Analyzing Amazon Product Reviews Based on Helpfulness Ratings Through Content and Language Style

Ashwin Nitnaware

Oakland University

MIS-6940: Project Seminar

Abstract

This study analyzes the review sentiment analysis problem through the lens of the e-commerce industry's demand for better online product reviews by examining what factors influence their helpfulness. The study analyzes Amazon reviews from the consumer electronics section with regards to how content, emotional tone, and style contribute to a review's helpfulness. Natural language processing techniques were applied to extract and evaluate linguistic and sentimental attributes using linear regression models. The research concludes that writing with conviction, authenticity, and emotional transparency as well as acting in alignment with achievement biases significantly aids the perceived helpfulness of reviews, which provides useful strategies for improving review quality and enabling better decisions for consumers and businesses.

Keywords: *E-commerce, Helpfulness Rating, Text Analysis, Linguistic Features, Sentiment Analysis, Machine Learning.*

Main Report

Project Title

Analyzing Amazon Product Reviews Based on Helpfulness Ratings Through Content and Language Style

Organization/Industry Description

The project is conducted in the context of the e-commerce industry, with a particular focus on Amazon product reviews. E-commerce platforms heavily rely on user-generated content, especially reviews, to guide customer purchasing decisions, establish product trust, and increase conversions. Amazon, being a leader in this space, benefits from reviews that are perceived as credible and helpful by other shoppers.

Through psycholinguistic analysis using LIWC, this project analyzes patterns in review content and writing style to decide what makes certain reviews stand out in terms of helpfulness. The analysis includes sentiment detection, tone profiling, and identification of humor and sarcasm — all aimed at better understanding how language influences consumer behavior.

This data-driven approach empowers e-commerce platforms to not only find high-quality reviews, but also to recommend or prioritize them for better user engagement and sales optimization.

Research Questions

Sentiment, humor, sarcasm, and psycholinguistic features have been analyzed in the reviews using Python. In relation to this, the following questions will be answered:

1. Which of the linguistic and psychological features from LIWC have the strongest relationships with greater helpfulness ratings in Amazon reviews?
2. Does the use of humor or sarcasm in a review enhance or detract its perceived helpfulness?
3. What is the impact of sentiment (positive or negative) on helpful scores?
4. Is there a differentiating writing style (e.g., analytical, authentic, or confident) associated with helpful versus unhelpful reviews that form distinct differences?
5. Can a machine learning model be constructed (i.e., using logistic or linear regression) that determines whether a review is marked as helpful using LIWC features and textual patterns as a basis?
6. In what ways can the identification of humor and sarcasm enhance their ranking or display helpful reviews on e-commerce platforms such as Amazon?

Sponsors

The principal stakeholder guiding and providing insight during the definition of research problems and practical solutions is Professor Venugopal Balijepally. He had shaped to some extent the scope of the project by proposing certain factors such as the linguistic factors of the review helpfulness and suggesting the application of analytical methods of LIWC and machine learning for actionable insight generation. His comments have been instrumental in achieving the alignment of the research outcomes with the goals of academic and industry interests in e-commerce.

Data Collection and Analytical Methodology

Data Collection Sources and Process

The dataset for this project comprises reviews of products listed on Amazon, concentrating primarily on the consumer electronics products. This set of reviews contains metadata such as review title, review content, rating, and count of how many users found the review helpful. The data was obtained from a cleaned dataset of Amazon reviews that was set aside for linguistic and content analysis.

LIWC-22 or Linguistic Inquiry and Word Count software, which evaluates text for over 90 psychological, emotional, and structural language categories, was used to process the text data generated from the reviews to investigate the psycholinguistic features of the reviews.

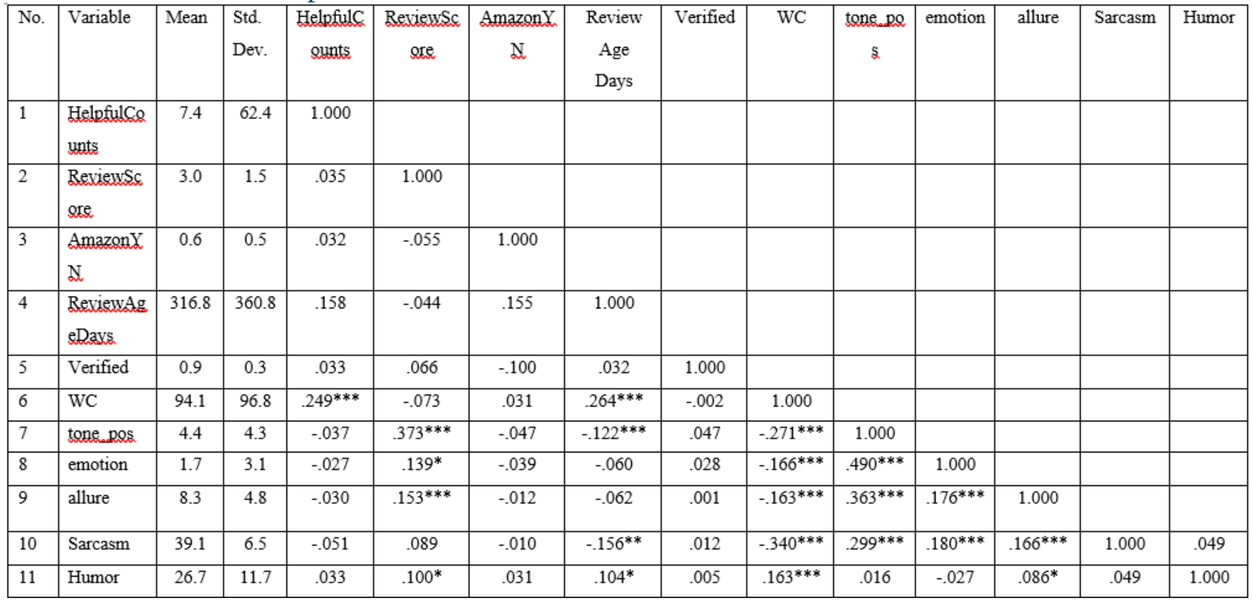
Tools Used

The different steps that include collection of data, cleansing as well as data analysis, modeling and data visualization utilize the following technologies and tools:

