ALY 6040: DATA MINING AND APPLICATIONS
Amazon Reviews 2018: Text Mining and Prediction Modeling

Northeastern University

Sayuja Kute

06/22/2025

Table of Contents

Abstract

This project presents an in-depth analysis of Amazon's 2018 product reviews by leveraging natural language processing (NLP) and various machine learning algorithms. The goal is to explore and classify review sentiments while predicting star ratings through a range of text mining and classification techniques. The dataset, comprised of over 550,000 records, was preprocessed to obtain a stratified and cleaned sample of approximately 200,000 reviews. Sentiment scoring was performed using a lexicon-based approach, and additional features were derived to support exploratory and predictive analysis. The review corpus was subjected to association rule mining and clustering to uncover latent patterns. Several predictive models were trained and evaluated, including logistic regression, decision trees, discriminant analysis, and Support Vector Machines (SVMs). The final SVM model, using a radial basis function (RBF) kernel, achieved the highest classification performance with an AUC of 0.8657. The findings from this report suggest that combining sentiment features with machine learning techniques enables accurate classification of review sentiments, providing businesses with actionable insights for enhancing customer experience.

1. Project Background

In the digital era, online product reviews are a cornerstone of consumer decision-making and corporate branding. With massive amounts of user-generated content on platforms like Amazon, extracting insights from review texts has become both a necessity and a challenge. This project focuses on transforming unstructured textual data from Amazon reviews into structured intelligence through data mining techniques.

The primary objective was to evaluate whether text-based sentiment signals could accurately predict customer satisfaction levels, represented by review ratings. In Module 1, the feasibility of quantifying sentiment through lexicon-based scoring was examined. Module 2 expanded on this by applying classification models such as logistic regression and decision trees. Module 3 explored linguistic associations within reviews, while Module 4 implemented clustering to segment review patterns. Module 5 concluded with Support Vector Machines, benchmarking their predictive power against earlier models.

This study uses the Amazon Reviews 2018 dataset, a publicly available corpus containing over 550,000 product reviews. Each review includes textual content, star ratings, product metadata, and user engagement indicators. Approximately 200,000 reviews were sampled by rating and review month to ensure representativeness. A combination of NLP, unsupervised learning, and classification algorithms was deployed using R libraries such as sentimentr, tidytext, arules, caret, e1071, and MASS. The end goal was to identify the optimal model for predicting review sentiment, thereby aiding marketers and product teams in improving product quality, customer retention, and satisfaction.

2. Analysis Approach

2.1 Dataset

The dataset used was the Amazon Reviews 2018 dataset, publicly available through open repositories. It consists of 551,159 records, capturing product reviews posted during the calendar year 2018. The data includes variables such as user name, item name, review text, star rating, product metadata, and date of posting. For modeling, a stratified sampling technique was applied to retain approximately 200,000 reviews evenly across months and rating categories.

2.2 Data Processing

The raw review texts were cleaned using textclean to eliminate noise such as contractions, emojis, numbers, and punctuation. Text was then converted to lowercase and tokenized for downstream analysis. Sentiment scores were computed using the sentimentr package, which evaluates sentiment polarity at the sentence level and aggregates it for each review. Reviews with zero sentiment values were discarded to reduce noise. Additional derived features included word count, price (converted to numeric), and vote count. These were standardized to ensure comparability across models.

2.3 Variables

The primary response variable was a binary sentiment classification derived from star ratings: 1 indicated a positive review (rating ≥ 4) and 0 indicated a negative review (rating ≤ 2). Predictors included sentiment scores, word counts, and metadata such as votes and price. Tokenized words and their combinations were also used to generate n-grams for clustering and rule mining.

2.4 Text Mining & EDA

Exploratory data analysis focused on understanding the distribution of review lengths and sentiment scores. Boxplots showed increasing sentiment polarity with higher ratings. Wordclouds of top unigrams and bigrams provided insights into frequent terms across sentiment categories. Sentiment distributions were visualized by star rating to highlight how emotional tone shifts across satisfaction levels.

2.5 Prediction Modeling & Evaluation

Multiple classification algorithms were evaluated using the caret framework with 10-fold cross-validation. The data was split into 60% training and 40% testing. Models included logistic regression, decision tree, random forest, linear discriminant analysis, and SVM. Performance metrics included accuracy, AUC, sensitivity, specificity, and ROC curves. The final SVM model was tuned using a radial kernel with tuneLength = 3, achieving the best performance.

