Sentiment Analysis over Goodreads Reviews

Matthew D. Branson, James R. Brown

Department of Computer Science

Missouri State University

Springfield, USA

branson773@live.missouristate.edu

brown926@live.missouristate.edu

Abstract

This project explores sentiment analysis using the UCSD Goodreads Book Graph datasets. We examine book reviews and the users who wrote them, focusing on predicting star ratings from review text and analyzing cross-domain adaptation of sentiment classifiers. Additional dimensions of analysis include review categorization, genre-specific sentiment patterns, and the development of a parametric search interface for the extensive dataset.

Index Terms

sentiment analysis, natural language processing, Goodreads, domain adaptation, classification, information retrieval

I. Project Description

Using the UCSD Goodreads Book Graph datasets, we propose to examine book reviews and the users who wrote them. The dataset contains information about 2,360,655 books (1,521,962 works, 400,390 book series, 829,529 authors); 876,145 users; 228,648,342 user-book interactions in users' shelves (including 112,131,203 reads and 104,551,549 ratings). A GitHub repo containing the source code of our project can be viewed at https://github.com/mdb42/goodreads.

II. Literature Review

Our project primarily builds upon the work of the curators of the UCSD Goodreads dataset. In their first paper, Wan and McAuley [1] introduce chainRec, a recommendation system that models user interactions as sequences of increasingly committed behaviors. This paper is significant to our work not only for the comprehensive Goodreads dataset it provides—comprising over 225 million user-item interactions—but also for its conceptual framework of viewing user engagement as structured sequences rather than independent events. The authors specifically request citation of this work when using their dataset.

In a follow-up study, Wan et al. [2] developed SpoilerNet, a neural model trained on 1.4 million Goodreads reviews to detect spoilers at the sentence level. This research demonstrates that users exhibit stable, quantifiable patterns in their reviewing behavior, offering a methodological precedent for our user-level sentiment analysis. The authors specifically request citation of this work when using their dataset.

For sentiment analysis methodology, Monika and Chooralil [3] offer a comprehensive survey of techniques, providing a systematic framework for the analytical pipeline from data collection through interpretation.

Reagle [4] examines online comment culture across platforms, developing a taxonomy of commenting behaviors with six functions: informing, improving, manipulating, alienating, shaping identity, and perplexing. While not focused specifically on literary reviewing, Reagle's typology offers a valuable framework for categorizing reviewers and understanding how digital environments may amplify certain reviewing behaviors.

Sentiment analysis has been used in machine learning and deep learning contexts for a variety of applications. Chandra and Jana have found that sentiment analysis can be an effective tool when assessing the general public's feelings about products and topics on social media [5]. Going beyond machine learning, it seems that using deep learning could lead to more accurate analysis, which presents an interesting area to extend this project into in the future. Additionally, given the social media-like nature of Goodreads, it will be interesting to see if sentiment analysis is just as effective for Goodreads as it has for Twitter, despite differences in approach.

Tangential to Goodreads, sentiment analysis seems to have some interesting use cases when determining structure of a novel, particularly novels that may not use a traditional plot structure. Elkins and Chun found that sentiment analysis when paired with manual close reading can result in new critiques of literature [6]. Despite some issues in the `Syuzhet.R` library used for sentiment analysis, the researchers still found interesting new insights and emotional arcs in the novel that may be missed by other readers. However, it was noted that the library used struggled with properly assigning sentimental scores to certain kinds of grammar, like negation, capitalization, and even emojis. Since we are expecting a much more informal style of writing in Goodreads reviews, we will have to monitor how our approach handles situations where excessive punctuation, emojis, and incorrect spelling/grammar are present.

Beyond emotional analysis, sentiment analysis could be used to gain insights on those participating in the conversation(s). Sokolova and Bobicev attempted to use sentiment analysis on medical forums to evaluate the presence of the "echo chamber effect" in those forums [7]. However, the authors had some difficulty finding properly-labeled data to train their model effectively to make these sorts of analyses. While Goodreads reviews often come with a star rating that can be used as a label, great care needs to be taken in what kinds of conclusions can be drawn from our approach beyond predicting a star rating.