1. Apify.com Scraper – Serves one of the main functions of gathering data which is retrieving Amazon reviews for the time covering 2019 - 2025.
2. Tableau Prep
   a) Used for the purposes of data cleaning.
   b) Cleaning and converting review date fields.
   c) Created Calculated fields such as review age and review year.
   d) Finding and deleting all duplicate entries
   e) Removing blank fields
   f) Joining data formats for standards such as Numerical, Categorical, or even Dates
3. Utilizing Hugging Face Pretrained Models
   a) For identifying humor and sarcasm in Amazon reviews:
   b) Applied natural language understanding with transformers-based pretrained models.
   c) Deduced the percentage presence of sarcasm and humor within the review datasets.
4. LIWC-22 – Linguistic software for the inquiry and word count used for extracting title review and full review text to obtain psychological, cognitive, emotional, structural linguistic features.
5. SPSS – Used for building and interpreting regressions particularly determining the impact sentences and emotions have on how useful the reviews are perceived to be. SPSS allowed stepwise regression, coefficient testing, and evaluating regression models thus empowering robust statistical analysis.
6. Google Colab: It was the core environment where coding, documentation, and dataset analysis with Python was done in an iterative manner.
7. Excel: Utilized during the preliminary stages of the project for data inspection and the generation of summary statistics and validation checks for the data tables.

Summarizing Findings of Explanation/Prediction Models

Correlations and Descriptive Statistics



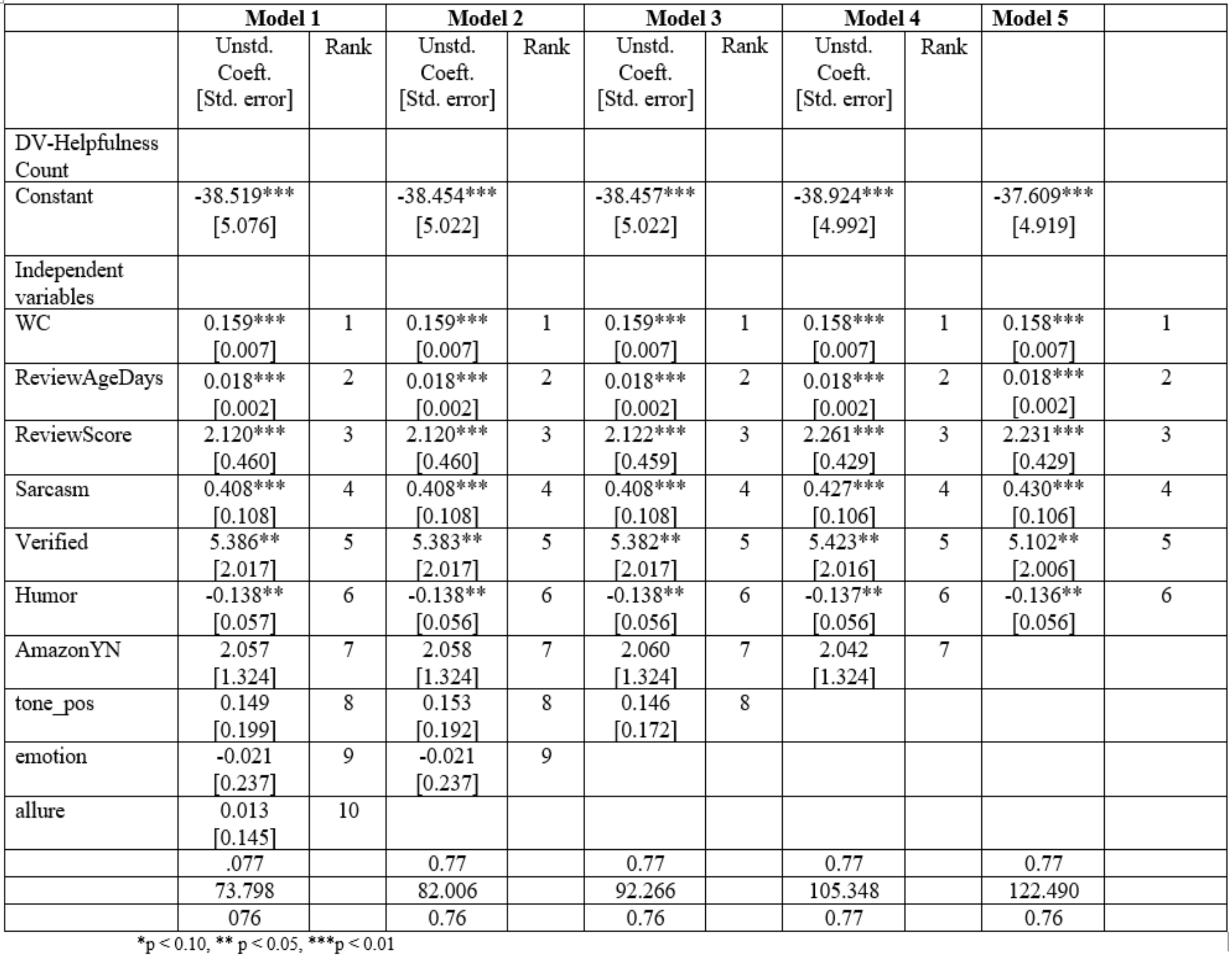
No.	Variable	Mean	Std. Dev.	HelpfulCounts	ReviewScore	AmazonYN	Review Age Days	Verified	WC	tone_pos	emotion	allure	Sarcasm	Humor
1	HelpfulCounts	7.4	62.4	1.000										
2	ReviewScore	3.0	1.5	.035	1.000									
3	AmazonYN	0.6	0.5	.032	-.055	1.000								
4	ReviewAgeDays	316.8	360.8	.158	-.044	.155	1.000							
5	Verified	0.9	0.3	.033	.066	-.100	.032	1.000						
6	WC	94.1	96.8	.249***	-.073	.031	.264***	-.002	1.000					
7	tone_pos	4.4	4.3	-.037	.373***	-.047	-.122***	.047	-.271***	1.000				
8	emotion	1.7	3.1	-.027	.139*	-.039	-.060	.028	-.166***	.490***	1.000			
9	allure	8.3	4.8	-.030	.153***	-.012	-.062	.001	-.163***	.363***	.176***	1.000		
10	Sarcasm	39.1	6.5	-.051	.089	-.010	-.156**	.012	-.340***	.299***	.180***	.166***	1.000	.049
11	Humor	26.7	11.7	.033	.100*	.031	.104*	.005	.163***	.016	-.027	.086*	.049	1.000

Table 1: Correlations and Descriptive Statistics

Table 1 shows correlations and summary statistics for all variables. Overall, the HelpfulCounts or the total number of votes considered helpful were rather low (M = 7.4, SD = 62.4) due to the extreme value in their distribution. Out of all variables, WC had the strongest correlation with HelpfulCounts (r = .249, p < .001) indicating that longer reviews are more likely to be considered helpful. Notably, tone_pos or positive tone of the review formed a negative correlation with helpfulness (r = –.271, p < .001) suggesting that overly positive phrases may lessen helpfulness. Likewise, sarcasm (r = –.340, p < .001) was negatively correlated but later in the analysis was able to positively correlate with helpfulness.

