NANYANG TECHNOLOGICAL UNIVERSITY

WEE KIM WEE SCHOOL OF COMMUNICATION AND INFORMATION

IS6750 Social Media Analytics

Mai Mingshan G2503141E
Liu Menglu G2503020K
Zhou Ziqi G2503378D

Selected files: Amazon Instant Video

Introduction

In the rapidly evolving e-commerce ecosystem, online reviews have become a crucial source of information for consumers' decisions and product iterations. Compared with traditional advertising or professional reviews, online reviews are regarded as more authentic and valuable "social evidence", which can significantly influence potential consumers' expectations of product quality and experience. Among various industries, the video game market stands out for its highly engaged user base and subjective experiences, which are often vividly expressed through customer reviews. A successful game is often closely related to the emotional feedback and experience perception of the player community. Therefore, conducting effective and large-scale analysis of these comments can provide valuable references for game developers and marketing personnel to improve game design and user engagement. As one of the largest e-commerce platforms in the world, Amazon has accumulated a vast amount of user review data, providing a representative foundation for in-depth research on user behavior and emotional tendencies in this market.

However, manual reading and classification of such comments are both time-consuming and subjective, making it difficult to implement on large-scale data. It is necessary to introduce natural language processing and machine learning technologies to automatically identify the sentiment polarity of comment texts. Our research lies in quantifying and classifying the emotional tendencies present in video game reviews. We utilized the "Amazon Review Data 2018" dataset provided by the Jianmo Ni team from the University of California, San Diego (UCSD), and selected the "Video Games" category subset from it as the research object. This dataset contains a wealth of key fields; based on this characteristic, it is necessary to adopt automated natural language processing technology.

In terms of model selection, this study adopted three classic supervised learning algorithms: Support Vector Machine (SVM), Naive Bayes, and Random Forest. SVM divides different sentiment categories in a high-dimensional space by constructing the optimal hyperplane and usually performs stably on sparse text features. Naive Bayes conducts probability inference based on the assumption of conditional independence; Random Forest enhances the generalization ability of the model by integrating multiple decision trees. By comparing the accuracy, balance accuracy, sensitivity (recall rate), and specificity of these models, this study aims to identify the most suitable model for handling the task of e-commerce review sentiment classification and reveal the performance differences of different algorithms when facing category imbalance (where positive reviews far outweigh negative ones).

To enhance the interpretability and reliability of the model results, this study also introduces the sentiment dictionary method based on Hu & Liu dictionaries for analysis, and conducts statistics and scoring of positive and negative words in the comment texts. By comparing the sentiment score of the dictionary method with the supervised labels based on star ratings, we can test whether the output of the supervised learning model is consistent with the actual sentiment expression in the text. Therefore, a mutually corroborating relationship is established with user ratings through model prediction and analysis of lexical polarity.

In conclusion, this study breaks through the traditional single dimension of focusing only on "model accuracy comparison", and places the research perspective in the complex situation composed of real e-commerce environments, real player reviews, and real data distribution. It focuses on exploring how different sentiment analysis methods can complement and collaborate to jointly deepen the understanding of user sentiment patterns. The research results are expected to provide game developers and platform operators with more refined player sentiment profiles, thereby assisting in optimizing product recommendation mechanisms, enhancing public opinion monitoring capabilities, and supporting the subsequent implementation of text mining tasks. Meanwhile, this study also provides a methodological reference of practical value for how to scientifically select and combine sentiment analysis methods in actual application scenarios such as e-commerce reviews.

Methodology

This study applied multiple text mining and machine learning techniques to classify the sentiment of Amazon Video Game reviews and to interpret the linguistic structure behind users' opinions. The workflow consisted of text preprocessing, vectorization, supervised classification, and lexicon-based validation.

Text Preprocessing and Vectorization

Because the original Amazon Video Games review dataset was extremely large, a random subset of 5,000 reviews was drawn for analysis. A fixed random seed (set.seed(123) was applied to ensure that the sampling process and all subsequent results were fully reproducible. From the original dataset, only two variables were retained: overall (star rating) and reviewText, as these were sufficient for the supervised sentiment classification task.

