2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

# Book Reading Behavior on Goodreads Can Predict the Amazon Best Sellers

Suman Kalyan Maity, Abhishek Panigrahi and Animesh Mukherjee
Dept. of CSE, IIT Kharagpur, India
Email: sumankalyan.maity@cse.iitkgp.ernet.in

***Abstract*—Success of a book depends on various intrinsic and extrinsic parameters. We perform a cross-platform study of book reading behavior on Goodreads and attempt to establish the connection between the Goodreads entities and the Amazon best sellers. We analyze the collective reading behavior on Goodreads platform and quantify various characteristic features of the Goodreads entities to identify differences between the Amazon best sellers (ABS) and the other non-best selling books. Using these features we devise a model to predict if a book shall become a best seller after one month (15 days) since its publication. On a balanced set, we are able to achieve a very high avg. accuracy of 88.72% (85.66%) for the prediction where the other competitive class contains books which are randomly selected from the Goodreads dataset. Our method primarily based on features derived from user posts and genre related characteristic properties achieves an improvement of 16.4% over the traditional popularity factors (ratings, reviews) based baseline methods.**

## I. Introduction

Goodreads is a popular social book-reading platform which allow book-lovers to share books they have read, review books, rate books and connect with other readers. On Goodreads website, users can add books to their personal bookshelves for reading, track the status of their readings and post a reading status, find which books their friends and favorite authors are reading, participate in discussions and take part in group activities on various topics, and, as well, get personalized book recommendations. Further, Goodreads provides a platform for social interactions among users; once an user adds another user as a friend, one can view his/her friends' shelves and reviews and comment on friends' pages. Goodreads features a 5-star rating system along with text reviews. The Goodreads readers also receive regular newsletter featuring new books, suggestions, author interviews etc.

Popularity of a book depends on various factors. They can be broadly classified into two groups - intrinsic or innate content factors and external factors. Intrinsic content factors mostly concern quality of books that include how interesting it is, the novelty, the writing style, the engaging story-line etc., in general. However, these content and quality factors of books are very different for different genres. For example, a successful thriller requires a credible, big story-line, complex twists and plots, escalating stakes and tension on every page whereas a popular romantic novel does not require complex twists and plots, tension or shock effect; what it requires are variety, demonstration of strong and healthy relationship, sexual tension etc. [1]. Therefore, it is very difficult to find common grounds for books belonging to various genres and to quantify those aspects. External factors driving books' popularity include the readers' reading behaviors, social contexts, reviews by the critics etc. Some of the early works [2], [3], [4] provide quantitative insights to stylistic aspects based on human expertise on literature. Ashok et al. [5] focus on writing styles of the novels and establish the connection between stylistic elements and the literary success of novels providing quantitative insights to them. In this work, we try to quantify the external factors of books' popularity by analyzing the characteristics of **book-reading habits** as reflected on the Goodreads platform. In particular, we are interested to understand **if the collective reading habits on Goodreads can distinguish the Amazon best sellers from the rest of the books**.

Analysis of reading habits has been an active area of research for quite long time [6], [7], [8], [9], [10]. While most of this research investigates blog reading behavior [6], [7], [9], there have been some work that also discuss about interactive and connected book reading behavior [8], [10]. Despite such active research, very little investigation has been done so far to understand the characteristics of social book reading sites and how the analysis of the collective reading phenomena can even influence the online sales of books. On Goodreads platform, there has been few attempts to understand message posting and users' reading characteristics [11], [12], [13]. Our work differs from these above in following ways. We study the characteristics of book reading behavior on Goodreads and try to establish whether these factors can discriminate between Amazon best sellers from other books. We use both the external characteristics of a book as well as the content of the reviews in our study. To the best of our knowledge, we are the first who try to explicitly provide quantitative insights, based on collective reading habits, on the unstudied connection between the entities of a book-reading platform (Goodreads) and the success of a book (best-sellers on Amazon).