Given that data availability can be an issue in just about any application where a sentiment analyzer may be used, some researchers have been looking in to making more general-purpose models to improve performance when used in subjects unrelated to the model's test set. SentiX, a cross-domain sentiment analysis model, was proposed by Zhou et. al. to be used on several domains of user reviews without the need for fine-tuning the model along the way [8]. This model beats most BERT-based models and several other models (excluding the domain that other models were trained on) while being trained on less samples than other models. This experiment seems *far* more complex than ours will be, demonstrating the difficulty in constructing such a model while avoiding overfitting on the domain that the model is trained on.

Naïve Bayes still seems to be a valid framework to build a sentiment analysis model around for internet user reviews. Quadri and Selvakumar used Naïve Bayes to develop a model that analyzes cross-domain reviews, achieving an accuracy range of 76% to 99% against a set of different domains and review websites [9]. It is noted that the results could be better if the model used other variants of Naïve Bayes for certain domains over others (Multinomial and Bernoulli specifically), but the results achieved here seem to make a promising case for the effectiveness of our approach in both the trained Goodreads domain and a different "target" domain.

Feature selection is another important component in developing a sentiment analysis model. These are used to make sure the model is getting relevant information from any given review to give it the best chance to make the correct prediction on the sentiment of the review. Several sentiment analysis models used in other published journals use feature selection methods that information retrieval systems tend to lean on, like document frequency, chi-squared ($\mathcal{X}^2$), odds ratio, and clustering [10]. While a classifier like Naïve Bayes is the decision maker of the model, feature selection reduces the amount of input to a more reasonable amount, ideally improving performance of the model and resulting in a more accurate model overall.

Shahsavari et al. developed a pipeline for extracting character and relationship networks from Goodreads book reviews using Greimasian actant theory and latent graphical models [11]. By aggregating thousands of reviews per novel, their system was able to reconstruct a "consensus narrative framework" with high accuracy. While their focus is narrative rather than sentiment, their work demonstrates the viability of mining large-scale Goodreads review data to infer structured latent knowledge from unstructured user text—supporting our motivation to extract aggregated reviewer patterns from the same platform.

Hajibayova performed a linguistic and psychological analysis of over 470,000 Goodreads reviews to examine how users express personal reactions through review text [12]. The findings suggest that reviews are not only evaluative but also performative—shaped by an intent to influence cultural consumption, in line with Bourdieu's theory of symbolic capital. The prevalence of highly positive language was noted as a potential reliability concern. This work helps ground our interpretation of reviewer sentiment and supports the idea that review tone may reflect social motives as well as genuine opinion.

III. Background

To construct the sentiment analyzer, we will be using a Naïve Bayes classifier to make decisions on the sentiment "rating" of a review. This classifier will be fed information from a feature extraction pipeline using *tf-idf* over tokenized and stemmed reviews. This feature extraction method helps determine how "important" a word is to any given review. From these, we are looking to predict a star rating on a discrete 1–5 scale, matching Goodreads' review system. After this, we will cluster reviewers based on their aggregate sentiment scores, review frequency, and other behavioral metrics.

Once the model has been trained on the Goodreads dataset, we will proceed with applying the model to a set of movie reviews and examine how effective the model is in cross-domain applications.

IV. Methods

A. Data Collection and Indexing

This project will use the UCSD Book Graph dataset, which includes over 2.3 million books, 1.3 million user reviews, and approximately 18,000 unique users. The raw data is provided in gzipped JSON-line format and will be processed using a custom PyQt6-based data exploration tool. Review records will be parsed into a relational SQLite database, with schema support for books, authors, users, reviews, and user-defined genres.

To support scalable document processing and analysis, we will adapt an existing modular indexing framework built for information retrieval tasks. Index construction will be performed using either the `StandardIndex` or `ParallelIndex` class, selected dynamically based on dataset size and available system resources. These indexers tokenize, stem, and filter review text, storing normalized term frequencies per document. Multiprocessing support will enable distributed indexing when appropriate, with fallback to sequential processing if necessary.