Explanation Model

Linear Regression



	Model 1		Model 2		Model 3		Model 4		Model 5	
	Unstd. Coeft. [Std. error]	Rank	Unstd. Coeft. [Std. error]	Rank	Unstd. Coeft. [Std. error]	Rank	Unstd. Coeft. [Std. error]	Rank		
DV-Helpfulness Count										
Constant	-38.519*** [5.076]		-38.454*** [5.022]		-38.457*** [5.022]		-38.924*** [4.992]		-37.609*** [4.919]	
Independent variables										
WC	0.159*** [0.007]	1	0.159*** [0.007]	1	0.159*** [0.007]	1	0.158*** [0.007]	1	0.158*** [0.007]	1
ReviewAgeDays	0.018*** [0.002]	2	0.018*** [0.002]	2	0.018*** [0.002]	2	0.018*** [0.002]	2	0.018*** [0.002]	2
ReviewScore	2.120*** [0.460]	3	2.120*** [0.460]	3	2.122*** [0.459]	3	2.261*** [0.429]	3	2.231*** [0.429]	3
Sarcasm	0.408*** [0.108]	4	0.408*** [0.108]	4	0.408*** [0.108]	4	0.427*** [0.106]	4	0.430*** [0.106]	4
Verified	5.386** [2.017]	5	5.383** [2.017]	5	5.382** [2.017]	5	5.423** [2.016]	5	5.102** [2.006]	5
Humor	-0.138** [0.057]	6	-0.138** [0.056]	6	-0.138** [0.056]	6	-0.137** [0.056]	6	-0.136** [0.056]	6
AmazonYN	2.057 [1.324]	7	2.058 [1.324]	7	2.060 [1.324]	7	2.042 [1.324]	7		
tone_pos	0.149 [0.199]	8	0.153 [0.192]	8	0.146 [0.172]	8				
emotion	-0.021 [0.237]	9	-0.021 [0.237]	9						
allure	0.013 [0.145]	10								
	.077		0.77		0.77		0.77		0.77	
	73.798		82.006		92.266		105.348		122.490	
	076		0.76		0.76		0.77		0.76	

*p < 0.10, ** p < 0.05, ***p < 0.01

Table 2: Linear Regression Model Summary

In the above table, Model 5 overall proved to be the best model among all others. This is based on the F-statistics of Model 5 which is the highest of 122.490 showing better overall fit and joint significance of the predictors. Even though R-square value (0.077) and adjusted R-square (0.076) remain constant across all models, the increase in F-statistic value from Model 1 to Model 5 suggests that the addition of variables improved the model’s performance without making it more complex or noisy. All predictors were added to Model 5. While some of the extra variables were individually insignificant, their overall inclusion provided the model with a better outcome, which shows that there is merit in adding them.

Models 1 through 3 all have only six base predictors; WC, ReviewAgeDays, ReviewScore, Sarcasm, Verified, and Humor. They rank and provide stable coefficients. Model 4 adds the variable AmazonYN but does not significantly change performance. Model 5 goes further by including tone_pos, emotion, and allure. While these variables are not significant on their own, their inclusion results in better F-statistics suggesting that they have some level of more variance in the data, However, this is still subtle.

In Model 5, a few variables are statistically significant and affect the dependent variable Helpfulness Count. The most influential variable is WC (Word Count) which has a coefficient of 0.158. This shows that longer reviews are considered more helpful because, probably, they offer more information. Next is ReviewAgeDays, where older reviews have higher helpfulness counts, probably due to longer exposure time. Third is ReviewScore where it shows that higher product ratings receive more helpful votes, and this might be the result of general consumer sentiment. Sarcasm is the fourth most important variable and is positively significant, suggesting that sarcasm helps to add some authenticity or personality that many readers appreciate. Fifth is verified purchase status and is also positively associated with helpfulness, probably because it is perceived as more credible. Finally, Humor, although significant, has a negative effect, meaning humorous reviews are perceived as less serious and thus lose ratings in perceived helpfulness.

To summarize, Model 5 is, by far, the best performing model because of its robust statistical performance and its inclusion of virtually all relevant factors. The most important determinants of review helpfulness are the factors word count, age of the review, score given in the review, sarcastic content, whether the review was verified, and humor since they all illustrate various aspects that can either enhance or diminish a review's utility to its readers.

Research Evaluation

The completed deliverables and research findings offer valuable insights for both e-commerce platforms and businesses that rely on user-generated content to influence consumer behavior. One of the key challenges in the online retail industry is helping customers find which product reviews are genuinely useful and trustworthy. This research directly addresses that issue by identifying the specific features that make reviews more likely to be rated as helpful. For example, reviews from verified purchasers, those with moderate to high review scores, and those that are longer and more detailed are consistently perceived as more helpful. Additionally, the findings highlight that emotional or overly positive language, as well as humor, can reduce perceived credibility, which can mislead or confuse consumers. These insights can help platforms like Amazon enhance their recommendation algorithms, prioritize high-quality content, and design review guidelines that encourage clarity, objectivity, and depth. For companies, this research can inform strategies on how to ask for and manage customer reviews to positively influence purchasing decisions. Overall, the study helps tackle the industry-wide problem of information overload and quality inconsistency in online reviews by offering data-driven criteria for review usefulness.

