Modeling Readers' Appreciation of Literary Narratives Through Sentiment Arcs and Semantic Profiles

Yuri Bizzoni
Center for Humanities Computing /
Aarhus University, Denmark
yuri.bizzoni@cc.au.dk

Pascale Feldkamp Moreira
Comparative Literature,
School of Communication and Culture /
Aarhus University, Denmark
pascale.moreira@cc.au.dk

Mads Rosendahl Thomsen
Comparative Literature,
School of Communication and Culture /
Aarhus University, Denmark
madsrt@cc.au.dk

Kristoffer L. Nielbo
Center for Humanities Computing /
Aarhus University, Denmark
kln@cas.au.dk

Abstract

Predicting the perception of literary quality and reader appreciation of narrative texts are highly complex challenges in quantitative and computational literary studies due to the fluid definitions of quality and the vast feature space that can be considered when modeling a literary work. This paper investigates the potential of sentiment arcs combined with topical-semantic profiling of literary narratives as indicators for their literary quality. Our experiments focus on a large corpus of 19th and 20^{the} century English language literary fiction, using GoodReads' ratings as an imperfect approximation of the diverse range of reader evaluations and preferences. By leveraging a stacked ensemble of regression models, we achieve a promising performance in predicting average readers' scores, indicating the potential of our approach in modeling perceived literary quality.

1 Introduction

Defining what contributes to the perceived literary quality of narrative texts (or lack thereof) is an ancient and highly complex challenge of quantitative literary studies. The versatility of narrative and the myriad of possible definitions of a text's quality ultimately complicate the issue. In addition, the diversity and size of the possible feature space for modeling a literary work contribute to the complexity of the matter. It can even be argued that the quality of a literary text is not systematic and that "quality" is an expression of noisy preferences, as it mostly encodes idiosyncratic tastes that depend on individual reader inclinations and capacities. However, various studies have shown that this 'literary preference as noise' position is not tenable because text-intrinsic features (e.g., text coherence, literary style) and text-extrinsic factors (e.g., reader demographics) systematically impact perceived literary quality (Mohseni et al., 2021; Koolen et al., 2020a; Bizzoni et al., 2022b). At the same time, the questions of how such features interplay and what kind of metric we should use to validate them remain open. Thus, current research on the perception of literary quality implicitly tries to answer two primary questions: 1) Is it possible to define literary quality at all, and 2) Is it possible to identify the intrinsic or extrinsic features that contribute to the perception of literary quality? While quality as a single measure may be impossible to agree on (Bizzoni et al., 2022a), it is hard to refute that reader preferences can be measured in different ways, both in terms of consistent attention given to literary works over time, and to valuations made by critics and readers. The intrinsic qualities of texts are more difficult to agree upon as the quality of a literary work consist of many elements, some that are virtually impossible to grasp by computational methods (e.g. the effect of metaphors or images). In addition, there are text-extrinsic features, such as the public image of the author or author-gender (Wang et al., 2019; Lassen et al., 2022), which influence reviews to a degree that is hard to account for. Still, as mentioned there is evidence that intrinsic models do have some predictive value when considering an array of different features , which pertain to both style and narrative. As such, the difficulty is not to only to model literary quality, as including intrinsic and extrinsic features such as genre and author-gender in a models of quality has resulted

Proceedings of the The 5th Workshop on Narrative Understanding, pages 25–35
July 14, 2023 ©2023 Association for Computational Linguistics

in good performances (Koolen et al., 2020a) – but in elucidating what to include in a feature-set and why, and in seeking a level of interpretability.

