Exploring the Relationship between Review Star Rating and Text Sentiment: Insights from an Ordered Probit Approach

22203640

Alexandros Doganis

COMP0047 Data Science, UCL.

Contents

1 Introduction 1

2 Methodology 2
2.1 Data . . . 2
2.2 Sentiment Analysis . . . 2
2.3 Statistical Analysis . . . 2
2.3.1 Visualization . . . 2
2.3.2 Correlation Matrix . . . 2
2.3.3 Classification . . . 3

3 Results 5
3.1 Visualization . . . 5
3.2 Correlation Matrix . . . 5
3.3 Classification . . . 6

4 Conclusion 7

1 Introduction

While it is widely understood that review ratings in most domains tend to be polarized to the extremes following a J-curve distribution [1], we sought to explore how this phenomenon relates to the sentiment of accompanying review text and ultimately determine how strongly the sentiment of review text relates to its accompanying star rating. We hypothesized that this would be a strong relationship given that a review's rating and text are respectively quantitative and qualitative representations measuring the same experience.

2 Methodology

2.1 Data

We analyzed the Yelp Dataset of business reviews [2]. To eliminate cultural bias, data was localized to reviews left on businesses within California. In preparation for sentiment analysis, the data was cleaned such that only non-empty, English reviews were considered.

2.2 Sentiment Analysis

In order to determine a text's sentiment, we used the Valence Aware Dictionary and sEntiment Reasoner (VADER) [3], a rule-based sentiment analysis tool pre-trained on social media data. These rules are derived from its scored lexicon containing not only words, but emoticons, emojis, colloquialisms, and unconventional punctuations, each with a a pre-assigned sentiment value; this approach attempts to capture linguistic nuances more directly than probabilistic or Deep-Learning based approaches.

Using VADER, each review i was assigned a sentiment score comprised of the probability that the text is positive, negative, or neutral such that:

$$P_{pos}(i) + P_{neg}(i) + P_{neu}(i) = 1 \tag{1}$$

thus completing our feature selection.

2.3 Statistical Analysis

2.3.1 Visualization

Visualization of the distribution of text sentiment for each star rating (see Fig. 1) was achieved by plotting the polarity against the neutrality for every sample point such that:

$$Polarity(i) = P_{pos}(i) - P_{neg}(i) \tag{2}$$

To distinguish distributions, each point was color coded and alpha layered according to its star rating [4]. The mean of each distribution was additionally overlaid on the plot to aid in visual trend identification.

2.3.2 Correlation Matrix

In order to more formally assess the strength of our hypothesis, we computed a correlation matrix R using Spearman's rank correlation coefficient ρ_s [5], given by:

$$\rho_s(A, B) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{3}$$

$$R = \begin{bmatrix} 1 & \rho_s(A,B) \\ \rho_s(B,A) & 1 \end{bmatrix} \tag{4}$$

where d_i is the difference between the ranks of corresponding variables A and B for each observation i, and n is the number of observations [6].

Spearman's rank correlation coefficient was chosen over the more straightforward Pearson's correlation coefficient as Pearson's is only able to measure linear relationships between variables, while Spearman's adaptation of Pearson can capture the monotonic relationship present between our variables. Given that star rating is a categorical variable (integers 1 - 5), Spearman's was more suitable and likely to result in reliable insights.

2.3.3 Classification

Following from our initial analysis and building on the trends identified, we sought to construct a classification model using our sentiment scores to predict star rating; the performance and component parameters of such a model can be harnessed to definitively assess the strength of the relationship.

Our dependant star rating variable is not only categorical, but ordinal i.e., has a natural order. Following from this, we utilized an ordered probit model [5], which is specifically designed to capture the cumulative probability for each ordinal category; this is in contrast to linear or multinomial logistic regression models not intended to capture ordinal relationships.

Such a model functions by identifying a vector of parameters $\boldsymbol{\beta}$ from which we can identify the predictive power of each feature, such that:

$$y^* = \boldsymbol{\beta}\mathbf{x}^T - \epsilon \tag{5}$$

where y^* is an underlying, un-observable dependant variable (overarching rating in our case), $\mathbf{x}$ is a vector of independent variables (our sentiment scores), and ϵ is an error term [7].

Given that the un-observable y^* is in reality our ordered star rating categories 1 to 5, we can define these observable categories, y, via a series of thresholds μ_i such that:

$$y = \begin{cases} 1 & \text{if } y^* \leq \mu_0, \\ 2 & \text{if } \mu_0 < y^* \leq \mu_1, \\ 3 & \text{if } \mu_1 < y^* \leq \mu_2, \\ 4 & \text{if } \mu_2 < y^* \leq \mu_3, \\ 5 & \text{if } \mu_3 < y^* \end{cases} \tag{6}$$

Next we estimated β for each feature in our feature vector $\mathbf{x}$ and μ_i for each categorical threshold. As identified in Eq. 1, any one sentiment probability can be calculated given the other two, thus making inclusion of all three features in our model redundant; we considered only positive and negative scores. To make this estimation, we used Maximum Likelihood Estimation (MLE) to maximize the likelihood function L representing the joint probability of the observed data given the model, given by:

$$L(\boldsymbol{\beta}, \boldsymbol{\mu} | \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} P(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\mu}) \tag{7}$$

where n is the number of observations, and $P(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\mu})$ is the probability of observing a star rating y_i given the independent positive and negative sentiment score variables $\mathbf{x}_i$ and their corresponding parameters β and μ.