Lessons Learned

This research project has enhanced my skills in technical data analysis and decision-making based on data. I relied on linear regression techniques to identify relationships between variables, which taught me how critical proper assumption validation is to obtain meaningful model results. This experience taught me to evaluate model effectiveness as well as the limitations of various analytic techniques. I furthered my understanding of how products are perceived by consumers through emotional tone, humor, sarcasm, and even more complex linguistic aspects—elements that often have indirect, yet powerful, impact on results. I learned how to work with real-world data sets, which are typically dirty, inconsistent, and incomplete. Most notably, this project demonstrated how advanced analytics can reveal meaningful patterns in data that directly tend to industry issues, such as enhancing review systems and fostering trust across online marketplaces. The project was instrumental in showing me the importance of analytical approaches to problem solving and emphasizing the need for data literacy in our society.

Limitations & Future Work

Although the study offered valuable understanding of the determinants of online reviews, several limitations need to be considered. To begin with, data collection was performed on a single platform and a specific product category. This might restrict the extent to which the findings can be applied to different sectors or platforms. In addition, the quantitative approach employed in this analysis alongside the available linguistic features placed significant constraints on the analysis, potentially obscuring the nuanced context of the reviewer's intent and the consumer's interpretation. The models also overlook the impacts of popularity and seasonal trends alongside reviews, suggesting external influences on perceived helpfulness. From a methodological standpoint, negative binomial regression, while addressing overdispersion, still leaves room in the construct of hierarchical or mixed effect models.

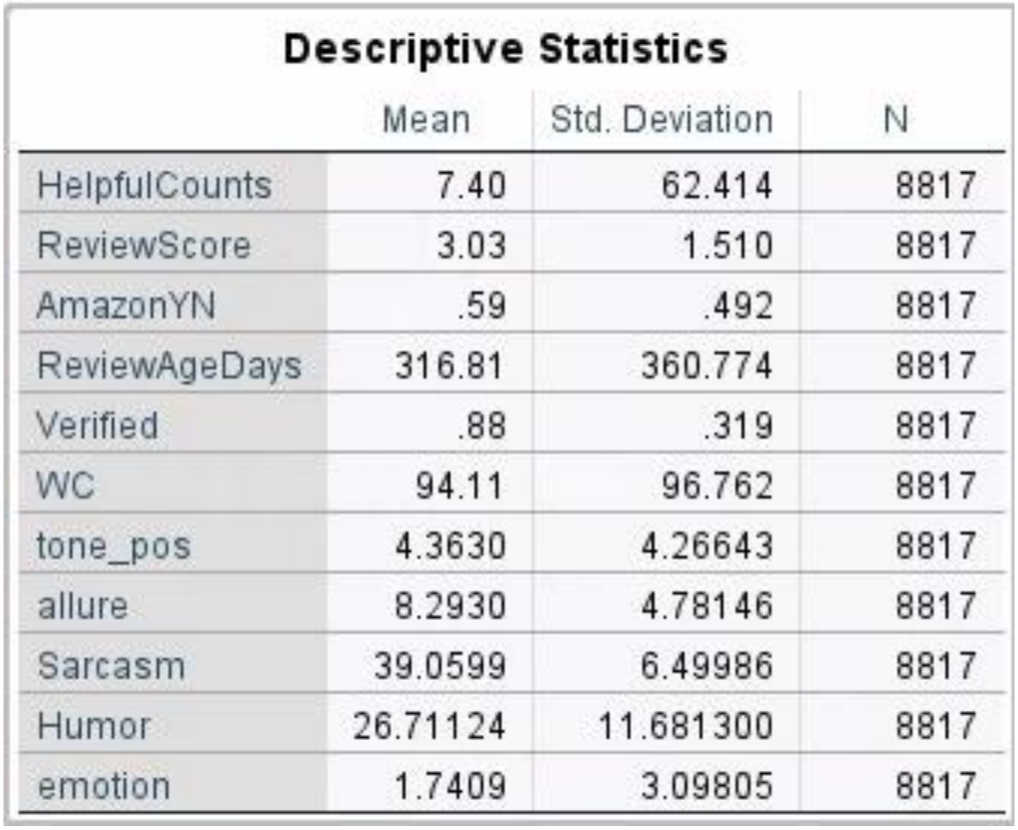
References (Khan n.d.)

1. Feng, X. 2020. "An analysis of online review helpfulness that integrates content and function words: An application to online reviews on Amazon.com. Marketing Science." *The Japan Society of Marketing and Distributio* 28(1), 49–68.
2. Haque, M. E., Tozal, M. E., & Islam, A. 2018. "Helpfulness prediction of online product reviews. ." *Proceedings of the 18th ACM Symposium on Document Engineering (DocEng '18)* 1–4.
3. Hjalmarsson, F. 2021. "Predicting the helpfulness of online product reviews." *Blekinge Institute of Technology.*
4. Kim, S.-M., Pantel, P., Chklovski, T., & Pennacchiotti, M. 2006. " Automatically assessing review helpfulness." *Conference on Empirical Methods in Natural Language Processing.* Association for Computatio. 423–430.
5. Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., & Roy, P. K. 2017. ". Predicting the “helpfulness” of online consumer reviews." *Journal of Business Research* 346–355.

Appendixes

Appendix A: Linear Regression SPSS Output

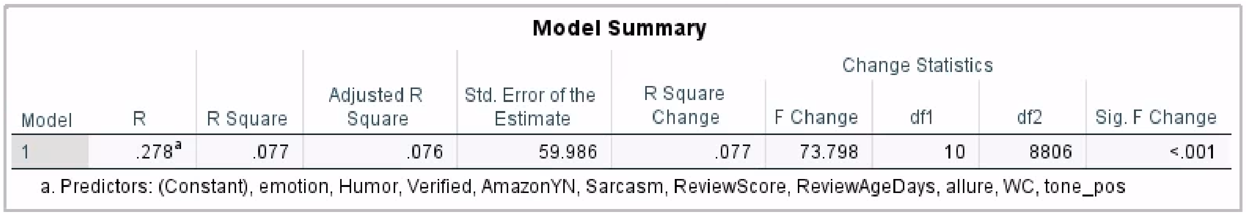
1. Descriptive Statistics



Descriptive Statistics

	Mean	Std. Deviation	N
HelpfulCounts	7.40	62.414	8817
ReviewScore	3.03	1.510	8817
AmazonYN	.59	.492	8817
ReviewAgeDays	316.81	360.774	8817
Verified	.88	.319	8817
WC	94.11	96.762	8817
tone_pos	4.3630	4.26643	8817
allure	8.2930	4.78146	8817
Sarcasm	39.0599	6.49986	8817
Humor	26.71124	11.681300	8817
emotion	1.7409	3.09805	8817