The sentiment labels were assigned based on the star ratings in the dataset. Reviews with a rating above 3 were placed in the positive category, while those below 3 were treated as negative. The 3-star group was removed because these reviews tend to fall in the middle and do not clearly indicate sentiment. Afterwards, missing entries were cleared out, and reviews that only contained empty strings were converted to NA using na_if() before being dropped.

For data cleaning and preprocessing, all reviews were converted to lowercase, and punctuation, numbers, and common English stopwords were removed. Extra spaces were also cleaned up so the text would be easier to process. After these steps, 4,581 reviews were left in the dataset. The data was then split into 70% for training and 30% for testing, which is a typical setup for model evaluation.

To turn the cleaned text into numerical features, the reviews were represented using a TF–IDF matrix created with the tm package. TF–IDF weights words based on how often they appear in a single review compared to the whole collection. This representation allows the text to be used by machine-learning methods such as SVM, Naive Bayes, and Random Forest.

Classification Models

Support Vector Machine (SVM)

The first classification model applied in this study was the Support Vector Machine (SVM) with a linear kernel. SVM constructs an optimal hyperplane to separate data points of different sentiment classes in a high-dimensional feature space. By maximizing the margin between the positive and negative classes, the linear SVM is highly effective for text classification tasks involving sparse TF–IDF inputs.

The model's results are 82.37% accuracy, 66.1% balanced accuracy, and a 0.36 Kappa coefficient. This means the model's predictions match the most true labels. Sensitivity was 0.40, reflecting the model's ability to detect negative reviews, while specificity reached 0.92, demonstrating strong performance in identifying positive reviews. However, the moderate sensitivity indicates that some negative reviews were misclassified as positive, likely due to class imbalance in the dataset.

Naive Bayes

The second model is Naive Bayes. In this study, however, the Naive Bayes classifier produced poor results, with an accuracy of only 19.8% and a balanced accuracy of 50%. It is far below the baseline accuracy (81%). The confusion matrix showed that the model predicted almost all reviews as negative, resulting in extremely high sensitivity (0.99) but extremely low specificity (0.01). This behavior indicates a failure to distinguish between positive and negative reviews. It might caused by the combination of TF–IDF weighting and strong class imbalance. As a result, Naive Bayes was unsuitable for this dataset.

Random Forest (RF)

The third is the Random Forest (RF) model. This implemented using the randomForest package. Two versions were trained: one with a small ensemble (ntree = 3) and another with a larger ensemble (ntree = 300) to assess the impact of increasing model complexity.

With ntree = 3, the model's accuracy is 81.9%, balanced accuracy is 64.1%, and a Kappa is 0.32. Increasing the number of trees to ntree = 300 improved overall accuracy to 85.1%, but the balanced accuracy decreased to 62.1%. The larger ntree has very high specificity (0.99), meaning it correctly identified positive reviews. However, the lower sensitivity (0.25) indicates the model cannot detecting negative reviews effectively. This again reflects the dataset's class imbalance. Despite this, the Random Forest (ntree = 300) achieved the highest raw accuracy among all models. Thus, the model has the strong capacity for capturing complex patterns in textual data.

Lexicon-Based Sentiment Analysis (Hu & Liu)

To complement the supervised machine learning models, a lexicon-based sentiment analysis using the Hu & Liu sentiment lexicon (Bing Liu, 2004) was conducted. This approach provides an interpretable, unsupervised evaluation by identifying the polarity of words appearing in each review. Each review was tokenized into individual words, which were then matched against predefined lists of positive and negative terms.

From all datasets, there are 67.97% positive reviews and 32.03% negative reviews, indicating that Amazon Video Games reviews generally express favorable sentiment. A sentiment score was computed for each review by subtracting the number of negative words from the number of positive ones. Positive reviews achieved a mean score of 0.44, whereas negative reviews had a mean score of 0.01.

These results demonstrate strong alignment between the textual polarity and the numerical star ratings, supporting the validity of the supervised labels. High-rated reviews contained substantially more positive expressions, while low-rated reviews included more neutral or mildly negative terms. This consistency reinforces the reliability of both the dataset and the preprocessing pipeline.