In this study, we aim to investigate the relationship between a narrative's emotional trajectory, its fine-grained semantic profile, and its perceived literary quality. Using the average of hundreds of thousands of readers' ratings, we examine how sentiment-arcs and semantic profiles of literary narratives influence their perceived quality, exploring the prediction of these factors through a machine learning model trained on multiple features, encompassing both sentiment-related aspects and their dynamic progression, as well as semantic categorization. We also claim that access to a diverse corpus of works with a significant representation of highly successful works in all genres is an essential prerequisite for developing models with a credible performance. Without the inclusion of the best regarded works, it is not possible to produce a model that relates to what is commonly understood as the highest level of literary achievement. The 9,000 novels corpus used in our study contains several of such works from 1880 to 2000, including major modernist and postmodernist writers as well as fiction from a range of popular genres.

2 Related works

Studies that predict the perception of literary quality from textual features have primarily relied on classical stylometric features, such as sentence-length or readability (Koolen et al., 2020b; Maharjan et al., 2017), the percentage of word classes, such as adverbs or nouns (Koolen et al., 2020b) or the frequencies of n-grams in the texts (van Cranenburgh and Koolen, 2020). More recent work has tested the potential of alternative text or narrative features such as sentiment analysis (Alm, 2008; Jain et al., 2017) as a proxy for meaningful aspects of the reading experience (Drobot, 2013; Cambria et al., 2017; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017; Reagan et al., 2016a). Such work has focused on sentiment valence, usually drawing scores from induced lexica (Islam et al., 2020) or human annotations (Mohammad and Turney, 2013), modeling, for instance, novels' sentiment arcs (Jockers, 2017), although without considering fundamental arc-dynamics (e.g., temporal structure of plot variability) or narrative progression. By simply clustering sentiment arcs, Reagan et al. (2016a) was however able to identify six fundamental narrative arcs that underly narrative construction, while more recently, Hu et al. (2021) and Bizzoni et al. (2022b) have modeled the persistence, coherence, and predictability of sentiment arcs by fractal analysis, a method to study the dynamics of complex systems (Hu et al., 2009; Gao and Xu, 2021) and to assess the predictability and self-similarity of arcs, in order to model the relation of sentiment arcs with reader evaluation (Bizzoni et al., 2021, 2022c). Similarly, Mohseni et al. (2021) conducted fractal analysis on classical stylometric and topical features to model the difference between canonical and non-canonical literary texts. Beyond sentiment analysis, the narrative content of texts has also been shown to impact perceived quality. Relying on topic modeling, Jautze et al. (2016) has shown that a higher topic diversity in texts corresponds to higher perceived literary quality, suggesting that works with a less diverse topical palette, like genre fiction, are perceived as having overall less literary quality, while van Cranenburgh et al. (2019) has claimed that words that refer to intimate and familiar relations are distinctive of lower-rated novels, which can be linked to the hypothesis that specific genres, especially those in which women authors are dominant, are perceived as less literary (Koolen, 2018). These studies suggest that the distribution of topics touched upon in texts impacts literary quality perception. Several works have widely used resources like LIWC to model such distributions (Luoto and van Cranenburgh, 2021a; Naber and Boot, 2019). However, building on the findings of Jarmasz (2012) – i.e., that Roget's thesaurus is an excellent resource for natural language processing – Jannatus Saba et al. (2021) has shown that Roget outperforms other dictionary resources (e.g., LIWC and NRC sentiment lexicons) in modeling literary quality by category frequency – which is an intriguing argument to use the Roget categories for modelling the perception of literary quality on a larger scale.

