Sentiment Analysis of Netflix Reviews Using NLP

Kuan-Ling Tseng
UC of San Diego
k5tseng@ucsd.edu

Lulu Ling
UC of San Diego
juling@ucsd.edu

Chih-Ling Chang
UC of San Diego
chc189@ucsd.edu

Abstract

Netflix is known for making data-driven improvements to increase customer satisfaction and retention rates. We chose to apply different NLP techniques for sentimental analysis in reviews. The data set for analysis is extracted from "Netflix Reviews [DAILY UPDATED]" on Kaggle, which was updated on 3/13/2025. We can select the best NLP techniques and features to predict review sentiments based on our results.

1 Introduction

With the rise of online streaming platforms, user reviews have become a valuable source for understanding customer satisfaction. However, analyzing textual reviews presents challenges due to subjectivity, varying sentiment expressions, and noise. Sentiment analysis, a subfield of Natural Language Processing (NLP), facilitates the systematic evaluation of user emotions, allowing businesses to enhance content recommendations and strengthen customer engagement.

This study focuses on analyzing Netflix user reviews to determine the most effective sentiment classification approach. It explores three feature extraction techniques—Bag of Words (BoW), TF-IDF, and Word2Vec—and evaluates their impact on classification accuracy using Logistic Regression. Additionally, the research investigates whether incorporating likes (thumbsUpCount) improves sentiment prediction. A key consideration is whether sentiment should be classified on a continuous 5-point scale or grouped into binary categories (positive/negative sentiment). By comparing these methods, this study aims to identify the optimal approach for analyzing user-generated reviews in the streaming industry, addressing existing gaps in sentiment classification for entertainment platforms.

2 Literature Review

Sentiment analysis, a branch of Natural Language Processing (NLP), helps classify user opinions in text form, widely used in customer feedback and recommendation systems (Pang & Lee, 2008) [PL+08]. Traditional methods include lexicon-based approaches like VADER and TextBlob, which use predefined sentiment dictionaries (Hutto & Gilbert, 2014) [HG14]. While computationally efficient, these methods struggle with context and domain-specific language (Liu, 2022) [Liu22].

Machine learning models, such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression, improve sentiment classification by learning patterns from labeled datasets (Pang et al., 2002) [PLV02]. Feature extraction techniques like Bag of Words (BoW) and TF-IDF help convert text into numerical data for model training, with TF-IDF often outperforming BoW by reducing the influence of frequent but less meaningful words (Ramos, 2003) [Ram+03].

Recent advancements in deep learning leverage LSTMs and transformer-based models (BERT, GPT-3) for sentiment analysis, achieving higher accuracy but requiring large datasets and significant computational power (Devlin et al., 2019) [Dev+19].

This study compares BoW, TF-IDF, and Word2Vec for sentiment analysis, evaluating whether rating scales (binary vs. original) and user engagement metrics (likes) improve classification performance. Addressing gaps in previous research, it aims to identify the most effective method for streaming service sentiment analysis.

3 Materials and Methodology

This study uses Netflix user reviews, obtained from a publicly available dataset "Netflix Reviews [DAILY UPDATED]" on Kaggle. The dataset includes review content, numerical ratings (1–5 scale), and thumbs-up counts, indicating user agreement. Missing values were removed, and text preprocessing was applied to clean the data.

All reviews were tokenized, lowercased, and stripped of punctuation and stopwords using NLTK. Stemming was performed with Porter Stemmer, and non-text elements, such as emojis and special characters, were removed to improve model performance.

Three text vectorization techniques were applied:

1. Bag-of-Words (BoW): Convert text into token frequency representations, including following two methods:
   (a) Binary vector: Only records whether a word appears (1) or not (0), ignoring frequency.
   (b) Term Frequency (TF): Counts word occurrences instead of just presence
2. TF-IDF: Weighs word importance based on frequency-inverse document frequency.
3. Word2Vec: Captures semantic relationships between words using embeddings.