Results

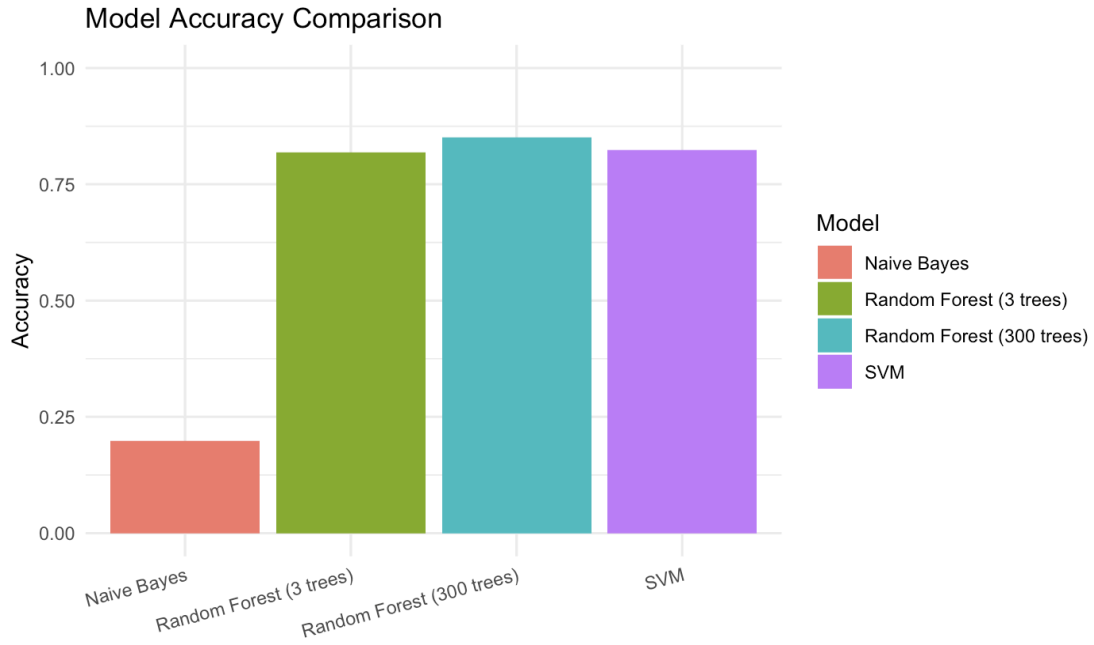


Figure shows the comparison of model accuracies across four classifiers: Naive Bayes, Random Forest (3 trees), Random Forest (300 trees), and Support Vector Machine (SVM).
All models except Naive Bayes achieved accuracy above 80%, which means both SVM and Random Forest are effective for sentiment classification on TF-IDF features.

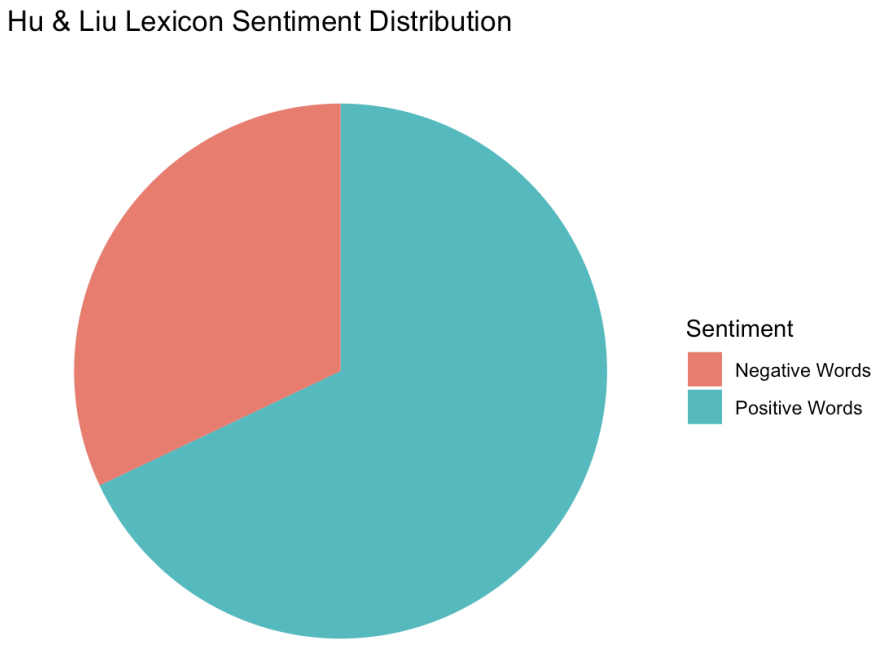
SVM achieved an accuracy of 82.37% with a balanced accuracy of 66.1%. The Random Forest model with 300 trees produced the highest overall accuracy (85.07%), but its sensitivity dropped to 0.25, indicating strong bias toward the majority positive class.