3 Quality measures

While it is clear that various studies have recently used conceptually different features as a basis for understanding or predicting perceived literary quality, reader appreciation, or success, it should be noted that each study has a slightly different take on "quality" and that terms like "prestige", "popularity", or "canonicity" are not synonymous - although they could all be argued as aspects of quality, and a more comprehensive study would benefit from taking a stronger perspectivist approach, considering multiple definitions of quality together (Bizzoni et al., 2022a). For any study trying to assess the factors contributing to the perception of literary quality, determining the quality judgments themselves is often one, if not the first, of the most challenging tasks. Computational studies assessing literary quality often use a single standard of evaluation, which may not capture the diverse preferences of various groups of readers. Various quality measures have been used, such as readers' ratings on platforms such as GoodReads (Kousha et al., 2017), or a text's presence in established literary canons (Wilkens, 2012). Despite their diversity, different conceptions of quality can display significant convergences (Walsh and Antoniak, 2021). In this work, we have employed average book-ratings on **Goodreads**, a popular online social platform for readers that allows users, among other things, to comment, recommend, and review a book on a scale. [1] This metric possesses obvious limitations: it doesn't explicitly represent "literary quality" but arguably an aspect of it, it potentially conflates genre-specific value-judgements, and it forces GoodReads' users to reduce their literary evaluations to a mono-dimensional scale. The latter issue might also obscure important differences in rating behaviour. For example, readers of Sci-fi may be inclined to give a higher average rating on GoodReads, something that we do not take into account when using average rating as a quality metric. Nevertheless, this limitation can also be an advantage: the simplicity of the GoodReads rating system offers a streamlined approach to a problem that frequently proves overly complex for quantitative analysis. The single GoodReads' rating, representing readers' impressions on a single scale, offers a practical starting point for identifying patterns or trends across a wide range of books, genres, and authors.

On the other hand, with its 90 million users, GoodReads is argued to offer a particularly valuable insight into reading culture "in the wild" (Nakamura, 2013), as it collects books from widely different genres and curricula (Walsh and Antoniak, 2021), and derives ratings from a notably heterogeneous pool of readers in regard to backgrounds, gender, age, native languages and reading preferences (Kousha et al., 2017).

4 Data

We have used the Chicago Corpus as a dataset, encompassing more than 9,000 English-language novels penned or translated into English between 1880 and 2000. The selection criterion for these works is based on each novel's number of libraries holdings. This results in a diverse compilation that spans various literary styles, from popular fiction genres to highly esteemed works of literature. It comprises novels written by Nobel Prize laureates (Bizzoni et al., 2022c) and recipients of other highly regarded literary awards, as well as texts featured in canonical collections such as the Norton Anthology (Shesgreen, 2009). However, it is important to acknowledge the cultural and geographical bias present in the corpus, which exhibits a significant over-representation of Anglophone authors, limiting the scope of the analysis to a predominantly English-speaking context.

	Titles	**Authors**
Number	9089	3150
Avg. rating below 2.5	140	118
Avg. ratings	3.74	3.69

Table 1: Number of titles and authors in the corpus and below the rating of 2.5, and avg. number of ratings

5 Features

We employ three types of features, representing three distinct approaches to modeling a literary narrative. See Table 1 for a summary.

5.1 Sentiment features

We perform a simple sentiment analysis of the novels, extracting the **VADER** (Hutto and Gilbert, 2014) compound sentimental score of each sentence after tokenizing the texts with nltk (Bird, 2006). We selected this model as it is based on a lexicon and set of rules, and so remains relatively transparent. Although it was developed for social media analysis, VADER is widely employed and exhibits a good performance and consistency across domains (Ribeiro et al., 2016; Reagan et al., 2016b). When dealing with narrative, this versatility is especially valuable, as considering our corpus, we are also comparing texts across widely different (literary) genres. Moreover, the sentiment arcs resulting from VADER appear comparable to those of the **Syuzet-package** (Elkins and Chun, 2019),

[1] https://www.goodreads.com

which was developed for literary texts (Jockers, 2017). Yet, in using VADER we side-step some of the problems of the Syuzet-package, like of word-based annotation (Swafford, 2015). To ensure the validity of our annotation, we manually inspected a selection of novels both at the sentence and arc level (e.g., fig. 1). Using VADER, the result is a rather fine-grained sentiment arc that, when detrended, roughly describes the overall evolution of the storyline, as shown in Figure 1 (also see Hu et al. (2021) and Bizzoni et al. (2021) for more details on this method).