Upon completion of the MLE, we calculated their test statistic (z-value) given by:

$$z_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \tag{8}$$

where $\hat{\beta}_i$ is the estimated coefficient for the i-th variable and $SE(\hat{\beta}_i)$ is the standard error of the estimated coefficient.

We then determined the significance of the estimated coefficients by computing the p-value of each using the cumulative distribution function of the standard normal distribution, $\Phi(z)$, and calculating the probability that a random variable from $\Phi(z)$ would be less than or equal to z, given by:

$$p_i = 2 \times (1 - \Phi(|z_i|)) \tag{9}$$

Note: As our star ratings distribution followed a J-curve, we under-sampled the data to ensure each distribution was inline with the least common star rating (2 stars) in order to avoid over-fitting the data and to increase the model's predictive power.

The model's performance was evaluated using the precision, recall, and F1 score for each star rating, as well the model's overall accuracy and weighted F1 score, respectively given by:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{10}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{11}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Observations}} \tag{13}$$

$$\text{Weighted F1} = \frac{1}{\sum_{i=1}^{5} w_i} \sum_{i=1}^{5} w_i \cdot \text{F1}_i \tag{14}$$

where w_i is the weight for the star rating i proportionate to its representation in the dataset, and F1_i is the F1 score for the star rating i.

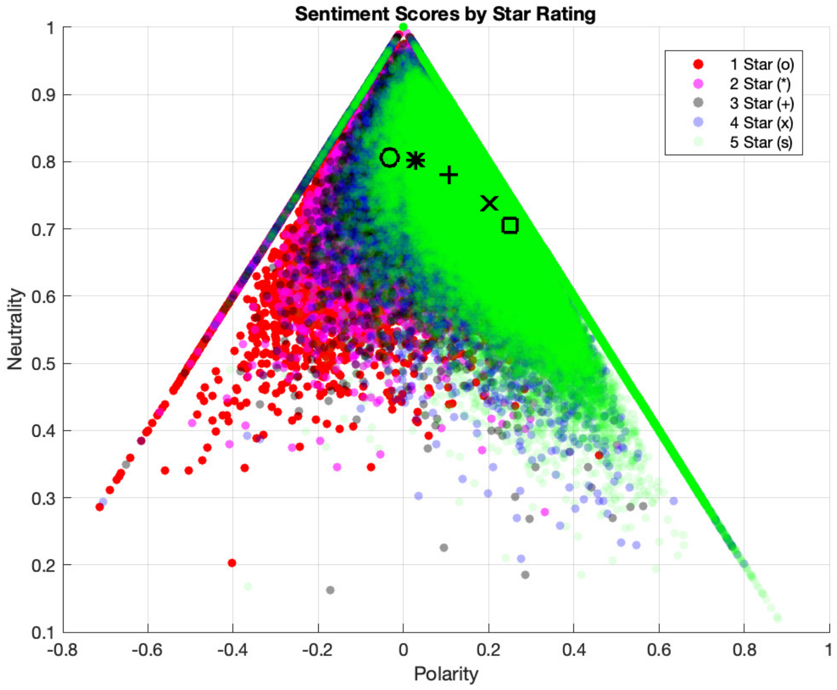

Fig. 1 Polarity vs Neutrality, grouped by star rating. The parenthetical beside each star rating in the legend denotes the symbol marking the mean of the respective distribution. Note that the samples are triangularly bounded due to their relationship (see Eq. 1).

3 Results

3.1 Visualization

Fig. 1 visually identifies that the means of each star rating's sentiment distributions become more positive and less neutral as star rating increases and vice versa, inline with our hypothesis. This effect becomes more pronounced the higher the star rating.

3.2 Correlation Matrix

	stars	neg	neu	pos
stars	1.00	-0.50	-0.38	0.58
neg	-0.50	1.00	0.17	-0.55
neu	-0.38	0.17	1.00	-0.88
pos	0.58	-0.55	-0.88	1.00

Table 1 Spearman's Correlation Matrix

Table 1 shows moderately positive correlation (0.58) between positive sentiment and high star rating. Similarly, it shows moderately negative correlation (-0.50) between negative sentiment and high star rating. Neutral sentiment also shows a slight negative correlation (-0.38) to high star rating.

