example, Zhang et al. analyzed mobile application reviews using traditional machine learning algorithms, showing that these techniques can adequately capture general opinion, finding differences based on text length [3]. This is exact experimental documentation in terms of algorithm evaluation, feature representation, and cross-validation; however, they present limitations in handling irony and specific jargon.

In recent years, the introduction of advanced deep learning models, such as BERT (Bidirectional Encoder Representations from Transformers), has significantly improved the accuracy in sentiment classification. A recent study employed BERT models and thematic modeling techniques with sentence transformers to classify sentiments and detect topics in smartphone reviews, revealing important details about recurring consumer preferences and concerns [4]. In a way, it is noteworthy that the study presented technological robustness, including processing and clustering techniques. However, they emphasized the difficulty of these models in adequately classifying negative reviews due to the imbalanced distribution of data towards positive opinions.

Predictive analysis has also been used to estimate numerical ratings directly from review text, which is considered an

Citation: Sebastián Chamarravi, Daniel Duque, Bilal Faraj. Amazon Phone Reviews Analysis: Using an Econometric and Deeplearning Approach. J Bus Econo Stud. 2026. 3(3): 1-4. DOI: doi.org/10.61440/JBES.2026.v3.141

emerging field, but with a long way to go. A broad analysis performed on unlocked mobile phone reviews sold on Amazon showed that advanced models like neural networks and Transformers achieve high accuracy in rating prediction, significantly outperforming the performance of simple linear models. Chen et al [5]. Introduced a hybrid model called TSPRA that integrates topics, sentiments, and user preferences to further improve rating prediction, identifying critical product aspects.

Following the same line of integrating econometric approaches, in this case using advanced natural language processing (NLP) techniques. Adamopoulos, Ghose, and Tuzhilin clearly exemplify this by using vector representations (embeddings) generated with both deep learning models and econometric methods (logit and nested logit) to analyze and examine the heterogeneous effects caused by different recommendation strategies in mobile applications.

However, these advances in research present certain critical limitations, according to Xiao, Hou, and Johnson, highlighting the need for the appropriate and justified use of control variables in these studies to avoid biases and ensure valid inferences [6]. Even more so in the context of training and prediction models, since for correct functioning, introducing noise or inadequate controls could contaminate the estimation. Other challenges are attributed to the inherent biases in review data, typically skewed towards positive ratings, limiting the generalization of models to less frequent negative cases [4].

Additionally, the existence of problems in identifying irony or exaggeration in important variables can lead to adverse effects or weakened statistical power. Also, for certain analyses, such as price, they have proven difficult to predict exclusively from reviews due to the absence of detailed technical information and relevant metadata in the reviewed texts.

Data Analysis

The database used for this study includes variables such as Product Name, Brand Name, Price, Rating (1–5 stars), Reviews (text), and Review Votes (utility votes). A description of what they contain would be, for example, the Brand Name field, which indicates the product manufacturer (e.g., Samsung). According to official documentation, price ranges between $1.73 and $2598, averaging about $226.86, and review votes between 0 and 645, with an average of 1.50. Ratings are from 1 to 5 stars, tending to concentrate around 4 and 3 stars on average.

The descriptive analysis performed showed that most reviews are relatively brief, with an average length of 230 characters, and that there is a positive correlation between price and rating. For sentiment analysis, TextBlob was applied, a Python library for natural language processing that provides a simple API for tasks like sentiment analysis. Each review was processed to obtain a polarity that was used as a Sentiment Score.