By examining the mean sentiment and its standard deviation for an entire novel and its subsections (e.g., the first or last ten percent), we can create a coarse representation of the narrative's emotional profile. In this study, we divide each sentiment arc into 20 segments and calculate the mean sentiment for each segment. Additionally, we include the overall sentiment mean and standard deviation as features. This approach allows for a rudimentary characterization of the sentiment-profile of the novel.

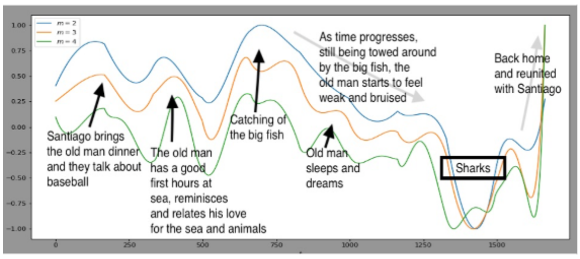

Figure 1: Sentiment arc of Hemingway's *The Old Man and the Sea* with different polynomial fits (m = polynomial degree). Values on the y-axis represent compound sentiment score as annotated with VADER, while values on the x-axis represent the narrative progression of the book by the number of sentences.

5.2 Dynamic features

As the most important aspect of a narrative arguably relies on its dynamic development rather than in its global characteristic, we relied on two measures to try and capture the high-level properties of the narratives' sentiment arcs, rather than their simple states: For each sentiment arc we computed its Hurst exponent, which represents the degree of time series persistence; and its approximate entropy, which represents the level of predictability of a series. The Hurst exponent is a measure that quantifies the persistence, or long-range dependence, of a time series, where a higher value indicates stronger trend-following behavior and a lower value represents a more anti-persistent or mean-reverting pattern. Hurst estimates of several time-dependent textual features, including narrative sentiment arcs, have been proven predictive of literary quality perception in several recent studies (Bizzoni et al., 2022b,c; Mohseni et al., 2021). Approximate entropy is a metric that evaluates the predictability of a time series by assessing the regularity and complexity of its fluctuations, with lower values indicating more predictable and repetitive patterns. In comparison, higher entropy values suggest greater randomness and unpredictability in the series. Approximate entropy has also been linked to aspects of literary quality perception (Mohseni et al., 2022).

5.3 Roget features

The aim of Roget's thesaurus was semantic classification, closely related to similar projects in areas like biology during the Victorian era, by scientists who – like Roget – were members of the Royal Society (Liddy et al., 1990). Yet the thesaurus also had an explicitly literary aim: to aid literary composition, not only as a tool to query for words and synonyms, but also as a tool for grasping "the relation which these symbols [i.e., words] bear to their corresponding ideas" (Roget, 1962). The classification scheme of the thesaurus follows six major divisions: affection, volition, intellect, abstract relations, space, and matter (Roget, 1997); each of these subdivided into three to eight subheadings, and further divided into "paragraphs". For example, "memory" with its connected words is a paragraph in the subdivision "extension of thought" within the major category of "intellect". As such, Roget-categories are semi-topical and do in a sense reflect the distribution of ideational content in literary works.

We used the Roget thesaurus of English words to construct topical representations of each narrative as the interplay of different themes with different strengths. In other words, we used the Roget thesaurus, that links each word in its collection to one or more topical-semantic categories, to derive a word-based representation of the topics "touched" by a novel (even through one single metaphoric word) and with which frequency they were mentioned. For example, the sentence

```
He walked the dog
```

would be linked to the categories of *Motion* (walked), *Animal* (dog) and so forth. While the Roget thesaurus is in this respect not dissimilar from several other thesauri built to attempt a rough hyerarchization of words into concepts (see WordNet for a more modern example) we chose it due to its apparent suitability to model literary texts, as discussed in Section 2. The thesaurus was originally built around 1805 by M. R. Roget as a compilation of English language words into hyerarchical semantic clusters that would help a writer find the most apt words for their ideas. The thesaurus was partly inspired by Leibniz's symoblic languages and by Aristotle's categories, and has since its appeareance been regularly revised and increased; its most recent edition contains more than 400.000 words.