3.3 Classification

Dep. Variable	Log-Likelihood	AIC	BIC	No. Observations	Df Residuals	Df Model
stars	-1.6000e+05	3.200e+05	3.201e+05	122,131	122,125	6

	coef	std err	z	P>|z|	[0.025	0.975]
neg	-7.26	0.062	-117.308	0.000	-7.38	-7.14
pos	6.12	0.035	174.264	0.000	6.06	6.19
μ_0	-0.639	0.009				
μ_1	-0.212	0.006				
μ_2	-0.117	0.005				
μ_3	-0.215	0.006				

Table 2 Ordered model results star rating classification on sentiment scores. Note that $\boldsymbol{\mu}$ are ancillary variables and therefore do not carry testable significance

Table 2 shows the estimated $\boldsymbol{\beta}$ coefficients. Both positive and negative sentiment scores carry highly positive (6.12) and highly negative (-7.26) coefficients, respectively. Both β values carry p-values < 0.05, indicating that these values are significant and that we can reject the null hypothesis that sentiment scores and star ratings have no relationship.

The $\boldsymbol{\mu}$ threshold values also shown in Table 2 show that the distance between rating categories follows an ordinal structure between the first four categories, with each being larger than the next, with the exception of μ_3 (4/5).

	Precision	Recall	F1 Score
Star Rating 1	0.532	0.581	0.555
Star Rating 2	0.341	0.155	0.214
Star Rating 3	0.331	0.602	0.427
Star Rating 4	0.350	0.121	0.180
Star Rating 5	0.495	0.541	0.517
Overall Accuracy:	0.409		
Weighted F1 Score:	0.381		

Table 3 Ordered model performance metrics for each star rating.

Table 3 indicates that the model can correctly predict star rating based on sentiment scores with 40.9% accuracy. The model has the highest precision (≈ 0.5) at the extreme star ratings and is less precise (≈ 0.3) in the middling ratings. Similarly, the model has high recall (≈ 0.5) at the extremes, however has the highest recall (≈ 0.6) at a star rating of 3, with star ratings 2 and 4 having quite low recall (≈ 0.1). The model's weighted F1 score (see Eq. 14) is a low 38.1%.

4 Conclusion

Both the plot in Fig. 1 and the correlation values shown in Table 1 indicate that as ratings increase, sentiment becomes strictly more positive. In addition, the strong negative correlation between positive and neutral sentiment indicates that neutrality decreases with an increase in star rating inline with the trend identified in Fig. 1. When combined with the similar high positive and negative β coefficients in 2, these results collectively support our initial hypothesis of a strong relationship between review text sentiment and star rating.

Despite this, the ordered probit model did not perform well; the inconsistency in the categories' ordinal relationship in conjunction with the model's low accuracy and weighted F1 score suggest that it is generally a poor classifier. A notable exception, the model was better at predicting extreme values which follows given that these categories were the least bounded, each having only one μ threshold bound. Ultimately, the performance of this classifier suggests that while there is indeed a relationship between the review text sentiment and star rating, it may not be a strong and simple one; the intra-correlations between sentiment scores and distances between ordinal thresholds indicate that their relationship is rich with nuance.

This analysis has so far considered VADER to be a trustworthy ground-truth for sentiment classification, however the study might be improved through involvement of a more sophisticated or purpose-built sentiment classifier; this classifier might consider additional factors such as context, sarcasm, or idiomatic expressions, or may learn sentimental relationships completely independently, each of these possibly providing a more accurate representation of sentiment.

Additionally, it is essential to acknowledge that star ratings are subjective and can vary significantly between individuals, making them an inherently imprecise measure. This subjectivity may contribute to our poor ordinal relationship in Table 2 the difficulty in accurately predicting star ratings based solely on sentiment scores without the inclusion of exogenous data. A potentially more accurate measure of a business's performance could be the average star rating across all reviews, which would smooth out individual variations and provide a more representative view of overall customer sentiment. Future research could investigate the relationship between average star ratings and aggregated sentiment scores, potentially providing more reliable and actionable insights for businesses seeking to understand and improve their customer experience.

References

[1] Hu, N., Pavlou, P.A., Zhang, J.: Can online reviews reveal a product's true quality? empirical findings and analytical modeling of online word-of-mouth communication. In: Proceedings of the 7th ACM Conference on Electronic Commerce, pp. 324–330 (2006)

[2] Yelp: Yelp Open Dataset. https://www.yelp.com/dataset. [Online; accessed March 24, 2023] (2022)

[3] Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, pp. 216–225 (2014)

[4] The MathWorks, Inc.: MATLAB, R2022b edn. The MathWorks, Inc., Natick, MA (2022). The MathWorks, Inc. https://www.mathworks.com/products/matlab.html

[5] Seabold, S., Perktold, J.: Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference **57**, 92–96 (2010)

[6] Fisher, R.A.: Statistical Methods for Research Workers, 13th edn. Hafner, Boston (1958)

[7] Greene, W.H.: Econometric Analysis, 7th edn., pp. 827–831. Pearson Education, Boston (2012)