B. Feature Extraction

We will extract both document-level and user-level features to support classification and clustering. Review text will be preprocessed using NLTK's `word_tokenize`, with custom logic to retain negation terms (e.g., *not*, *never*) and filter out non-informative stopwords. Tokens will be lowercased and stemmed using Porter's algorithm. *tf-idf* vectors will be constructed using a filtered vocabulary with document frequency thresholds between 5 and 85%. Both unigrams and bigrams will be included in the final term space.

For user-level feature extraction, we will aggregate document vectors to create per-user representations. These feature vectors will include:

- Mean rating across all reviews authored by the user
- Rating variance and review count
- Average sentiment score (derived from external model predictions)
- Mean *tf-idf* vector across the user's reviews

These features will be used as inputs to both supervised and unsupervised modeling. Scalar attributes will be standardized to unit variance, and all sparse vectors will be L2-normalized to ensure compatibility with cosine-based similarity metrics.

C. Modeling Framework

We will integrate two modeling approaches using the existing retrieval infrastructure: (1) supervised sentiment classification using Naïve Bayes, and (2) unsupervised clustering using K-means. The modeling layer is implemented on top of a shared vector space model (VSM) abstraction that supports multiple weighting schemes and backends. Three concrete VSM implementations are available—standard, parallel, and sparse—each compatible with the same indexing interface. These models are selected dynamically using a factory pattern, enabling flexible experimentation without duplicating preprocessing logic.

1) Sentiment Classification with Naïve Bayes: To predict user star ratings based on review text, we will train a Multinomial Naïve Bayes classifier. *tf-idf* vectors computed during indexing will serve as the input feature space. Class priors and likelihoods will be estimated from the training set, and Laplace smoothing ($\alpha = 0.3$) will be applied to account for unseen terms. The conditional probability of a rating class c given a document d will be computed as:

$$P(c \mid d) \propto P(c) \prod_{i=1}^{n} P(t_i \mid c)^{f_{i,d}},$$

where $f_{i,d}$ denotes the frequency of token t_i in document d. Evaluation will be conducted using five-fold cross-validation, and we will report accuracy and F_1-scores across all five rating categories. Comparative testing with support vector machines and logistic regression will be used to evaluate model performance on both extreme and ambiguous reviews.

2) Reviewer Clustering with K-means: To identify latent reviewer archetypes, we will apply K-means clustering to the aggregated user feature vectors. Each user will be represented by a hybrid vector comprising behavioral features (e.g., mean rating, review frequency) and semantic features derived from *tf-idf* vectors of their authored reviews. All features will be standardized to ensure compatibility with the Euclidean distance metric used by K-means.

Clustering will be performed over a range of k values from 2 to 10. Optimal cluster count will be selected using the elbow method and silhouette score analysis. Post-hoc interpretation will rely on cluster centroids and representative users, with an eye toward identifying groups such as critical reviewers, casual enthusiasts, or sentiment-divergent users. Future work may involve experimenting with alternative clustering algorithms or dimensionality reduction techniques to improve interpretability and stability of the clusters.

V. Results

TABLE I
Classification Performance Metrics

Metric	**Value**
Accuracy	0.6074
Precision	0.6359
Recall	0.4954
F_1 Score	0.5173

Table I displays the aggregated accuracy, precision, recall, and F_1 scores across all classes when classifying a set of 120,000 reviews. Both metrics for accuracy and precision returned at or above 60%, while recall and F_1 returned closer to 50%. While our model is able to determine the correct label for a review over 50% of the time, this demonstrates that only using a multinomial

TABLE II
SUMMARY OF USER CLUSTERS

Cluster	Size (Users)	Avg. Word Length	Avg. Rating	Num Reviews	Rating Variance
0	779	3442.32	3.7354	7.7895	0.6028
1	4712	1104.07	3.7789	8.8326	0.6937
2	2304	2055.78	3.7747	9.1563	0.6473
3	7483	416.56	3.8417	6.7879	0.5923
4	116	6322.63	3.4145	3.6207	0.3552

Naïve Bayes classifier may not be sufficient enough for certain contexts where effectiveness matters more.