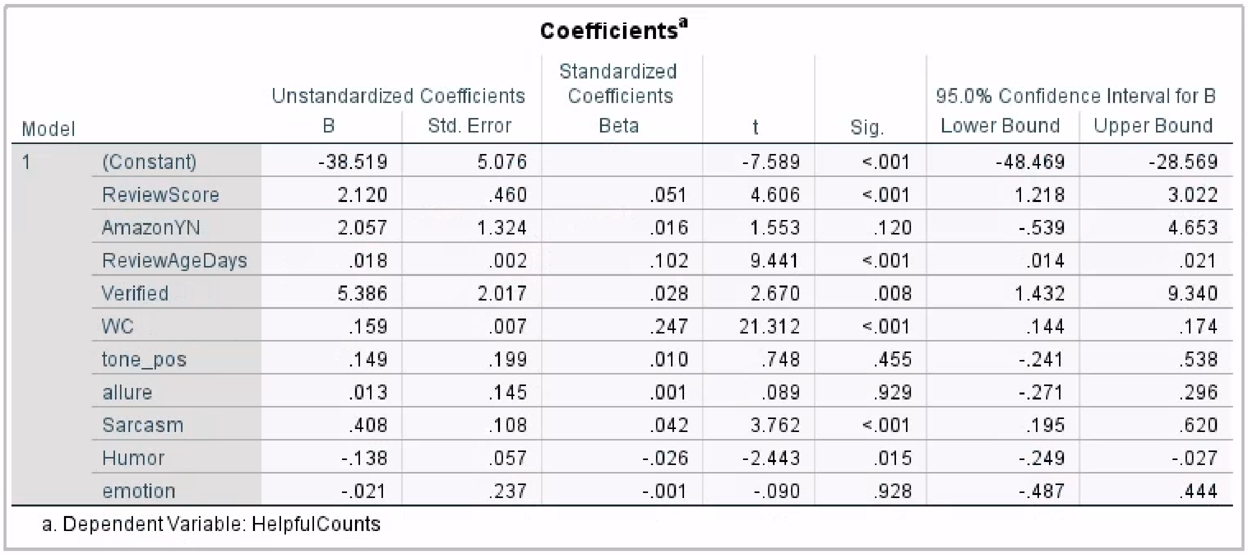
2. Model 1 Output



Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.278[a]	.077	.076	59.986	.077	73.798	10	8806	<.001

a. Predictors: (Constant), emotion, Humor, Verified, AmazonYN, Sarcasm, ReviewScore, ReviewAgeDays, allure, WC, tone_pos

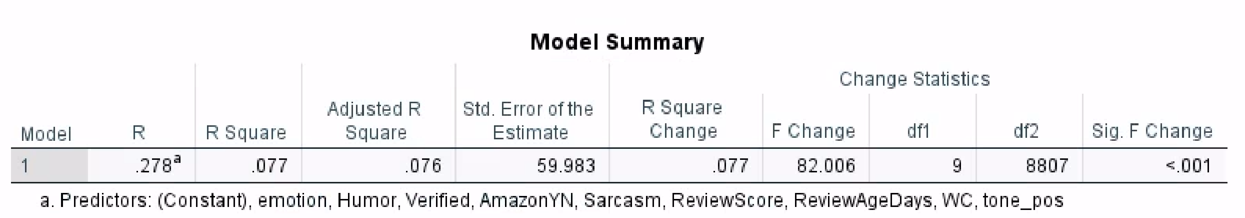


Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-38.519	5.076		-7.589	<.001	-48.469	-28.569
	ReviewScore	2.120	.460	.051	4.606	<.001	1.218	3.022
	AmazonYN	2.057	1.324	.016	1.553	.120	-.539	4.653
	ReviewAgeDays	.018	.002	.102	9.441	<.001	.014	.021
	Verified	5.386	2.017	.028	2.670	.008	1.432	9.340
	WC	.159	.007	.247	21.312	<.001	.144	.174
	tone_pos	.149	.199	.010	.748	.455	-.241	.538
	allure	.013	.145	.001	.089	.929	-.271	.296
	Sarcasm	.408	.108	.042	3.762	<.001	.195	.620
	Humor	-.138	.057	-.026	-2.443	.015	-.249	-.027
	emotion	-.021	.237	-.001	-.090	.928	-.487	.444

a. Dependent Variable: HelpfulCounts

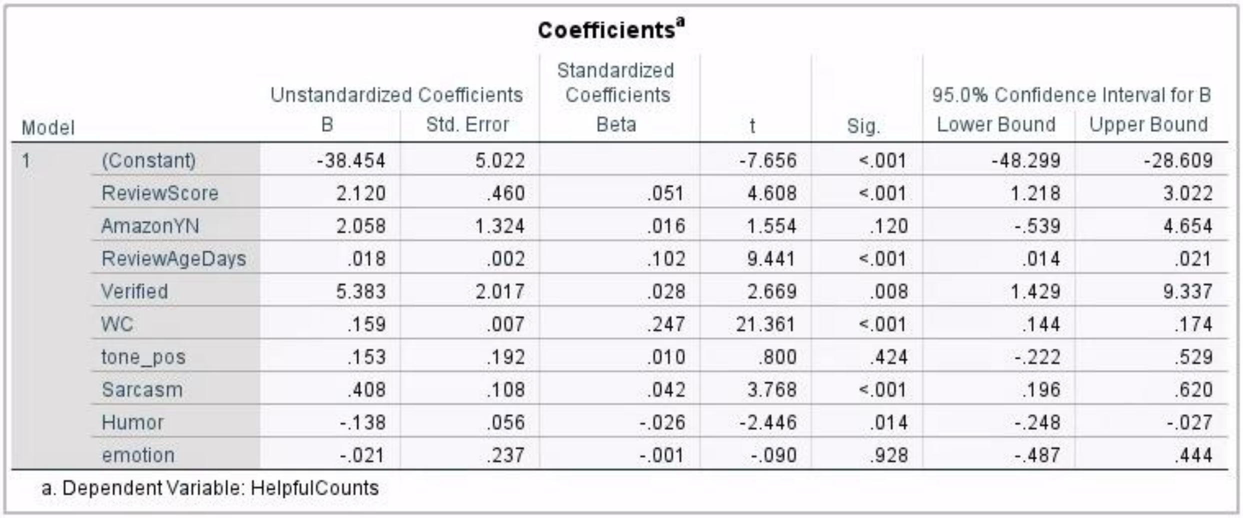
3. Model 2 Output



Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.278[a]	.077	.076	59.983	.077	82.006	9	8807	<.001

a. Predictors: (Constant), emotion, Humor, Verified, AmazonYN, Sarcasm, ReviewScore, ReviewAgeDays, WC, tone_pos