- **Brand distribution:** Although our analysis could not obtain an exact frequency f igure, other sources indicate that brands such as Samsung, Apple, Motorola, and BLU dominate unlocked mobile phone sales on Amazon (for example, BLU frequently appears in Kaggle analyses). However, this analysis focuses on the numerical and text variables we have available.
- **Statistical summary:** The average price was $226.86, the average rating close to 4 stars, and the average sentiment score is relatively high, showing a positive trend. Review votes showed high asymmetry, with a mean of 1.5 and a median of 0. These findings are consistent with previous dataset reports.
- **Initial correlations:** Pearson correlation was evaluated between sentiment score and Rating, yielding an $r \approx 0.5538$. This moderately positive correlation indicates that reviews with a 'more positive' sentiment tend to be associated with higher ratings, aligning with both logic and literature, which points to the linear dependence between opinion variables and their ratings.

Methodology for Prediction and Hypothesis Analysis

For the predictive part, we decided to formulate the following hypotheses: H1:
sentiment score positively correlates with Rating; H2: price can be predicted by variables such as Rating and sentiment score; H3: Rating can be predicted by review characteristics (sentiment, price, votes). With this defined, we proceeded with the following methodological steps:

- **Linear regression for price:** For this regression, an ordinary linear regression model using Scikit-learn estimators was used. Three variants were considered: simple regression (e.g., Price ~ Rating), multiple regression (Price ~ Sentiment_Score + Rating + ReviewVotes), and regression with an interaction term (e.g., Price ~ Sentiment_Score * Rating). With LinearRegression, we minimized the sum of residual squares to obtain the coefficients. R^2, MSE, β coefficients, and their p-values were found to evaluate the statistical significance of the different models.
- **Linear regression for rating:** Analogously, a multiple regression model was executed with Rating as the target variable to predict, using sentiment score, price, and review votes as predictor variables. Fit (R^2), MSE, and the relevance of each coefficient were measured to check the linear relationship between sentiment and rating.
- **MLP neural network for price:** Scikit-learn's MLPRegressor was used, which implements a multilayer perceptron neural network trained by backpropagation. For these models, we decided to use the default configuration (one hidden layer with 100 neurons, ReLU activation function, Adam solver), although modifying hidden layers as required. The entire model was trained on the complete dataset, with division into test and train sets. This network attempts to model nonlinear relationships between attributes and price. R^2 and MSE were reported for training and testing.
- **MLP neural network and Transformer for rating:** To predict Rating from review text, two approaches were tested: (a) an MLP similar to the previous one, which received numerical inputs derived from reviews (e.g., simple embeddings), and (b) a Transformer-based model from Hugging Face. The Transformers library was used, which provides state-of-the-art NLP architectures. In this case, a pre trained model was fine-tuned for this regression (rating 1–5). Basic parameters were optimized, and the final fit (R^2, MSE) was calculated.
- **Transformer model for price:** Similar to the previous case, a Hugging Face model was trained to predict price based on user reviews. Since price is a continuous variable

in a wide range, it was treated as a continuous regression (e.g., baseline: linear regressor over the last BERT layers, MLP on top of embeddings, etc.).

- **For each case, standard Python tools were used:** TextBlob for sentiment analysis, scikit-learn for regressions and MLP, and Hugging Face Transformers for deep learning models. These frameworks are documented and have been widely applied in data mining and NLP.

Results

Next, we will present the key results, comparing the performance of the models and commenting on their differences:

- **Price vs. Linear models:** Linear regressions, whether simple or multiple, yielded non-significant coefficients and R^2 very close to zero, practically no explained variance. For example, the Price ~ Rating + Sentiment Score + Review Votes model obtained an R^2 0.01 and extremely high MSE (on the order of tens of thousands). This indicates that the linear relationship between the available predictors and price is almost nonexistent. Individual coefficients, for example, for sentiment score, resulted in high p-values, suggesting that they cannot reliably predict price. This means that linear models did not capture useful price variations in the data.
- **Rating vs. Linear regression:** The linear model for Rating, using sentiment score, price, and review votes, showed a moderate fit, yielding an R^2 of ≈ 0.308. This implies that around 31% of the rating variance is explained by these three variables. The resulting MSE was around ~ 0.9 (given the Rating scale 1–5). Sentiment score emerged with a significant positive coefficient ($p < 0.01$), reflecting a positive correlation with the observed rating ($r \approx 0.554$). Likewise, price showed a slightly positive coefficient, compatible with the correlation reported in the dataset análisis. Review votes were les relevant.
- **MLP for price:** The neural network model obtained an R^2 ≈ 0.012 in both sets (training and testing), with an MSE quite similar to the linear case. This confirms that the MLP failed to capture patterns to predict price; the difference between test and train suggests that there was no noticeable overfitting (both R^2 are equally low). The non-linear capacity of the MLP did not improve the fit compared to linearity, indicating that the review text alone does not provide sufficient information to predict price.
- **MLP and Transformer for rating:** Both approaches achieved remarkably better performance. The MLP with enriched features and the Transformer model achieved an $R^2 \approx 0.727$ on test, with an MSE ≈ 0.65. That is, they explain nearly 72% of the rating variance. This far exceeds the 30% of linear regression. The result suggests that text-based models captured semantic nuances of reviews that correlate with product rating. Transformers allowed learning complex language representations that translated into more precise customer satisfaction predictions.
- **Transformer vs. MLP for rating:** Although both obtained a similar R^2, the Transformer, based on BERT, tends to generalize complex linguistic contexts better. In principle, its deep architecture increases modeling capacity. However, in this case, the numerical fit values were practically identical, suggesting that the MLP with the data used also managed to effectively approximate the text → rating mapping.
- **MLP and Transformer for price:** The Transformer model for price improved slightly ($R^2 \approx 0.086$, MSE ≈ 71592) compared to the basic MLP, but the fit remains very low. The explanation lies in the fact that, although the Transformer processed textual information better, reviews do not substantially explain price differences. Instead, the high price dispersion (average $226 with very large standard deviation) makes any precise prediction difficult with these limited variables.
- **MLP and Transformer for price with Poisson error type:** The Transformer model for price improved slightly ($R^2 \approx 0.107$, MSE ≈ 69956) compared to the previous Transformer model. We hypothesize that this type of error allowed for a better fit of the data due to the similar distribution with the predicted values.

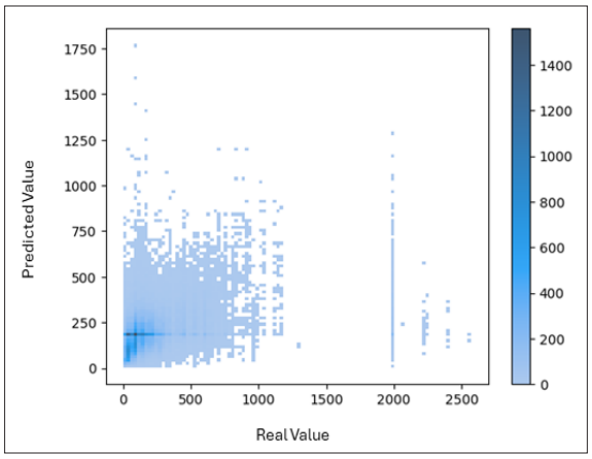

Figure 1: 2D Histogram of Actual Values vs. Predicted Values

The results shown in Figure 1 display the quality of the fit between the predicted and actual values. Firstly, it is interesting to see the increasing component in the histogram, which shows a correlation of the data, evidence of the positive but limited predictive capacity of the model. Secondly, the model has a strong bias towards predicting data around $200, probably because it is the most typical value within the distribution. Additionally, we do not see predictive capacity greater than $1750, which is concerning given the number of phones presented near this value.

Overall, the results indicate that product rating can be reasonably predicted from review content and basic attributes, especially with sophisticated deep learning models, while price remains a variable very poorly correlated with the text of the reviews. This contrast is because reviews explicitly express satisfaction (and therefore rating), but only indirectly reflect price, which depends on technical specifications, brand, and market demand not necessarily explained in user reviews. The literature supports that there is some price-rating correlation, but apparently not linearly or sufficiently for accurate prediction with this class of models.