We computed how many words in a book belonged to each Roget "paragraph" (i.e., topics in each subcategory), adding the result to our feature set. While the validity of the Roget categories is questionable at linguistic and cognitive levels – like any single-handed categorization of semantics – we selected this representation due to the somewhat surprising accuracy it has demonstrated in modeling the success of literary narratives in recent studies (Saba et al., 2021; Luoto and van Cranenburgh, 2021b).

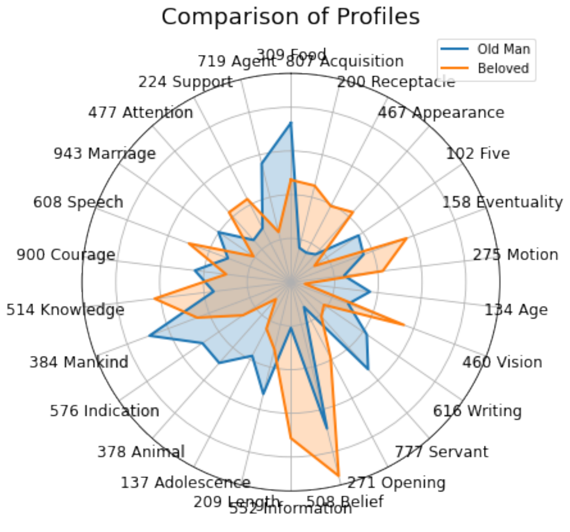

Figure 2: Profiles of Hemingway's *The Old Man and the Sea* and Morrison's *Beloved* along their most frequent categories. Hemingway's masterpiece draws on categories of food, age, animals and adolescence more than Morrison's novel, that instead peaks on speech, belief, vision and appearance.

5.4 Feature Selection

Before training a supervised prediction model on the dataset, we perform feature selection to reduce the size of the feature set and improve the interpretability of the final results. We use a filter method for feature selection (John et al., 1994), which ranks each possible feature based on a relevance weight. It then optimizes the list, shortening it to improve the model selection. The filter method of feature selection evaluates each feature independently based on a specific criterion, such as information gain or correlation with the target variable, and thus allows for the identification of the most relevant features and discarding the less important ones, ultimately leading to a reduced and more meaningful feature set for model training.

Category	Description	Number
Sentiment	mean, std SA	22
Dynamic	Hurst, AppEnt	2
Semantic	Roget categories	1044

Table 2: Feature categories and corresponding numbers.

6 Models

For our prediction task, we used a stacked ensemble model featuring a Support Vector Machine-based regressor (SVR) (Cortes and Vapnik, 1995) and a Random Forest regressor (Breiman, 2001), with a Ridge regressor as a meta-classifier (Hoerl and Kennard, 1970). The SVR is a popular choice for its ability to handle high-dimensional data and its robustness against overfitting, while the Random Forest is an ensemble method that constructs multiple decision trees to yield more accurate and stable predictions. They both outperformed other models in preliminary tests, demonstrating their promise as suitable candidates for this task. The Ridge regressor, acting as a meta-classifier in our stacked ensemble, takes the predictions from the base models as input and generates a final prediction, leveraging regularization to minimize multicollinearity issues and prevent overfitting. As we didn't find benefits in using grid search for parameter tuning, possibly due to the high computational cost and time-consuming nature of the method, we report only the results of the experiments that did not include a pre-grid search for parameter optimization, opting for a more efficient approach to model selection and training. All models were trained on