*ASONAM '17*, July 31 - August 03, 2017, Sydney, Australia

ACM ISBN 978-1-4503-4993-2/17/07/$15.00
http://dx.doi.org/10.1145/3110025.3110138

## II. Dataset Preparation

We obtain our Goodreads dataset through APIs and web-based crawls over a period of 9 months. This crawling exercise has resulted in the accumulation of a massive dataset spanning a period of around nine years. We first identify the unique genres from https://www.goodreads.com/genres/list. Next we collect unique books from the above list of genres and different information regarding these books are crawled via Goodreads APIs. Each book has information like the name of the author, the published year, no. of ratings it received, avg. rating, no. of reviews etc. In total, we could retrieve information of 558,563 books. We then find out the authors of these books and their information like no. of distinct works, avg. rating, no. of ratings, no. of fans etc. In total, we have information of 332,253 authors. We separately collect the yearly Amazon best sellers[1] from 1995 to 2016 and their ISBNs and then re-crawl Goodreads if relevant information about some of them is not already present in the crawled dataset. For these books, we separately crawl upto 2000 reviews and ratings in chronological order.

## III. Characteristics of book reading behavior

In this section, we shall study the characteristic properties of various book reading aspects for the Amazon best sellers and compare them with the rest of the books. We have 1468 Amazon best sellers in our dataset. To compare with the rest of the books, we choose random samples of books from the entire set of books minus the Amazon best sellers. We obtain ten such random samples and report averaged results for them.

### *Goodreads users' status posts*

While reading, a Goodreads user can post status updates regarding the book. We separately crawl the first 2000 user status posts for each book in our dataset. These book reading status postings for a book also drive its popularity. We attempt to differentiate the Amazon best sellers with the other Goodreads books through the reading status postings. In fig 1 (a), we show the distribution of the number of status update posts per users. The results show that while reading the Amazon best seller books, readers tend to post status updates more often as compared to the readers of other books. Fig 1 (b) presents the distribution of unique readers posting status updates. The Amazon best sellers engage more readers in posting status updates compared to other Goodreads books. Also, the Amazon best sellers engages better the same readers in posting multiple status updates compared to other Goodreads books (see fig 1 (c)). We also study the distribution of avg. inter-status arrival time (see fig 1 (d)) which shows that readers of Amazon best sellers post status updates more frequently (for more than $35\%$ books, avg inter-status arrival time is less than a day) than the readers of other Goodreads books. Fig 1 (e) shows the distribution of avg. maximum concentrated reading efforts at a stretch (in terms of percentage read). Readers of $\sim 80\%$ of the Amazon best selling books show maximum percentage stretch of read as 20-40%. Though, for the other Goodreads books also, the fraction is largest for the same zone, the relative number of books in this stretch is lesser than that of the best sellers. In fig 1 (f), we show the distribution of avg. time to finish book reading. We observe that for $\sim 63\%$ of the Amazon best sellers there are no readers who have completed reading the whole book whereas for the other books, this number is quite high ($\sim 94\%$). Among the books where one of the readers have at least finished reading the book, the fraction of books for Amazon best sellers are higher compared to the other books in all the time buckets.

[1]http://www.amazon.com/gp/bestsellers/1995/books

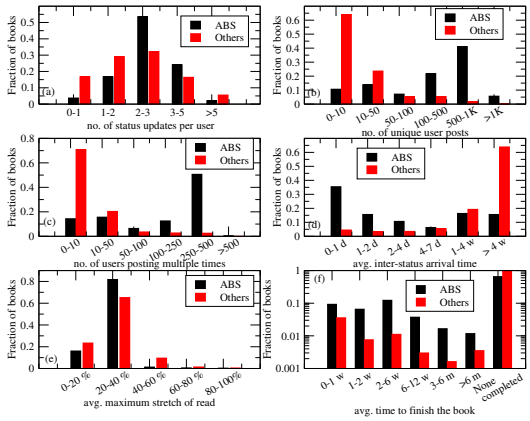


**Fig. 1:** Characteristic properties of Goodreads users' status posts: distribution of a) no. of status updates per user b) no. of unique users updating status c) no. of users updating multiple times d) inter-status arrival time e) avg. maximum stretch of reading f) avg. time to finish reading for ABS vs other books.