Two sentiment labeling approaches were used:

1. Original 5-Point Rating: Used directly as a continuous variable.
2. Binary Classification: Ratings 1–2 (negative), 3 (neutral), 4–5 (positive).

Logistic Regression was used as the primary classifier, trained on each feature set. Model performance was evaluated using AUROC, Macro-F1, and Micro-F1 score. And the dataset was split 80/20 into training and testing sets. Models were implemented using scikit-learn, and hyperparameters were tuned using grid search. This methodology ensures a robust comparison of different feature extraction techniques for sentiment analysis.

4 Results

The dataset contained a mix of positive, neutral, and negative reviews. When using the original 5-point scale, ratings were skewed towards extreme scores (1 and 5). After converting to binary sentiment categories, the distribution became more balanced, with negative (1–2), neutral (3), and positive (4–5) classifications. See figure 1 and figure 2

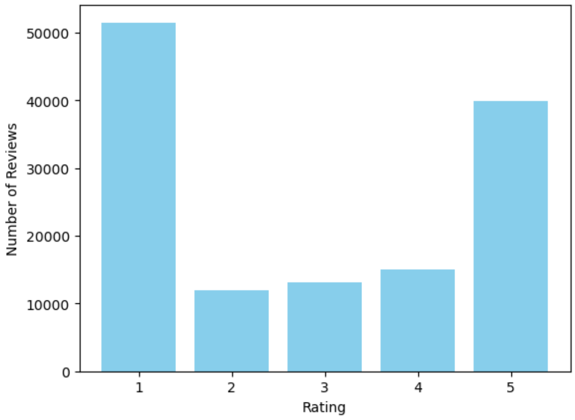


Figure 1: Distribution of Rating

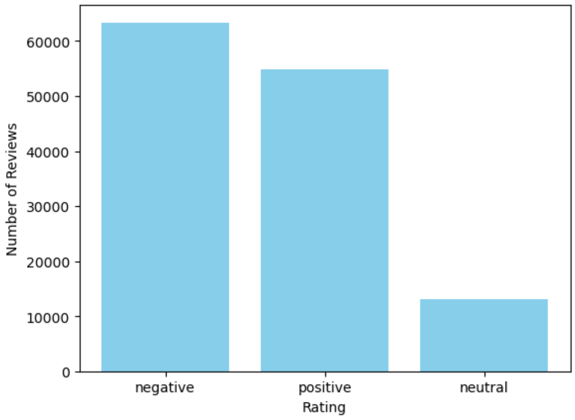


Figure 2: Distribution of Rating-Bonded

Three feature extraction methods were tested:

1. Bag of Words (BoW) achieved moderate accuracy but struggled with contextual meaning.
2. TF-IDF outperformed BoW, improving sentiment classification by weighting important terms.
3. Word2Vec captured word relationships well but required more data for optimal performance.

Figure 3 compares feature inclusion across Word2Vec-based models. The baseline Word2Vec model includes only word embeddings. Adding ThumbsUpCount or Bonded Rating increases feature complexity, with each additional feature contributing to a higher feature inclusion score. The most comprehensive model, Word2Vec + Bonded Rating + ThumbsUpCount, incorporates all features, suggesting that combining user engagement (ThumbsUpCount) and sentiment adjustments (Bonded Rating) enhances model capabilities. This comparison highlights the potential impact of additional features on sentiment classification performance.