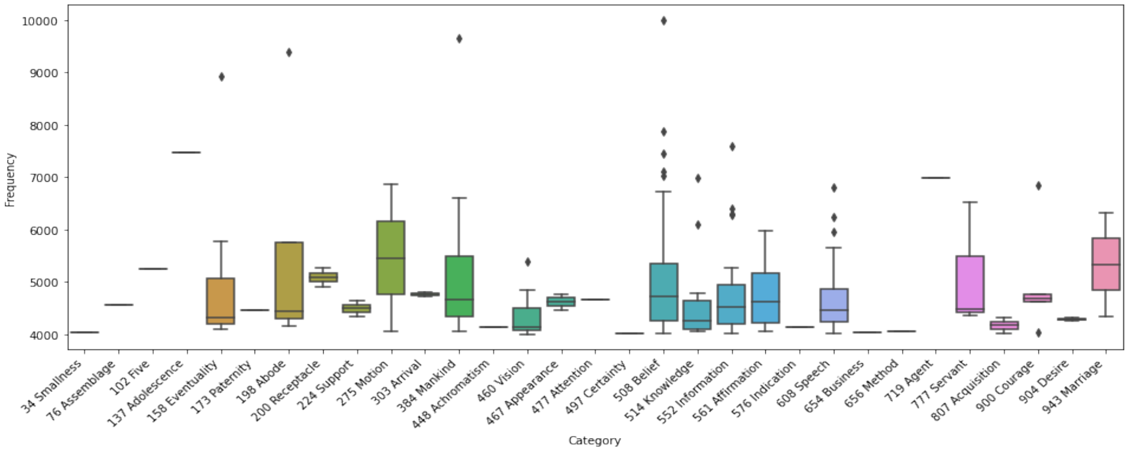

Figure 3: Raw frequencies of the most common categories in the corpus.

	Whole (9089)		score>2.5 (8949)		readers>130 (5827)	
Model	r2	MSE	r2	MSE	r2	MSE
Baseline	-1.1	0.8	-0.0041	0.11	0.0003	0.07
Sentiment Features	0.42	0.14	0.03	0.09	0.07	0.06
Roget Features	0.49	0.13	0.17	0.09	0.23	0.05
Sentiment + Roget Features	0.50	0.13	0.18	0.08	0.24	0.04
Feature selection max=500	0.41	0.14	0.16	0.09	0.22	0.06

Table 3: Model performance comparison with different features and subsets of the dataset. In parenthesis the number of titles in each subset.

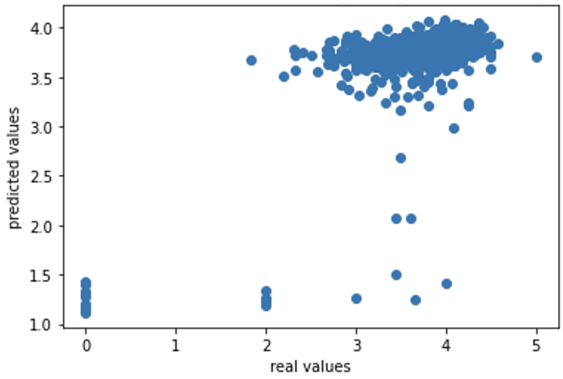

Figure 4: Distribution of real and predicted avg. rating values. Notice how ratings under 2.5 appear particularly predictable, despite their scarcity.

80% and tested on 20% of the corpus.

7 Results

7.1 Baseline

As a baseline, we used only the novels' average sentiment. This baseline relies on the intentionally simplistic idea that overall happier or sadder novels might correspond to reader-appreciation. We include this baseline to provide the reader with a comparison with a "poor model", since understanding the quality of regressor-outputs can be far from intuitive.

7.2 Using Sentiment

Using exclusively sentimental features as a basis for analysis, our model already demonstrates a notable capacity to predict GoodReaders' ratings of various literary works. However, upon closer inspection, it is evident that the high performance across the entire dataset may be somewhat misleading: a small number of exceptionally low-rated titles within the dataset exhibit a marked predictability when sentiment scores are employed as the sole predictive factor. Perhaps surprisingly, these low-rated titles seem to have overwhelmingly predictable sentimental profiles, which in turn make it relatively simple for the models to accurately predict the corresponding ratings. When we control for the low-scoring titles, sentiment analysis still appears to provide some degree of predictive power, although lower than what is achieved when bringing the Roget scores onto the scene.