## IV. Will a book become an Amazon best seller?

From the discussions in the previous sections, it is evident that there exist differences among various Goodreads book reading aspects for Amazon best sellers and the other Goodreads books. In this section, we attempt to build a prediction framework that can early predict whether a book will be an Amazon best sellers or not. Our goal is to predict whether a book will be an Amazon best seller or not just by observing data upto various time periods from Goodreads starting from the date of publication of the book ($t$ = 15 days, 1 month). As Goodreads was launched in 2007, we consider only those Amazon best seller that are published on or after 2007. To compare against these set of books, we consider a same sized random sample of books from Goodreads, none of which ever became an Amazon best seller and all of which have been published after 2007. In total, we have $\sim 380$ books in each class. For the task of prediction, we consider the following set of features:

- **Novelty of the book:** Novelty of a book is a key for its acceptability/success in the readers' circle. For each book in Goodreads, a short summary about the book is provided. We separately crawl this 'About' information of the book (say, the document containing the summary of the book be $A$) and of all the other books that are published before this book in question (say, the concatenated summary of all those books be $B$). We then

extract keywords[2] from documents $A$ and $B$ respectively ($Keywords_A$ and $Keywords_B$) and design a metric of keyword overlap as follows:

$$Overlap(A,B) = \frac{|Keywords_A \cap Keywords_B|}{min(|Keywords_A|, |Keywords_B|)}$$

Higher the keyword overlap, lower is the novelty score. We also define another novelty feature which is measured as the Kullback-Leibler (KL) divergence between the unigram language model inferred from the document $A$ containing all the words except the stop words from the book's summary for the $i^{th}$ book and the background language model from document $B$ and formally defined this as follows:

$$KLDiv(i) = -\sum_{w \in A} p(w|A) \times \log \frac{p(w|A)}{p(w|B)}$$

Higher the divergence value, higher is the novelty of the book. Once again this is used a feature in the prediction model.

- **Genres** - Genres of a book is an important feature in deciding the destiny of it. In this prediction model, we consider top 15 genres for Amazon best sellers and other books. In total we have 24 genres - fantasy, fiction, nonfiction, children, romance etc. We use each of them as a binary feature.

**Goodreads book reading status posts:** We compute several features out these book reading posts. The list is as follows:

- number of status updates of the readers (mean, min, max and variance),
- number of unique readers posting status updates,
- number of readers posting status updates more than once (twice/thrice/five times),
- inter-status arrival time (mean, min, max and variance),
- maximum percentage stretch of read (mean, min, max and variance). We also use the maximum stretch of read in terms of time.
- rate of reading of the readers (mean, min, max and variance),
- fastest rate of reading (mean, min, max and variance),
- time taken to finish reading the book (mean, min, max and variance),
- average positive and negative sentiments from the status posts.

***Baseline features:***

Ratings and reviews of book are the most common indicators of popularity and in this paper, we shall consider features extracted from them as baselines and compare them with the more non-trivial features related to reading behavior of users for the task of prediction. The ratings and reviews related features are described below

- average number of rating
- number of 1-star ratings, 2-star ratings, 3-star ratings, 4-star ratings, 5-star ratings
- rating entropy
- number of reviews received
- **Sentiment of the reviews:** For each book, we concatenate all the reviews in one month and find out the fraction of positive sentiment words (positive sentiment score) and the fraction of negative sentiment words (negative sentiment score) by using MPQA sentiment lexicon[14]. We use these two sentiment scores as two separate features.
- **Cognitive dimension of the reviews:** There could be differences in the cognitive dimension (linguistic and psychological) for the two category of books. To quantify this, we consider Linguistic Inquiry and Word Count (LIWC [15]) software. LIWC takes a text document as input and outputs a score for the input over all the categories based on the writing style and psychometric properties of the document.