Figure 1 displays a confusion matrix depicting the effectiveness of our model's multinomial Naïve Bayes classifier over a set of 120,000 user reviews from Goodreads. At a glance, this confusion matrix seems to suggest that a sentiment analysis using multinomial Naïve Bayes is effective when labeling strongly-rated reviews. However, this seems to lose a bit of fidelity when more lukewarm reviews are processed, particularly for labels of 2-, 3-, and 4-star ratings. This indicates that the classifier has a bit more difficulty in determining which terms belong to a review that's expressing a moderated sentiment.

Table II shows the results of the K-means cluster analysis of Goodreads reviewers, which gives us insight into what archetypes may exist among reviewers.

Reveiwers were ultimately separated into four clusters. Cluster 0 seems to encapsulate fairly verbose reviewers with lower ratings relative to the rest of the clusters. Cluster 4 seems to contain the most verbose reviewers within the dataset. This seems to suggest that more verbose reviewers tend to leave more negative comments relative to other reviewers. While the middle clusters contain the grand majority of reviewers in the dataset, it seems that these reviewers are much less likely to post long, negative reviews

Figure 2 exhibits a principle components analysis visualization of the clustering analysis, which better demonstrates the size of each cluster and their exhibited variance. Consistent with the metrics shown in Table II, we can see that both cluster 0 and 4 are the furthest right along the PC1 axis, indicating that these clusters contain much larger reviews. The remaining clusters also have a greater variance along the PC2 axis, indicating that these reviewers are, in fact, leaving more positive reviews than those leaving longer reviews in general.

VI. Conclusion

While using a Naïve Bayes classifier alone may not be the most reliable classifier when used in isolation, it still offers some meaningful insight into the characteristics of our Goodreads dataset. The clustering analysis used in this experiment gives us an interesting perspective about users on Goodreads, which could be used to improve recommendation algorithms over time. Overall, while both the classification and clustering approaches may require further refinement and external validation, this work establishes a strong foundation for continued exploration of sentiment analysis and user modeling in literary domains.

References

[1] M. Wan and J. McAuley, "Item recommendation on monotonic behavior chains," in *Proceedings of the 12th ACM Conference on Recommender Systems*, ser. RecSys '18. Association for Computing Machinery, 2018, pp. 86–94.

[2] M. Wan, R. Misra, N. Nakashole, and J. McAuley, "Fine-grained spoiler detection from large-scale review corpora," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2605–2610.

[3] L. Monika and V. Chooralil, "Sentiment analysis: A survey on design framework, applications and future scopes," *Artificial Intelligence Review*, vol. 56, pp. 12 505–12 560, 2023.

[4] J. M. Reagle, *Reading the Comments: Likers, Haters, and Manipulators at the Bottom of the Web*. MIT Press, 2015.

[5] Y. Chandra and A. Jana, "Sentiment analysis using machine learning and deep learning," in *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2020, pp. 1–4.

[6] K. Elkins and J. Chun, "Can sentiment analysis reveal structure in a plotless novel?" *ArXiv*, vol. abs/1910.01441, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:203641988

[7] M. S. V. Bobicev, "Machine learning evaluation of the echo-chamber effect in medical forums," *CoRR*, vol. abs/2010.09574, 2020. [Online]. Available: https://arxiv.org/abs/2010.09574

[8] J. Zhou, J. Tian, R. Wang, Y. Wu, W. Xiao, and L. He, "SentiX: A sentiment-aware pre-trained model for cross-domain sentiment analysis," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 568–579. [Online]. Available: https://aclanthology.org/2020.coling-main.49/