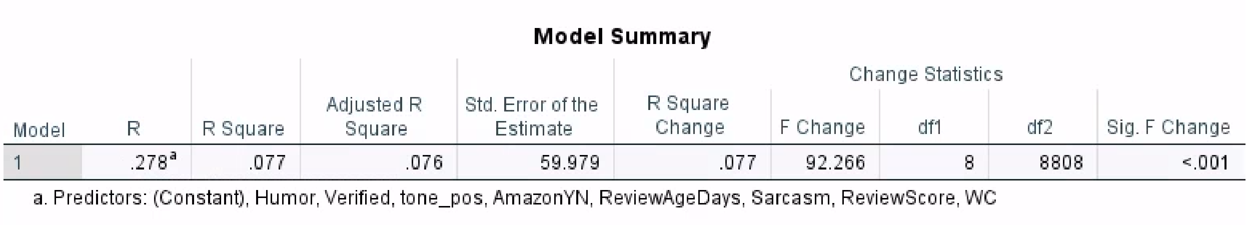


Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-38.454	5.022		-7.656	<.001	-48.299	-28.609
	ReviewScore	2.120	.460	.051	4.608	<.001	1.218	3.022
	AmazonYN	2.058	1.324	.016	1.554	.120	-.539	4.654
	ReviewAgeDays	.018	.002	.102	9.441	<.001	.014	.021
	Verified	5.383	2.017	.028	2.669	.008	1.429	9.337
	WC	.159	.007	.247	21.361	<.001	.144	.174
	tone_pos	.153	.192	.010	.800	.424	-.222	.529
	Sarcasm	.408	.108	.042	3.768	<.001	.196	.620
	Humor	-.138	.056	-.026	-2.446	.014	-.248	-.027
	emotion	-.021	.237	-.001	-.090	.928	-.487	.444

a. Dependent Variable: HelpfulCounts

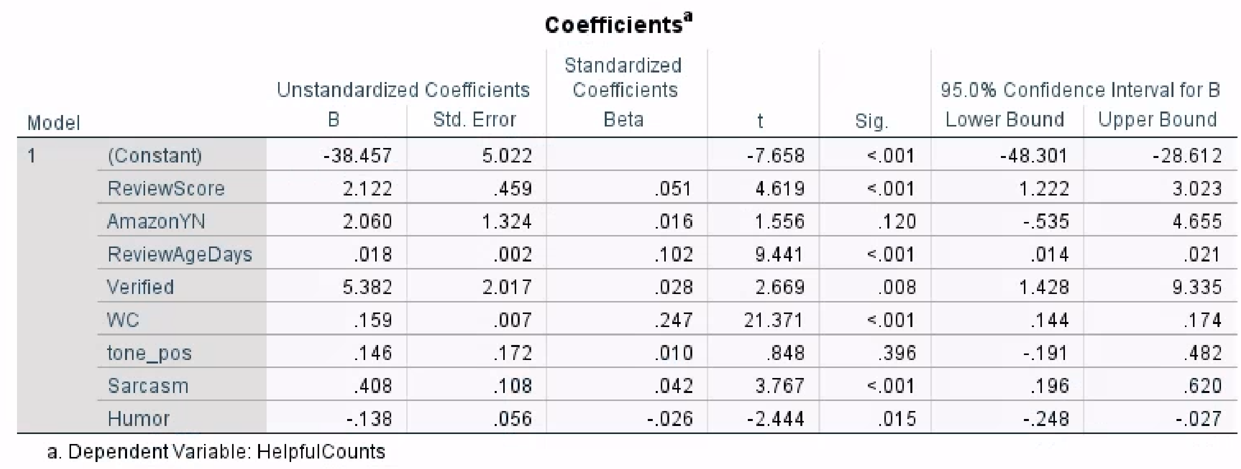
4. Model 3 Output



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics R Square Change	F Change	df1	df2	Sig. F Change
1	.278[a]	.077	.076	59.979	.077	92.266	8	8808	<.001

a. Predictors: (Constant), Humor, Verified, tone_pos, AmazonYN, ReviewAgeDays, Sarcasm, ReviewScore, WC

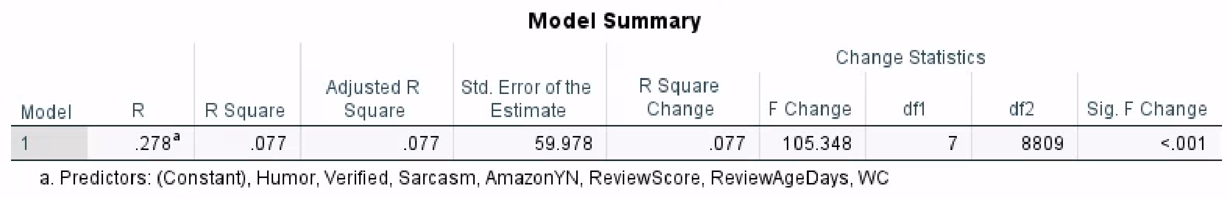


Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-38.457	5.022		-7.658	<.001	-48.301	-28.612
	ReviewScore	2.122	.459	.051	4.619	<.001	1.222	3.023
	AmazonYN	2.060	1.324	.016	1.556	.120	-.535	4.655
	ReviewAgeDays	.018	.002	.102	9.441	<.001	.014	.021
	Verified	5.382	2.017	.028	2.669	.008	1.428	9.335
	WC	.159	.007	.247	21.371	<.001	.144	.174
	tone_pos	.146	.172	.010	.848	.396	-.191	.482
	Sarcasm	.408	.108	.042	3.767	<.001	.196	.620
	Humor	-.138	.056	-.026	-2.444	.015	-.248	-.027