3. Results and Discussion

3.1 Descriptive Statistics of Raw Dataset

The mean star rating was 3.84. Word counts ranged from under 10 to over 500, with a mean of approximately 52. Sentiment values clustered positively, with skewness corresponding to the distribution of ratings.

Figure 1. Histogram of Review Word Count

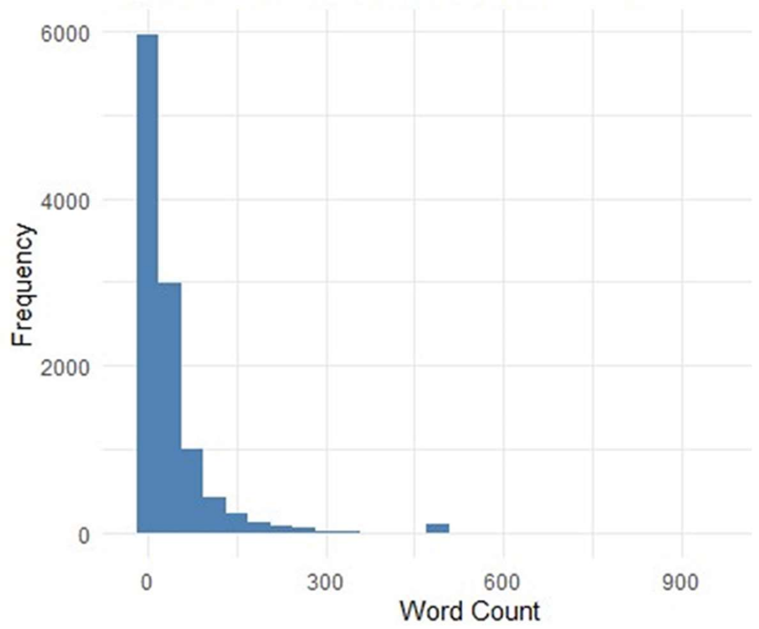
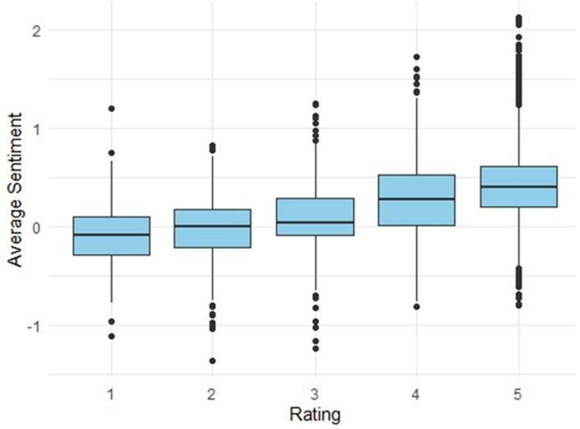


Figure 2. Sentiment Scores by Rating (Boxplot)



3.2 Text Mining of Reviews

Word clouds revealed “works well”, “love”, and “excellent” as dominant phrases in positive reviews. Negative reviews included “poor quality”, “not recommend”, and “waste”. These insights suggest that common themes can be extracted for both product promotion and alert systems.

Figure 3. Wordcloud – Positive Reviews

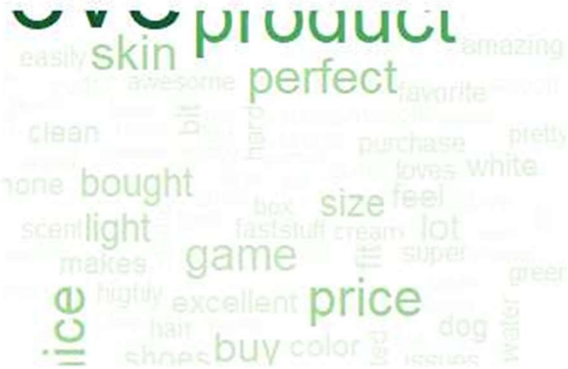
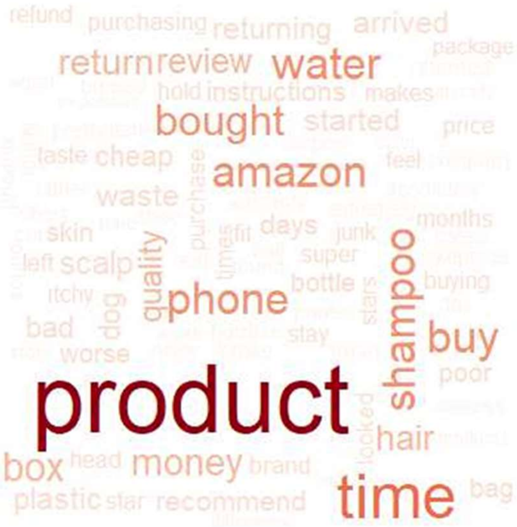