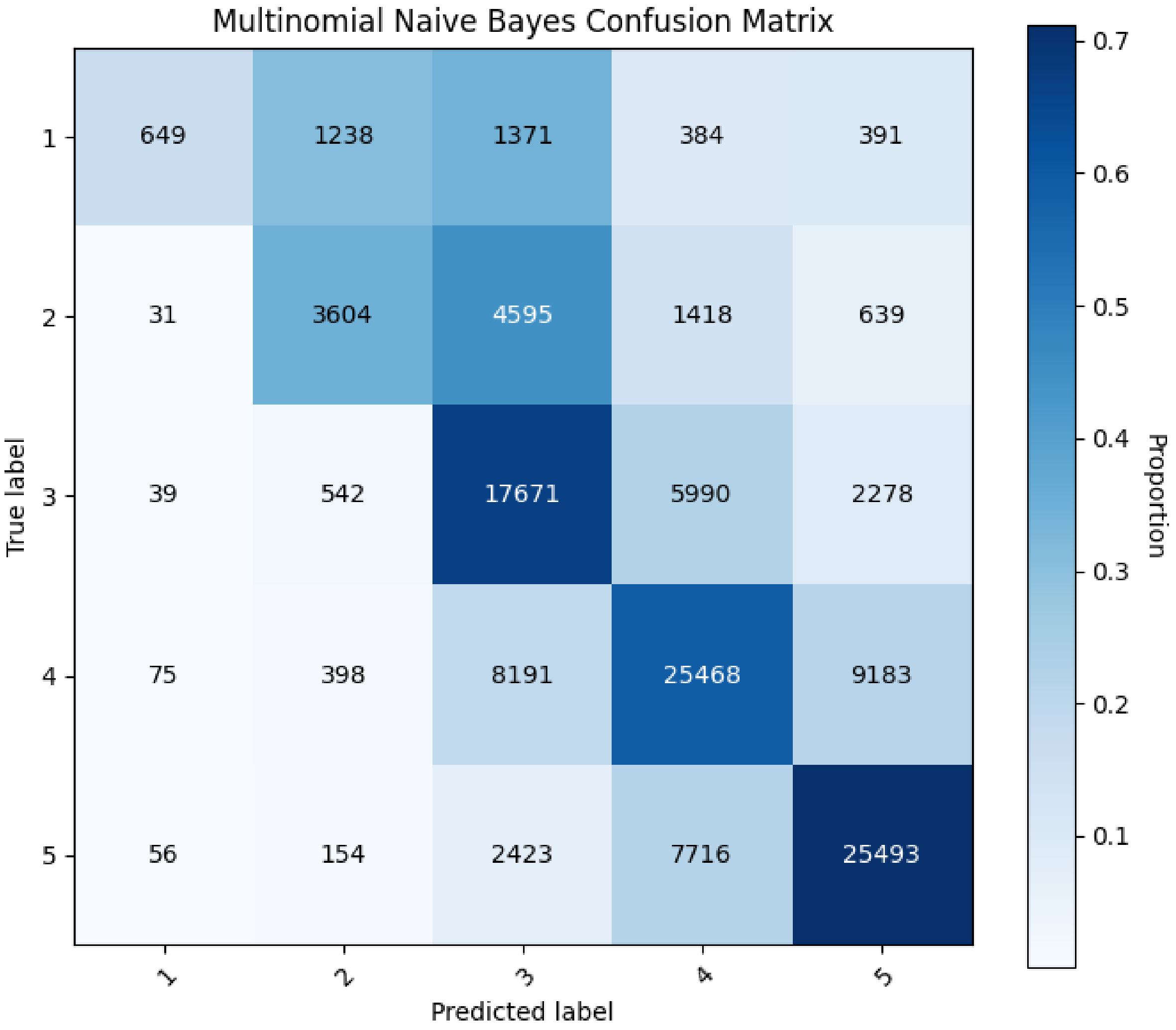
[9] M. Quadri and R. Selvakumar, "Performance of naïve bayes in sentiment analysis of user reviews online," *International Journal of Innovative Technology and Exploring Engineering*, vol. 10, pp. 64–68, 12 2020.

[10] L. Hung, R. Alfred, and M. Hijazi, "A review on feature selection methods for sentiment analysis," *Advanced Science Letters*, vol. 21, pp. 2952–2956, 10 2015.

[11] S. Shahsavari, E. Ebrahimzadeh, B. Shahbazi, M. Falahi, P. Holur, R. Bandari, T. R. Tangherlini, and V. Roychowdhury, "An automated pipeline for character and relationship extraction from readers' literary book reviews on goodreads.com," *arXiv preprint*, vol. arXiv:2004.09601, 2020. [Online]. Available: https://arxiv.org/abs/2004.09601

[12] L. Hajibayova, "Investigation of goodreads' reviews: Kakutanied, deceived or simply honest?" *ResearchGate*, 2019, available at: https://www.researchgate.net/publication/331384350.