a. Dependent Variable: HelpfulCounts

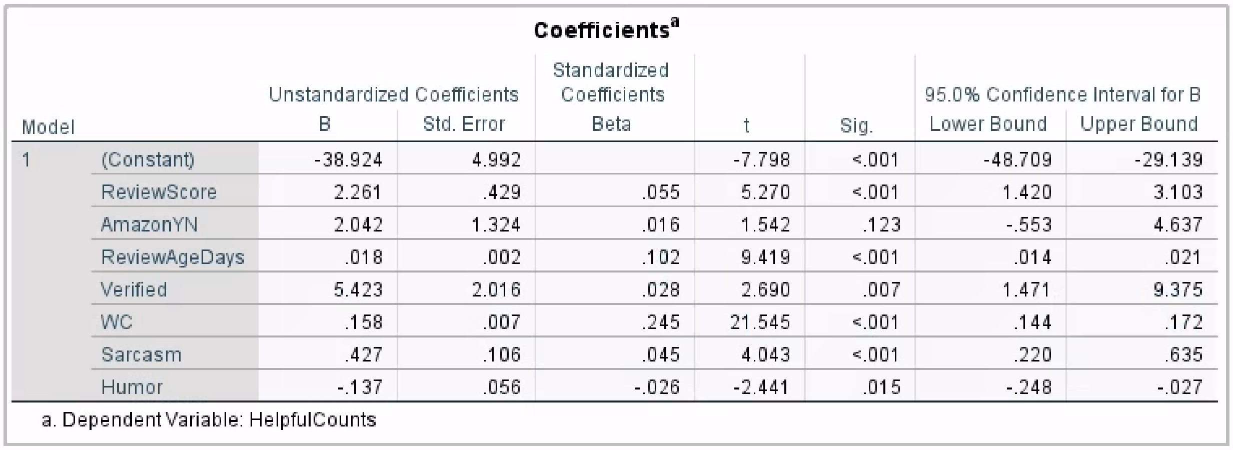
5. Model 4 Output



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics R Square Change	F Change	df1	df2	Sig. F Change
1	.278[a]	.077	.077	59.978	.077	105.348	7	8809	<.001

a. Predictors: (Constant), Humor, Verified, Sarcasm, AmazonYN, ReviewScore, ReviewAgeDays, WC

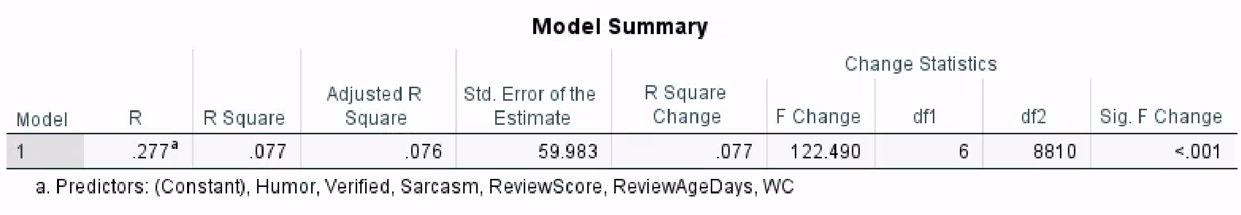


Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-38.924	4.992		-7.798	<.001	-48.709	-29.139
	ReviewScore	2.261	.429	.055	5.270	<.001	1.420	3.103
	AmazonYN	2.042	1.324	.016	1.542	.123	-.553	4.637
	ReviewAgeDays	.018	.002	.102	9.419	<.001	.014	.021
	Verified	5.423	2.016	.028	2.690	.007	1.471	9.375
	WC	.158	.007	.245	21.545	<.001	.144	.172
	Sarcasm	.427	.106	.045	4.043	<.001	.220	.635
	Humor	-.137	.056	-.026	-2.441	.015	-.248	-.027

a. Dependent Variable: HelpfulCounts

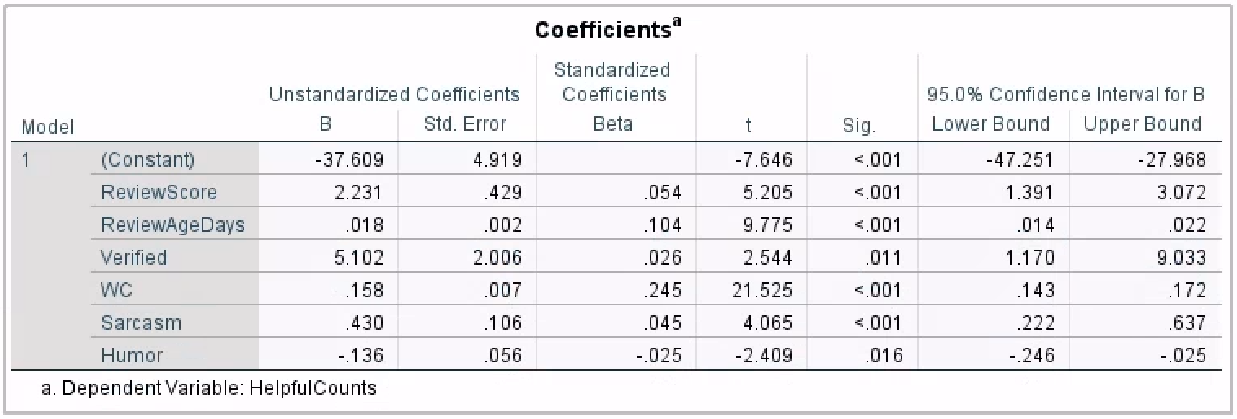
6. Model 5 Output



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.277[a]	.077	.076	59.983	.077	122.490	6	8810	<.001

a. Predictors: (Constant), Humor, Verified, Sarcasm, ReviewScore, ReviewAgeDays, WC

