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

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'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 [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.

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 [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.

III. Background

To construct the sentiment analyzer, we'll 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. From these, we're looking to have the model generate an average sentiment score on a scale from 1 through 5, matching up with the five-star review score that users can give a book on Goodreads. After this, these reviews can be clustered by sentiment score.

IV. Methods

V. Task Assignments

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.