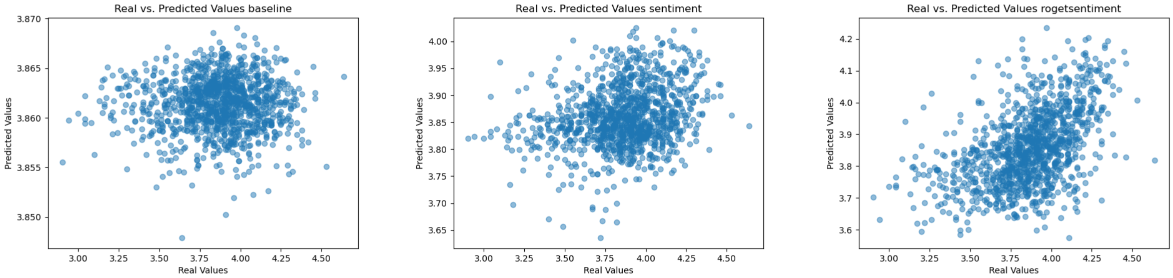

Figure 5: Distribution of real and predicted avg. rating values for all titles with more than 130 different ratings, from left to right: 1) Baseline, modelled on only one feature, the mean sentiment of arcs. 2) Using only the sentiment-arc based features. 3) Using our whole feature set: Roget features and sentiment-arc based features.

7.3 Adding Roget

Adding Roget-category frequencies in our regression model demonstrates significant improvement in predicting novels' ratings. It seems that by using these categories, we can model a broad range of linguistic and thematic elements present within the narratives, which in turn can provide valuable insights into their quality and reception. This enhancement to the model is particularly beneficial as it allows us to move beyond the limitations of relying solely on sentiment analysis. Interestingly, feature selection does not necessarily help the model. It appears that the interplay of "minor" categories maintains an important role in the overall reception of the text, and cutting the max number of features down to 500 decreases the performance of the model. On the other hand, almost halving the number of predictors reduces the r2 of “only" two points, which could be a valid tradeoff in practical applications.

7.4 Rating count thresholds

We experiment with training only the texts that have more than a given rating count (number of raters), using a threshold of 130. This represents the 0.000001 of all readers that rated books in our corpus - leaving us with 5827 titles. We find that in all cases, relying on higher scores systematically helps the models' performance. We find this particularly intriguing, as it shows that as the number of raters of a book increases, the final score may become more reliable, leading to improved predictability. This phenomenon can be likened to a larger sample size in a statistical study, where increasing the number of data points tends to produce more accurate and consistent results. The fact that our models perform better when relying on a higher number of reader scores seems to imply that there is a discernible, shared perception of literary quality among readers. This collective assessment, in turn, hints at the existence of certain objective criteria that contribute to the evaluation of a book's merit.

8 Inspecting the most rated individual titles

To better understand and analyse the strength and weakness of our model, we inspected the works that elicited its most accurate and the least predictions, considering only the "elite" of the most widely read (and often canonical) titles setting, a rating count threshold at 90,000. We provide an example of the very top and bottom of the list in Table 4. On top of the list of the **worst predicted** are both famous and infamous novels: Ayn Rand's *Atlas Shrugged*, William Gibson's *Neuromancer*, James Joyce's *Ulysses*, and Isaac Asimov's *I: Robot*. One possible explanation for this phenomenon is that these are all works that have a devoted following. As the model is solely fed with text-intrinsic features it would not be able to predict a more cult-like admiration of works that may otherwise be considered to be either very complex stylistically, like *Ulysses*, less literary, like *Atlas Shrugged*, or particularly simple in style, like *I: Robot*. Having a reputation that makes these "more than just novels", but cultural beacons of various kinds, may affect users' grading behaviour. Looking at instances with the lowest error in predicting average GoodReads rating, the **best predicted** titles in our model, it is clear that these are popular and accessible works rather than highly canonized works. Genre fiction, such as Sci-fi (Dick, Card, Butler), Fantasy (Gabaldon), and Mystery (Evanovich) dominate the list of best-predicted titles. A bit further down the list, below rank 13th, authors such as Toni Morri-