When the number of trees was reduced to 3, accuracy decreased slightly to 81.9%, but the model achieved better class balance (balanced accuracy = 64.1%). In contrast, Naive Bayes performed poorly (accuracy = 19.8%), classifying nearly all reviews as negative due to the incompatibility between TF-IDF features and its discrete probability assumptions.

The comparison highlights how different algorithms respond to high-dimensional text data.

The linear SVM demonstrates strong generalization and robustness with sparse TF-IDF vectors, confirming its suitability for sentiment classification. The Random Forest ensemble captures complex feature interactions but tends to overfit the dominant class, resulting in higher precision for positive reviews but weaker recall for negatives. This pattern reflects the inherent class imbalance of the dataset, where positive reviews represent more than 80% of the total.

Despite this imbalance, the overall accuracy and Cohen's Kappa values (0.32–0.36) suggest that both SVM and Random Forest learned meaningful sentiment distinctions rather than random noise. Therefore, these two models can be considered the most reliable for predicting review polarity in this dataset.



To further validate the classification results, a lexicon-based sentiment analysis was conducted using the Hu & Liu sentiment lexicon (Bing Liu, 2004). The figure illustrates the overall distribution of positive and negative words across all reviews. Approximately 68% of the words were positive, while 32% were negative, indicating a general tendency toward favorable opinions in Amazon Video Game reviews.

When aggregated by review label, positive reviews had an average sentiment score of 0.44, compared to 0.01 for negative reviews. This finding aligns with the supervised model predictions, confirming that the linguistic polarity of the text is consistent with the assigned sentiment labels based on user ratings.

In other words, both the machine learning models and the lexicon-based analysis identified similar emotional patterns: users tend to express more positive words in higher-rated reviews, while negative reviews contain fewer emotional cues.

Discussion

The results of this study highlight several important insights into the effectiveness of different sentiment classification methods applied to Amazon video game reviews. By comparing three supervised

machine-learning models—Support Vector Machine (SVM), Naive Bayes, and Random Forest—as well as a lexicon-based approach, this research provides a comprehensive understanding of how various techniques interpret user sentiment in a high-dimensional and imbalanced text dataset.

The classification outcomes demonstrate that both SVM and Random Forest perform strongly on TF–IDF transformed text, achieving accuracies above 80%. This reinforces the suitability of these models for sentiment analysis tasks involving sparse feature representations. The SVM model, in particular, showed stable and balanced performance. Its accuracy of 82.37% and balanced accuracy of 66.1% indicate that it can effectively distinguish between positive and negative reviews, even when the dataset is skewed toward positive sentiment. Its relatively high specificity (0.92) suggests that the model is highly reliable in detecting positive sentiment, while its more modest sensitivity (0.40) shows that it occasionally misclassifies negative reviews. This behaviour is expected given the dominance of positive reviews in the dataset.

Random Forest produced the highest overall accuracy when trained with 300 trees, reaching 85.07%. This suggests that the model is capable of capturing complex patterns and interactions between features that simpler models may not detect. However, this improvement in accuracy comes with a cost: the model exhibited very low sensitivity (0.25), demonstrating a strong bias toward predicting positive sentiment. This bias became even more pronounced as the number of trees increased, showing that although a larger ensemble allows the model to learn richer patterns, it also amplifies the imbalance already present in the training data. When using only three trees, Random Forest achieved slightly lower accuracy (81.9%) but demonstrated better balance between sensitivity and specificity. This contrast illustrates how model complexity can influence not only accuracy but also class balance.