VII. Appendix



True label \ Predicted label	1	2	3	4	5
1	649	1238	1371	384	391
2	31	3604	4595	1418	639
3	39	542	17671	5990	2278
4	75	398	8191	25468	9183
5	56	154	2423	7716	25493

Fig. 1. A confusion matrix displaying the effectiveness of a sentiment analyzer using multinomial Naïve Bayes to assign star ratings to user reviews.

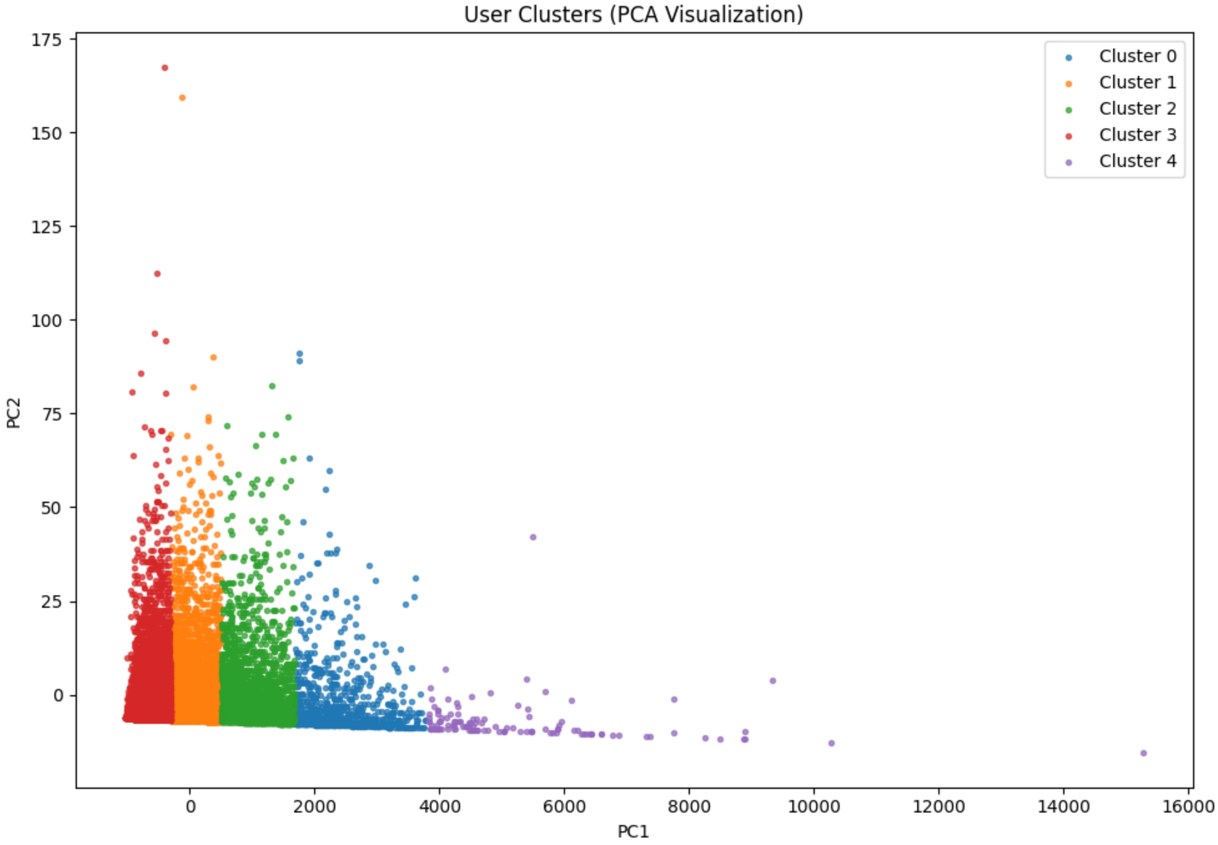


Fig. 2. A rinciple components analysis (PCA) of the clustering component of our model. PC1 seems to reflect the frequency of reviews/length of reviews, while PC2 seems to reflect variance in rating behavior.