### A. Performance of the prediction model

In this subsection, we shall discuss the performance of our prediction model. We use 10-fold cross-validation technique and use SVM and logistic regression classifier [16]. For the prediction task, we consider $t$ time periods ($t$) - 15 days, 1 month from the publication date. We compute all the feature values from the data available only within the time period $t$ from the publication date. We ensure that all the books that we select in both the classes are published after 2007 since Goodreads was launched in 2007. Table I shows the various classification techniques we employ and the evaluation results. The classifiers yield very similar classification performance with logistic regression performing little better; with logistic regression classifier, we obtain avg. accuracy of 88.72% with avg. precision and recall of 0.887 each and the avg. area under the ROC curve as 0.925 for $t = 1$ month on a balanced dataset with 10-fold cross-validation method. Note that the classification results for other time period also give very similar results. The user status and genre based features are most prominent ones and significantly outperforms the ratings and review feature based baselines. For $t$ = 1 month, our method yields 16.4% improvement over the best performing baseline (for $t$ = 15 days, we also yield similar improvement) suggesting that user's status on Goodreads are very important indicators of popularity and are, in fact, much better indicators than reviews or ratings. In other words, this shows that all Amazon best seller books might not necessarily have high quality reviews or a high volume of ratings; however, a large majority of them have user status post patterns very different from the other set of books.

Since in real-life, the proportion of the Amazon best sellers is far lower than the other types of books, we also consider testing our model on an unbalanced test set. Here, the training and test sample sets are taken in 3:1 ratio. In training set, both the class samples are taken in equal proportion (to guarantee fair learning) whereas in test sample the Amazon best sellers and the other books are taken in 1:9 ratio. We then train our classifiers on the balanced training set and test on the unbalanced one. We report the weighted avg. values for all the metrics in table I. For an observation period of even as small as 15 days, we achieve weighted avg. accuracy of $\sim 86.67\%$ with weighted avg. precision of $0.901$ and recall of $0.876$. Note that compared to the balanced set, the performance is

[2]https://github.com/aneesha/RAKE

slightly better. The weighted avg. ROC area under the curve is quite high (0.963 on weighted averaging - but same value is found for the individual classes also).

**TABLE I:** Evaluation results with comparison with baselines (baseline1 - ratings, baseline2 - reviews)

| $t$ | Method | Accuracy | Precision | Recall | F-Score | ROC Area |
|---|---|---|---|---|---|---|
| 15 days | LR | **85.66%** | 0.857 | 0.857 | 0.857 | **0.917** |
| | SVM | 85.3% | 0.853 | 0.853 | 0.853 | 0.853 |
| | Baseline1 - LR | 76.7% | 0.774 | 0.767 | 0.766 | 0.826 |
| | Baseline2 - LR | 75.98% | 0.76 | 0.76 | 0.76 | 0.829 |
| 1 month | LR | **88.72%** | 0.888 | 0.887 | 0.887 | **0.925** |
| | SVM | 88.71% | 0.888 | 0.887 | 0.887 | 0.887 |
| | Baseline1 - LR | 76.22% | 0.766 | 0.762 | 0.762 | 0.808 |
| | Baseline2 - LR | 75.91% | 0.76 | 0.759 | 0.759 | 0.812 |
| Unbalanced testset (t = 15 d) | LR | **86.67%** | 0.901 | 0.867 | 0.876 | **0.963** |
| | SVM | 86.67% | 0.924 | 0.867 | 0.879 | 0.919 |
| | Baseline1 - LR | 75.56% | 0.897 | 0.756 | 0.784 | 0.851 |
| | Baseline2 - LR | 75.5% | 0.839 | 0.756 | 0.78 | 0.78 |

*Discriminative power of the features*

Here, we shall discuss about the importance of the individual features (i.e., the discriminative power of the individual features). In order to determine the discriminative power of each feature, we compute the chi-square ($\chi^2$) value and the information gain. Table II shows the order of all features based on the $\chi^2$ value, where larger the value, higher is the discriminative power. The ranks of the features are very similar when ranked by information gain (Kullback-Leibler divergence). The most prominent features on individual level are the user status features. There are several user status features that come in list of top 15. Among those status features, the most discriminative ones are mean reading rate, no. of status posts, fastest rate of reading, inter-status post time etc.