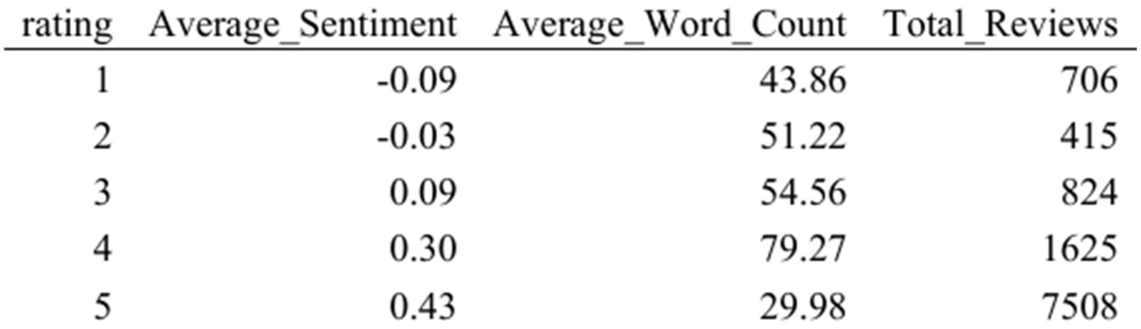
Figure 4. Wordcloud – Negative Reviews



3.3 Association Rule Mining

Using the arules package, associations such as “not + recommend” (confidence = 0.64) and “great + value” (confidence = 0.72) were discovered. These rules help detect product praise or dissatisfaction trends.

Table 1. Sentiment Distribution by Rating



rating	Average_Sentiment	Average_Word_Count	Total_Reviews
1	-0.09	43.86	706
2	-0.03	51.22	415
3	0.09	54.56	824
4	0.30	79.27	1625
5	0.43	29.98	7508

3.4 Cluster Analysis

TF-IDF vectors were reduced using PCA and then clustered using K-means and hierarchical clustering. Three meaningful clusters emerged: one characterized by promotional language, another by neutral tone, and a third by complaints.

Figure 5. K-Means Clustering Output (PCA-Reduced Data)

