Goodreads Report 1

Matthew D. Branson, James R. Brown

Department of Computer Science

Missouri State University

Springfield, USA

branson773@live.missouristate.edu

Abstract

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). We are currently in an exploratory analysis phase, using a PyQt6 application for dataset downloading and normalization into a SQL database. Our preparation can be observed at https://github.com/mdb42/goodreads.

II. Literature Review

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 [1]. 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 [2]. 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're expecting a much more informal style of writing in Goodreads reviews, we'll 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), too. Sokolova and Bobicev attempted to use sentiment analysis on medical forums to evaluate the presence of the "echo chamber effect" in those forums [3]. 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 [4]. 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.

That being said, 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 [5]. 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 [6]. 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.

III. Background

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. Term frequency–inverse document frequency (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 Naive 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 Naive Bayes: To predict user star ratings based on review text, we will train a Multinomial Naive 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 F1-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. Task Assignments

References

[1] 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.

[2] 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

[3] 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

[4] 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/

[5] 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.

[6] 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.