In comparison, Naive Bayes performed poorly, with an accuracy of only 19.8%. The model heavily favoured the negative class, predicting nearly all reviews as negative despite positive reviews representing the majority of the dataset. This outcome suggests that the combination of TF–IDF weighting and the strong assumption of conditional independence between words is incompatible with the structure of video game reviews. Furthermore, TF–IDF features tend to diminish the relative importance of highly frequent sentiment words, which are often essential for the Naive Bayes model to function effectively. The outcome indicates that Naive Bayes is not suitable for this dataset without significant adjustments such as rebalancing the classes or experimenting with alternative feature-engineering techniques.

Beyond the supervised models, the lexicon-based sentiment analysis provided an additional layer of validation. Using the Hu & Liu sentiment dictionary, the results showed that approximately 68% of the words identified were positive, while 32% were negative. This distribution is consistent with the predominance of high star ratings in online game reviews. When comparing lexicon-derived sentiment scores with the rating-based labels, the patterns aligned closely: positive reviews had an average sentiment score of 0.44, while negative reviews had a near-neutral score of 0.01. This suggests that negative reviews tend to be less emotionally expressive, containing fewer explicitly negative words. Instead of harsh criticism, many negative reviews take on a more descriptive or neutral tone, which may explain why models struggled to classify them accurately.

The close alignment between lexicon-based polarity and model predictions strengthens the validity of both the preprocessing steps and the supervised learning outcomes. The lexicon results also provide useful interpretability: while machine-learning models capture statistical relationships, lexicon analysis directly reveals the distribution of emotional terms. Together, these methods give a fuller understanding of user sentiment and show that Amazon video game reviews are generally characterised by positive user experiences, with dissatisfaction often expressed in more subtle ways.

Overall, the findings demonstrate that sentiment analysis of video game reviews is feasible and reliable using appropriate machine-learning techniques. However, the comparison also highlights the importance of considering class imbalance, modelling assumptions, and feature representations when selecting a method. The combination of supervised models and lexicon-based analysis provides a more robust and comprehensive interpretation than relying on a single technique alone.

Conclusion

This study set out to evaluate the effectiveness of different sentiment analysis techniques for classifying Amazon video game reviews, using a combination of supervised machine-learning models and a lexicon-based approach. Through a structured workflow of sampling, cleaning, TF–IDF vectorisation, supervised classification, and sentiment scoring, the research achieved a comprehensive understanding of how user opinions in the video-game market can be automatically analysed at scale.

The results show that sentiment classification using TF–IDF features is both feasible and reliable when suitable machine-learning models are applied. Among the three supervised methods tested, Random Forest with 300 trees achieved the highest overall accuracy, demonstrating its ability to capture complex patterns within the text. However, its tendency to over-predict the majority class highlights the practical limitations of accuracy-sensitive metrics, especially when dealing with imbalanced datasets. In contrast, the SVM classifier produced more balanced results and was more consistent at identifying both positive and negative reviews, making it a more stable model for real-world applications where a mix of sentiment must be correctly captured.

The poor performance of Naive Bayes serves as a reminder that not all classical text-classification models generalise well to large, sparse, and skewed datasets. Its misclassification patterns emphasise the importance of choosing modelling techniques that align with the structural characteristics of the data.

The lexicon-based sentiment analysis added valuable interpretability to the findings. Its consistency with rating-derived labels confirmed the quality of the dataset and validated the supervised models' predictions. The polarity patterns uncovered in this analysis suggest that positive reviews tend to contain more direct emotional expressions, while negative reviews often rely on neutral language, making them harder to detect through automated methods.

Overall, this study contributes meaningful insights into the strengths and limitations of different sentiment-analysis approaches. The findings demonstrate that automated sentiment classification can serve as a useful decision-support tool for understanding user experiences, guiding product improvement, and enhancing recommendation systems. Future work could incorporate balancing strategies, larger datasets, or deep-learning techniques to further refine model performance, especially for minority sentiment categories.