Goodreads Sentiment Analysis

Matthew D. Branson, James R. Brown

Department of Computer Science

Missouri State University

Springfield, USA

branson773@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). 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. Project Objectives

We intend to perform the following:

- Build a sentiment classifier, predicting star ratings given a body of review text.
- Analyze the effectiveness of the sentiment classifier not just on validation sets of Goodreads reviews, but also as a case study in domain adaptation when applied to movie reviews.
- Build other classifiers for multiple dimensions of reviews and reviewers themselves, examining features such as intensity, focus, constructiveness, subjectivity, and consistency.
- Perform clustering analysis from user-generated shelving descriptions to define formal categories for genre, performing then genre-specific sentiment analysis.
- Build a parametric search interface for the entire dataset, though this would rely more on SQLite indexing than any custom method in memory, simply for sake of the sheer immensity of the collection.

III. Literature Review

Our project primarily builds upon the work of Wan et al., 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] develop 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.

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.