Conclusions

This study shows the scope and limitations of product review data analysis, and we decided to conclude the following:

- **Limitations:** Simpler predictive models (linear regression) proved insufficient when underlying relationships are non-linear or influenced by unobserved variables. In particular, price prediction showed poor metrics across all models (R^2

very close to 0). This points to an intrinsic limitation of the current dataset: key attributes are missing (e.g., product technical specifications, launch date, brand reputation beyond the name) that would likely improve predictive power. Likewise, the use of TextBlob offers a sentiment score based on general lexicons, which may fail to capture irony, sarcasm, or technical terms specific to mobile reviews. Furthermore, although 334 thousand reviews were used, the distribution of ratings is skewed towards high values, which can induce biases in the models (e.g., lower accuracy for ratings 1–2 due to few examples).

- **Practical utility:** The findings suggest several applications. For ecommerce and marketing companies, sentiment analysis of reviews is useful for estimating the general satisfaction level (as sentiment correlates with rating) and detecting improvement opportunities. For example, by identifying products with negative average sentiment despite good sales, quality issues can be investigated. The poor predictability of price indicates that customer reviews are not enough to estimate market value: companies must supplement with cost data, competitor prices, and technical attributes. Advanced language models (like Transformers) can be used to automate textual feedback processing and enrich recommendation systems or satisfaction metric dashboards (Reimers & Gurevych, 2019).
- **Scientific contribution:** Although the model is not perfect, it shows us that there is the capacity to predict price or value using descriptions or reviews.
- Having the ability to use language modeling for commercial purposes can.
- **Future improvements:** It is recommended to explore more complex learning models and expanded data characterizations. For example, using sentence embeddings like Sentence-BERT to represent the full review text, or incorporating ensemble models (random forests, XGBoost) that integrate product and reviewer variables. It would also be valuable to perform more exhaustive cross-validation and adjust hyperparameters to avoid overfitting. Adding user demographic data (if available) or market conditions could improve price prediction. The integration of sentiment analysis with topic modeling could also refine hypotheses about which product aspects influence scores.

This report shows how NLP and machine learning techniques can extract quantifiable information from reviews to support business decision-making. Although not everything can be predicted (like price, with current data), the integrated approach of descriptive and predictive analysis offers a solid foundation for understanding the interaction between customer perception and market behavior [7,8].

References

1. Mariani MM, Perez-Vega R, Wirtz J. AI in marketing, consumer research and psychology: A systematic literature review and research agenda. Psychology & Marketing. 2022. 39: 755-776.
2. Kraus S, Durst S, Ferreira JJ, Veiga P, Kailer N, et al. Digital transformation in business and management research: An overview of the current status quo. International Journal of Information Management. 2022. 63: 102466.
3. Zhang L, Hua K, Wang H, Qian G. Sentiment analysis on reviews of mobile users. Procedia Computer Science. 2014. 34: 458-465.
4. Jabeen S. Decoding consumer sentiments: Advanced NLP techniques for analyzing smartphone reviews. Revista de Administração Contemporânea. 2024. 28: e-1648.
5. Chen Z, Zhang Y, Shang Y, Hu X. Unifying topic, sentiment & preference in an HDP-based rating regression model for online reviews (TSPRA). Proceedings of ACML 2018.
6. Xiao M, Hou J, Johnson NA. Have we controlled properly? Problems with and recommendations for the use of control variables in information systems research. International Journal of Information Management. 2024. 74: 102717.
7. Adamopoulos P, Ghose A, Tuzhilin A. Heterogeneous demand effects of recommendation strategies in a mobile application: Evidence from econometric models and machine-learning instruments. MIS Quarterly. 2022. 46: 101-150.
8. Google. NotebookLM [Herramienta de inteligencia artificial utilizada para asistencia en traducción]. 2025.