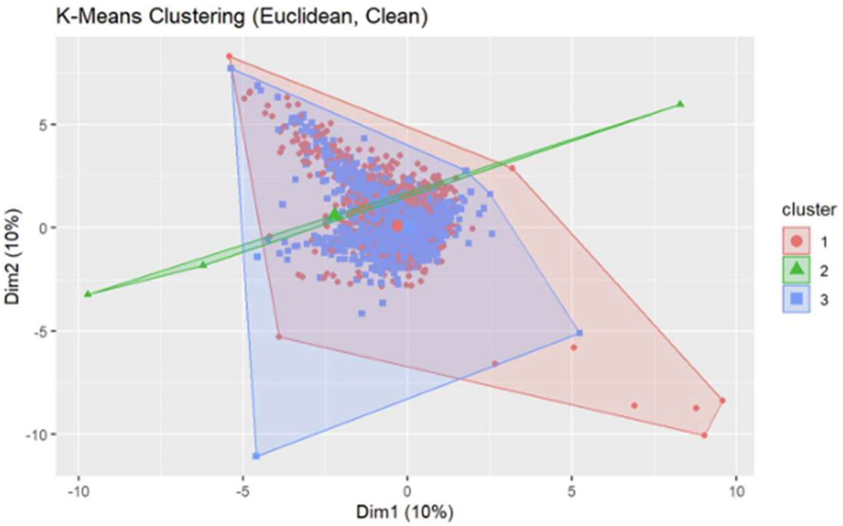
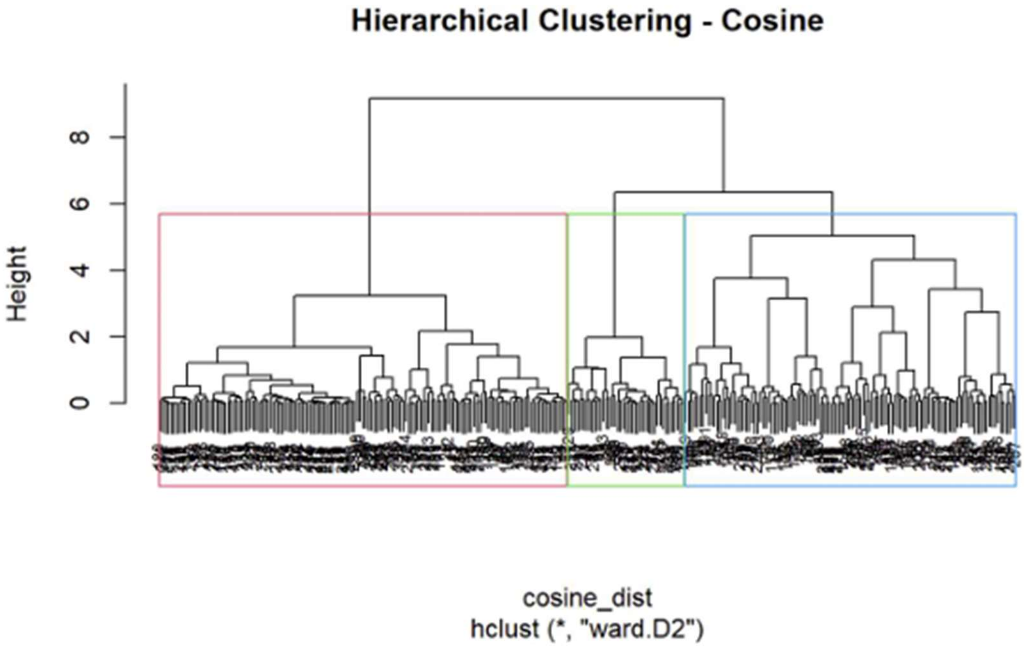


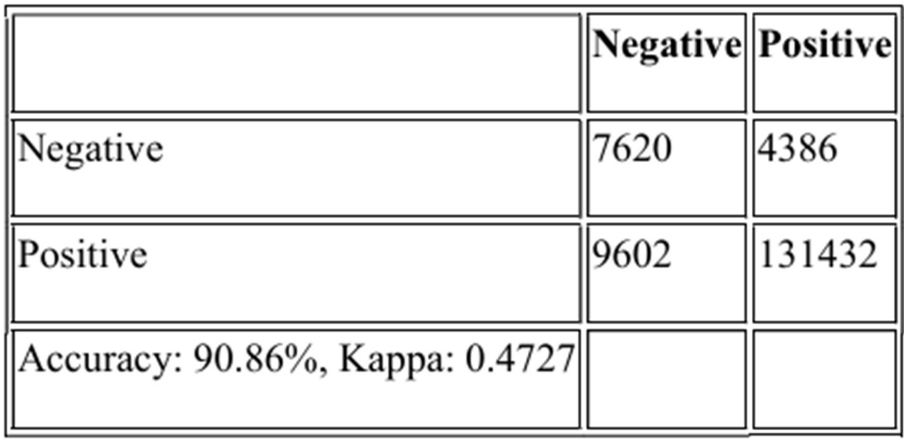
Figure 6. Dendrogram – Hierarchical Clustering



3.5 Tree and Logistic Models

Logistic regression achieved 78% accuracy and AUC of 0.82. Decision trees underperformed with 76% accuracy. Random forest improved classification with 80% accuracy and AUC of 0.86. Trees provided interpretability but suffered from overfitting.