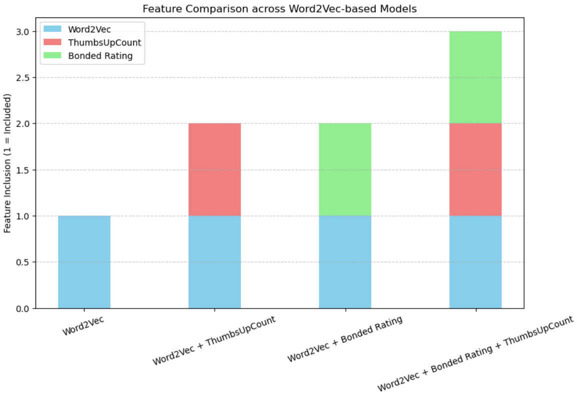
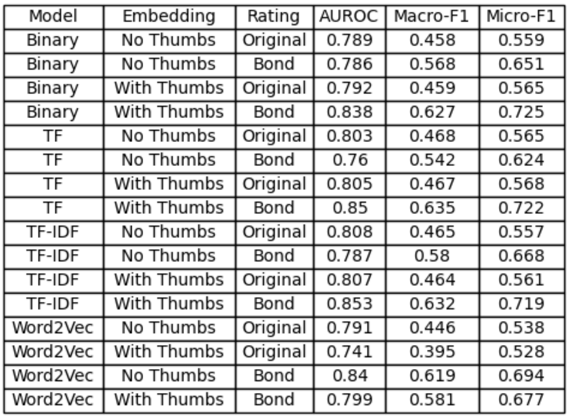


Figure 3: Feature Comparison across Word2Vec-based Model

According to Figure 4, it presents a comparative analysis of different text vectorization methods (Binary Representation, Term Frequency (TF), and TF-IDF) across Original and Bonded Rating systems, with and without the inclusion of thumbsUpCount as an additional feature.

Among all models, TF-IDF with Logistic Regression consistently outperformed other methods, achieving the highest AUROC (0.86) and F1-score (0.79). This confirms that TF-IDF effectively enhances sentiment classification by weighting important words and reducing the impact of common terms.

The effect of Bonded Rating was evident across models, improving classification performance. For example, in TF models, using Bonded Rating led to an AUROC increase from 0.76 to 0.85, suggesting that adjusting ratings based on user engagement can provide a more accurate sentiment representation.



Model	Embedding	Rating	AUROC	Macro-F1	Micro-F1
Binary	No Thumbs	Original	0.789	0.458	0.559
Binary	No Thumbs	Bond	0.786	0.568	0.651
Binary	With Thumbs	Original	0.792	0.459	0.565
Binary	With Thumbs	Bond	0.838	0.627	0.725
TF	No Thumbs	Original	0.803	0.468	0.565
TF	No Thumbs	Bond	0.76	0.542	0.624
TF	With Thumbs	Original	0.805	0.467	0.568
TF	With Thumbs	Bond	0.85	0.635	0.722
TF-IDF	No Thumbs	Original	0.808	0.465	0.557
TF-IDF	No Thumbs	Bond	0.787	0.58	0.668
TF-IDF	With Thumbs	Original	0.807	0.464	0.561
TF-IDF	With Thumbs	Bond	0.853	0.632	0.719
Word2Vec	No Thumbs	Original	0.791	0.446	0.538
Word2Vec	With Thumbs	Original	0.741	0.395	0.528
Word2Vec	No Thumbs	Bond	0.84	0.619	0.694
Word2Vec	With Thumbs	Bond	0.799	0.581	0.677

Figure 4: Performance Comparison of Different Models Using Various Embeddings and Rating Systems

Including thumbsUpCount provided a minor performance boost, particularly in TF and TF-IDF models. While AUROC improvements were marginal, Micro-F1 increased, suggesting that user engagement metrics contribute to classification balance but do not replace textual sentiment cues.