**TABLE II:** Features and their discriminative power.

| $\chi^2$ Value | Rank | Feature |
|---|---|---|
| 194.0324 | 1 | mean reading rate |
| 121.7847 | 2 | no. of readers posting status |
| 109.8861 | 3 | fastest rate of reading (min) |
| 96.0871 | 4 | inter-status post time (min) |
| 92.1124 | 5 | min. reading rate |
| 88.2171 | 6 | no. of readers posting status updates twice |
| 82.5114 | 7 | variance of reading rate |
| 75.9434 | 8 | maximum percentage stretch of read (min) |
| 75.0442 | 9 | no. of readers posting status updates thrice |
| 74.0946 | 10 | inter-status post time (mean) |
| 71.7962 | 11 | time taken to finish the book (max) |
| 71.7024 | 12 | maximum percentage stretch of read w.r.t time (max) |
| 69.5682 | 13 | no. of readers posting status updates five times |
| 63.3696 | 14 | genre (Romance) |
| 56.759 | 15 | time taken to finish the book (variance) |

## V. Conclusions and Future Works

In this paper, we study the characteristic properties of Amazon best sellers in terms of reading habits by analyzing a large Goodreads dataset. We observe that there exist characteristic differences between the Amazon best sellers and the other books. We then use these characteristic properties as features for a prediction model that attempts to predict whether a book will be an Amazon best seller or not.

Our proposed prediction framework achieves a very high avg. accuracy of **88.72**% with high avg. precision and recall (**0.887**) for observation time period $t$ = 1 month. Our results also hold true for an unbalanced test data set. We observe that the user status post features are the most discriminative ones.

There are quite a few other interesting directions that can be explored in future. One such direction could be to understand the detailed user reading dynamics focusing on various inter-dependent entities like shelves and user status posts. We are also interested in performing other cross-platform study to understand this unique dynamics between the platforms in more detail.

## References


[1] J. W. Hall, *Hit Lit: Cracking the Code of the Twentieth Century's Biggest Bestsellers*. Random House, 2012.

[2] A. Ellegård, *A Statistical method for determining authorship: the Junius Letters, 1769-1772*. Göteborg: Acta Universitatis Gothoburgensis, 1962, vol. 13.

[3] J. Harvey, "The content characteristics of best-selling novels," *Public Opinion Quarterly*, vol. 17, no. 1, pp. 91–114, 1953.

[4] J. J. McGann, *The poetics of sensibility: a revolution in literary style*. Oxford University Press, 1998.

[5] V. G. Ashok, S. Feng, and Y. Choi, "Success with style: Using writing style to predict the success of novels," in *Proc. of EMNLP*, 2013, pp. 1753–1764.

[6] E. Baumer, M. Sueyoshi, and B. Tomlinson, "Exploring the role of the reader in the activity of blogging," in *CHI*, 2008, pp. 1111–1120.

[7] E. P. Baumer, M. Sueyoshi, and B. Tomlinson, "Bloggers and readers blogging together: Collaborative co-creation of political blogs," *Comput. Supported Coop. Work*, vol. 20, no. 1-2, pp. 1–36, Apr. 2011.

[8] S. Follmer, R. T. Ballagas, H. Raffle, M. Spasojevic, and H. Ishii, "People in books: Using a flashcam to become part of an interactive book for connected reading," in *CSCW*, 2012, pp. 685–694.

[9] B. A. Nardi, D. J. Schiano, and M. Gumbrecht, "Blogging as social activity, or, would you let 900 million people read your diary?" in *CSCW*, 2004, pp. 222–231.

[10] H. Raffle, R. Ballagas, G. Revelle, H. Horii, S. Follmer, J. Go, E. Reardon, K. Mori, J. Kaye, and M. Spasojevic, "Family story play: Reading with young children (and elmo) over a distance," in *CHI*, 2010, pp. 1583–1592.

[11] S. Dimitrov, F. Zamal, A. Piper, and D. Ruths, "Goodreads vs amazon: The effect of decoupling book reviewing and book selling," in *Proc. of ICWSM '15*, 2015.

[12] A. Worrall, ""back onto the tracks": Convergent community boundaries in librarything and goodreads," in *9th Annual Social Informatics Research Symposium*, 2013.

[13] M. Thelwal and K. Kousha, "Goodreads: A social network site for book readers," *Journal of the Association for Information Science and Technology*, 2016.

[14] L. Deng and J. Wiebe, "Mpqa 3.0: An entity/event-level sentiment corpus," in *NAACL '15*, 2015.

[15] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.