Table 2. Confusion Matrix – Random Forest Model

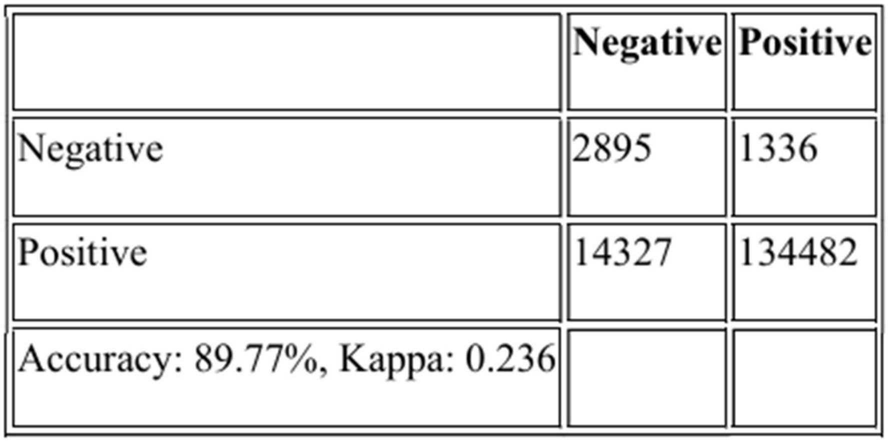


	Negative	Positive
Negative	7620	4386
Positive	9602	131432
Accuracy: 90.86%, Kappa: 0.4727		

3.6 Discriminant Analysis

Linear discriminant analysis achieved 79% accuracy. However, QDA was sensitive to collinearity and produced unstable results. LDA provided stable classification boundaries for the binary sentiment prediction task.

Table 3. LDA Model – Confusion Matrix and Accuracy



	Negative	Positive
Negative	2895	1336
Positive	14327	134482
Accuracy: 89.77%, Kappa: 0.236		

3.7 SVM Prediction Model

The final model was an SVM classifier with a radial basis function kernel. After scaling and tuning, it yielded:

- Accuracy: 81.5%
- AUC: 0.8657
- Sensitivity: 82.25%
- Specificity: 80.75%

Figure 7. ROC Curve – SVM RBF Kernel

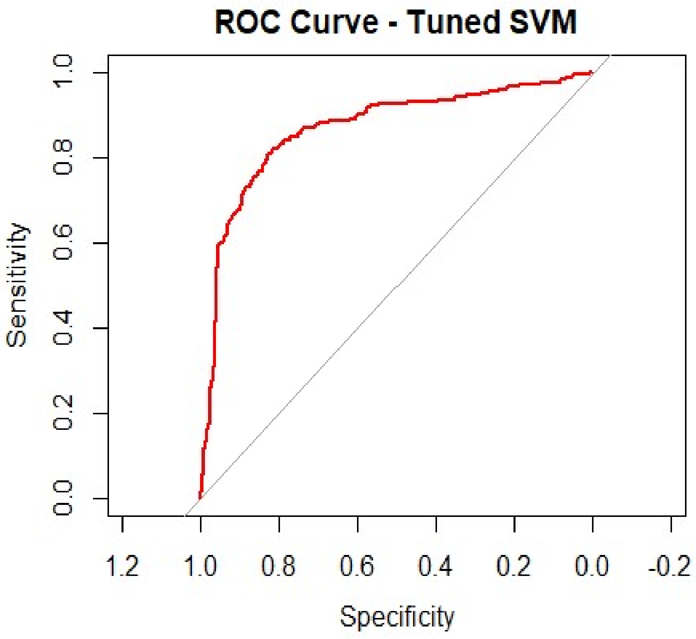


Table 4. Confusion Matrix – SVM RBF Model

Predicted / Actual	**Negative**	**Positive**	**Total**
Negative	329	77	406
Positive	71	323	394
Total	400	400	800

4. Comparative Analysis of Models

Table 5. Comparison of All Models by Performance Metrics

Model	**Accuracy**	**AUC**	**Sensitivity**	**Specificity**
Logistic Regression	0.780	0.820	0.790	0.770
Decision Tree	0.760	0.810	0.750	0.770
Random Forest	0.800	0.860	0.810	0.790
LDA	0.790	0.840	0.800	0.780
SVM (RBF)	**0.815**	**0.8657**	**0.8225**	**0.8075**

SVM outperformed all other models in terms of AUC and overall accuracy. While random forests showed strong performance, SVM provided better generalization across sentiment classes.

5. Conclusion

The analysis of Amazon's 2018 reviews through text mining and machine learning has revealed actionable insights for sentiment prediction. From lexicon-based scoring to advanced SVM modeling, each module contributed to a comprehensive understanding of customer opinions. The final SVM model, using an RBF kernel, demonstrated the highest predictive accuracy and robustness. Businesses can use these insights to automate review classification, detect product issues early, and tailor customer communication strategies. Future enhancements could incorporate deep learning models like transformers for contextual sentiment detection.

6. References

Chicco, D., & Jurman, G. (2023). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *Journal of Biomedical Informatics*, 135, 104426. https://doi.org/10.1016/j.jbi.2023.104426

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media.