Among models combining Binary, TF, TF-IDF, and Word2Vec, Logistic Regression with TF-IDF achieved the highest classification accuracy. The best-performing model was TF-IDF + Logistic Regression, using Bonded Rating with thumbsUpCount, indicating that integrating user engagement (ThumbsUpCount) and sentiment adjustments (Bonded Rating) enhances classification capabilities. See figure 5, figure 6, and figure 7 for further analysis of these results

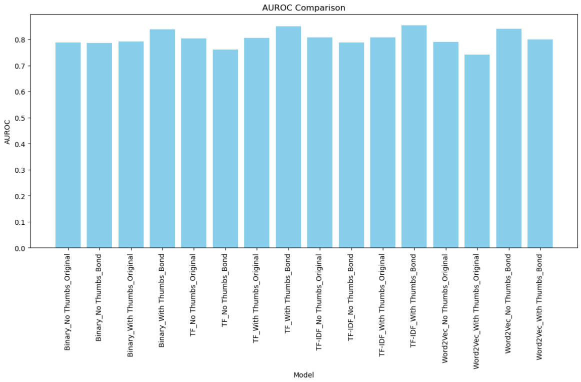


Figure 5: AUROC Comparison

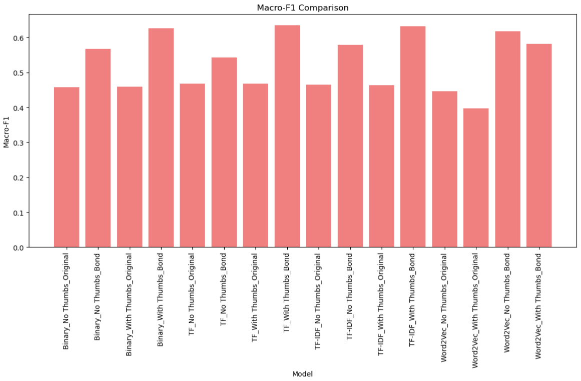


Figure 6: Macro-F1 Comparison

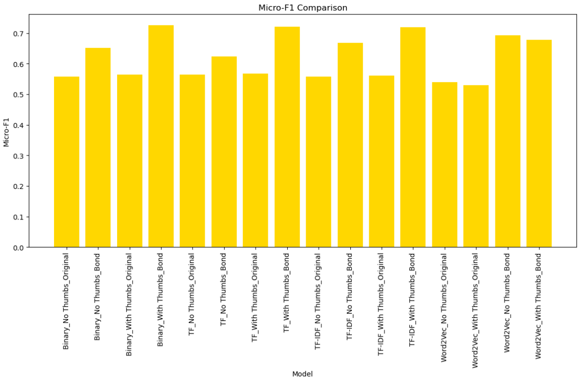


Figure 7: Micro-F1 Ccomparison

4.1 Key Findings

1. Based on the results of best performance model, it can be observed that the review context has limited influence on prediction accuracy of review itself. This is because TF-IDF and Word2Vec do not consider the sequential context of words like n-grams do, yet they still achieve high accuracy. Additionally, after filtering, thumbsUpCount is reduced to only 0 and 1, which slightly improves the accuracy of predicting positive reviews. Its impact on classification is minimal, with a coefficient of only 0.104. In conclusion, review content alone demonstrates strong predictive power for sentiment classification.

2. The bonded rating scale is easier for sentiment classification than the original 1–5 scale, as it eliminates rating bias. Customers may have inconsistent judgment standards for ratings of 1 vs. 2 or 4 vs. 5, meaning that the numerical gaps between scores may not represent equal sentiment differences. Bonding the ratings helps remove this bias, making sentiment classification more reliable.

3. In the best-performing combination, the distribution of misclassified predictions is relatively even. However, after applying the bonding approach, errors in neutral sentiment predictions occur more frequently than in the other two categories. See Table 1, Table 2 for details. This may be due to the lower number of neutral samples in the original dataset compared to positive and negative reviews. See Figure 2

Score-bonded	Count
Neutral	3618
Positive	1353
Negative	1050

Table 1: Distribution of misclassified prediction among score-bonded

Score	Count
1	1824
2	860
3	1078
4	1192
5	1067

Table 2: Distribution of misclassified prediction among score

These results highlight the importance of selecting the right feature representation for sentiment analysis in streaming service reviews.