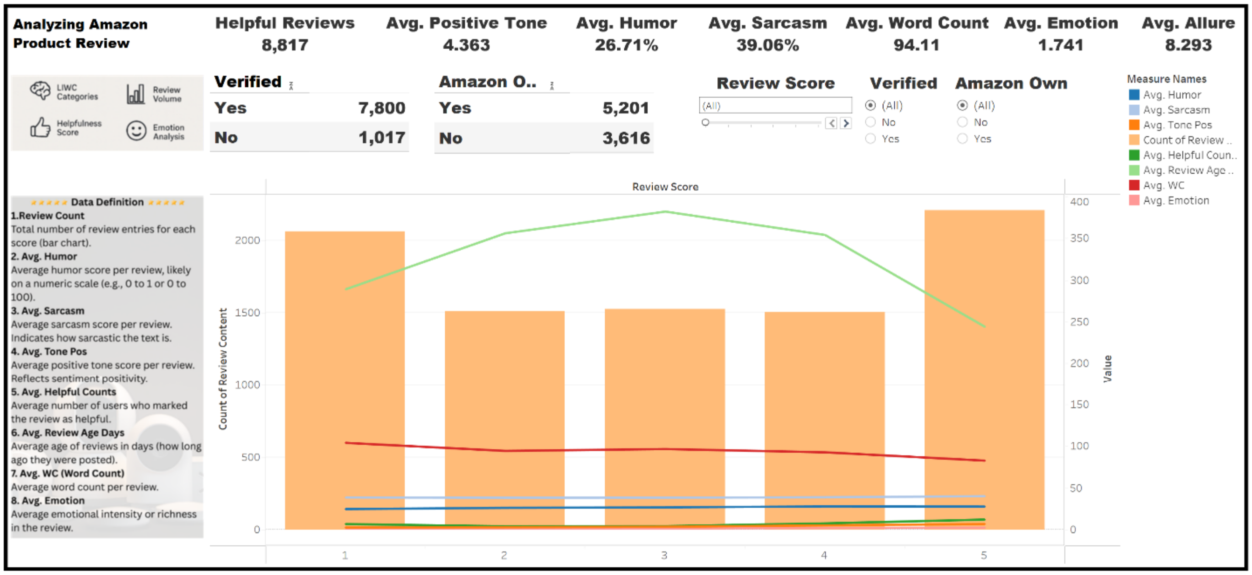


Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-37.609	4.919		-7.646	<.001	-47.251	-27.968
	ReviewScore	2.231	.429	.054	5.205	<.001	1.391	3.072
	ReviewAgeDays	.018	.002	.104	9.775	<.001	.014	.022
	Verified	5.102	2.006	.026	2.544	.011	1.170	9.033
	WC	.158	.007	.245	21.525	<.001	.143	.172
	Sarcasm	.430	.106	.045	4.065	<.001	.222	.637
	Humor	-.136	.056	-.025	-2.409	.016	-.246	-.025

a. Dependent Variable: HelpfulCounts

Appendix B: Tableau Dashboard



Analyzing Amazon Product Review

Helpful Reviews	Avg. Positive Tone	Avg. Humor	Avg. Sarcasm	Avg. Word Count	Avg. Emotion	Avg. Allure
8,817	4.363	26.71%	39.06%	94.11	1.741	8.293

Verified	
Yes	7,800
No	1,017

Amazon O..	
Yes	5,201
No	3,616

Review Score: (All)

Verified: ● (All) ○ No ○ Yes

Amazon Own: ● (All) ○ No ○ Yes

Data Definition

1.Review Count
Total number of review entries for each score (bar chart).

2. Avg. Humor
Average humor score per review, likely on a numeric scale (e.g., 0 to 1 or 0 to 100).

3. Avg. Sarcasm
Average sarcasm score per review. Indicates how sarcastic the text is.

4. Avg. Tone Pos
Average positive tone score per review. Reflects sentiment positivity.

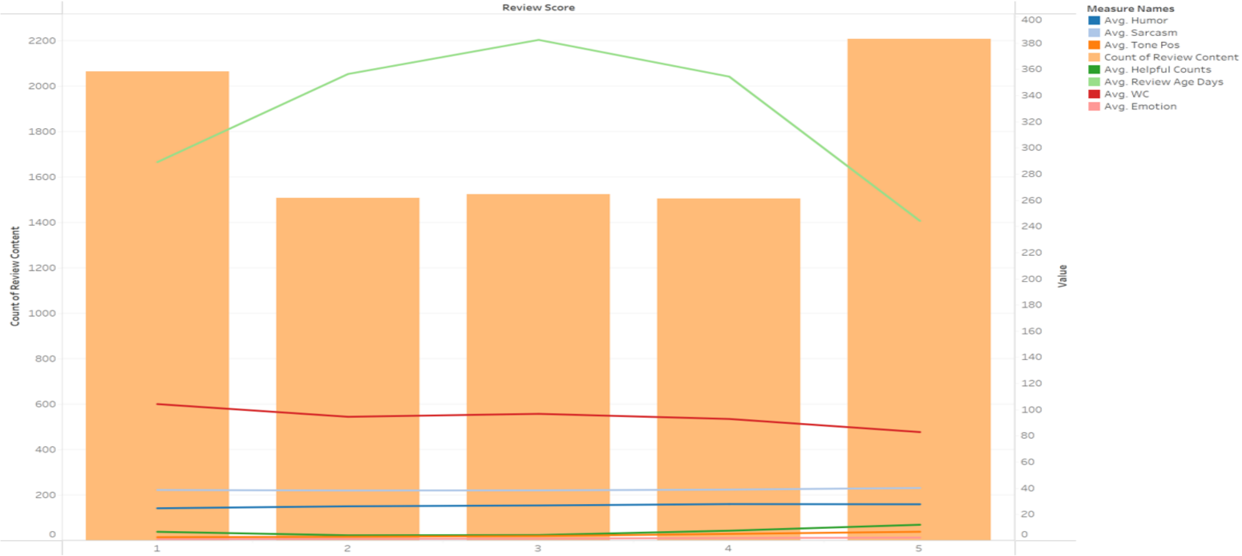
5. Avg. Helpful Counts
Average number of users who marked the review as helpful.

6. Avg. Review Age Days
Average age of reviews in days (how long ago they were posted).

7. Avg. WC (Word Count)
Average word count per review.

8. Avg. Emotion
Average emotional intensity or richness in the review.

Combo Variable Graph



The trends of count of Review Content, count of Review Content, Avg. Humor, Avg. Sarcasm, Avg. Tone Pos, Avg. Helpful Counts, Avg. Review Age Days, Avg. WC and Avg. Emotion for Review Score. Color shows details about count of Review Content, Avg. Humor, Avg. Sarcasm, Avg. Tone Pos, Avg. Helpful Counts, Avg. Review Age Days, Avg. WC and Avg. Emotion. The data is filtered on Verified and Amazon Own. The Verified filter keeps No and Yes. The Amazon Own filter keeps No and Yes. The view is filtered on Review Score, which keeps 1, 2, 3, 4 and 5.