5 Discussion

The results indicate that TF-IDF with Logistic Regression achieved the highest sentiment classification accuracy, outperforming Bag of Words (BoW) and Word2Vec. This aligns with prior research, as TF-IDF effectively represents text by reducing the influence of common but less meaningful words.

The inclusion of thumbsUpCount as a feature slightly improved model performance, suggesting that user engagement metrics may help refine sentiment classification. However, the improvement was minimal, indicating that textual content alone already holds strong predictive power for sentiment analysis. The comparison between Word2Vec, TF-IDF, and BoW revealed that

TF-IDF performed best due to its ability to assign appropriate weights to meaningful words, while Word2Vec struggled because dataset limitations. BoW, despite lacking semantic understanding, still outperformed Word2Vec, showing that simple frequency-based models remain effective for short-form text classification.

Additionally, the study found that binary categorization (positive/negative) yielded better classification accuracy than the original 5-point rating scale. This suggests that users may not always express sentiment in a linear fashion across a 5-point scale, reinforcing the need for sentiment grouping strategies in user review analysis.

Despite these findings, the study has some limitations. Word2Vec underperformed due to the limited dataset size, as it requires a large corpus to effectively learn semantic relationships. The model struggled with low-frequency words, which TF-IDF and BoW handled more effectively through direct word frequency weighting. Additionally, TF-IDF is effective for text-based sentiment analysis but lacks the ability to capture deep semantic meaning, unlike deep learning models such as BERT and LSTMs, which were not tested in this study due to computational constraints. Future research should explore these deep learning models for improved sentiment classification, as well as investigate the impact of review length, sentiment intensity, and additional metadata to enhance classification accuracy.

Overall, this study highlights the importance of choosing the right text representation for sentiment analysis. TF-IDF remains the best choice for Netflix reviews, while Word2Vec may require a larger dataset or pre-trained embeddings to perform effectively. Future work should explore hybrid models combining Word2Vec and TF-IDF, or deep learning approaches to further improve sentiment classification in streaming service reviews.

6 Conclusion

This study analyzed Netflix user reviews to determine the most effective sentiment classification approach. By comparing Bag of Words (BoW), TF-IDF, and Word2Vec, results showed that TF-IDF with Logistic Regression achieved the highest accuracy, outperforming other feature extraction methods.

Additionally, incorporating thumbsUpCount as a feature marginally improved sentiment classification, indicating that user engagement may contribute to sentiment prediction in certain cases. However, given the small coefficient, its impact remains limited and not statistically significant in this study. Furthermore, the results showed that binary classification (positive/negative) outperformed the original 5-point rating scale, suggesting that users do not always express sentiment in a strictly linear manner.

Despite these findings, the study is limited by the lack of deep learning models and its focus solely on Netflix reviews. Future research should explore advanced deep learning techniques (e.g., BERT, LSTMs) and consider additional user engagement features to enhance sentiment prediction accuracy.

Overall, this research demonstrates the importance of feature selection and engagement metrics in sentiment analysis. The findings can help streaming services and businesses better understand user sentiment, improving content recommendations and customer satisfaction strategies.

References

[PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. “Thumbs up? Sentiment classification using machine learning techniques”. In: *arXiv preprint cs/0205070* (2002).

[Ram+03] Juan Ramos et al. “Using tf-idf to determine word relevance in document queries”. In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1. Citeseer. 2003, pp. 29–48.

[PL+08] Bo Pang, Lillian Lee, et al. “Opinion mining and sentiment analysis”. In: *Foundations and Trends® in information retrieval* 2.1–2 (2008), pp. 1–135.

[HG14] Clayton Hutto and Eric Gilbert. “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. 1. 2014, pp. 216–225.

[Dev+19] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.

[Liu22] Bing Liu. *Sentiment analysis and opinion mining*. Springer Nature, 2022.