Best predicted				**Worst predicted**			
Error	**Title**	**Author**	**Rating count**	**Error**	**Title**	**Author**	**Rating count**
0,0013	*A Scanner Darkly*	Philip K. Dick	97963	0,1716	*Stoner*	John Williams	133814
0,0013	*The Big Sleep*	Raymond Chandler	144616	0,1415	*Robin*	Frances H. Burnett	1055312
0,0017	*The Color Purple*	Alice Walker	628511	0,1374	*And Then There Were None*	Agatha Christie	1124501
0,0018	*Xenocide*	Orson Scott Card	150601	0,1318	*Rebecca*	Daphne Du Maurier	557804
0,0019	*High Five*	Janet Evanovich	123615	0,1313	*Blood Meridian*	Cormac McCarthy	129364
0,002	*Kindred*	Octavia E. Butler	153340	0,128	*The Screwtape Letters*	C.S. Lewis	394394
0,0024	*Dragonfly In Amber*	Diana Gabaldon	327501	0,1193	*Atlas Shrugged*	Ayn Rand	375362
0,003	*Hatchet*	Gary Paulsen	356112	0,1102	*Ulysses*	James Joyce	120014

Table 4: Top 8 best and worst predicted titles of the best-performing model (all features), trained with a threshold of 130 readers. Error represents the difference between the real and predicted GoodReads' rating of titles.

son, Ernest Hemingway, John Steinbeck, Truman Capote, Aldous Huxley, and John Irving appear. All are known for solid craftsmanship and accessible stories.

Only conjectures can be made from inspecting these lists, but we do seem to see contours of a skewed grading that is based on more than text-intrinsic features, like a form of readerly devotion that may be playing a role in both the rating count and the average score of some titles.

Another possible interpretation of this distribution, sustained by the large amount of genre fiction among the best-predicted titles, is that the features we selected for our model, and in particular the Roget categories, behave in a characteristic way in works of genre-fiction, while more general works of literature might be distinguished better by considering stylistic features (wholly bypassed in our model). As such, Roget categories may be acting as a proxy for genre, which would be reasonable considering the ideational focus of the Roget thesaurus. The predictability of genre fiction especially may be explained if we assume that genre-fiction tends to place in a narrower grade-interval proper to their genre, while more general or "literary fiction" falls more consistently in a widert interval of ratings (from very low to very high).

An alternative hypothesis, not entirely incompatible with the above and in line with previous work (Jautze et al., 2016) , is that genre-fiction and lower-rated works tend to be more mono-topical, i.e., be less diverse in content, treating a smaller range of topics. As such, Roget categories may also to some extent be measuring topic-diversity, accurately predicting works lower that are more mono-topical. All in all, it is essential to bear in mind that our feature set does not include any stylometric features (such as word choice, sentence structure, and the use of punctuation), leaving it blind to a crucial aspect of literature – or even to "literariness" as such: stylistics contribute significantly to the experience of the uniqueness and richness of a literary work (Miall and Kuiken, 1998), and is a central part of the impact of fiction in non-genre-fiction in particular (Boot and Koolen, 2020). Since our feature-set only observes texts from the sentimental and semantic perspective, it is possible that elements central to the reading experience in some of these titles remain unobserved. Finally, the model's sensitivity to topical interplays might enable it to more accurately identify popular trends and themes and have a skewed performance towards books that follow popular topical patterns rather than those that exhibit exceptional style or depth.

9 Conclusions and future works

The present study has shown that a combination of sentiment arc features, including dynamic measures, and semantic profiling based on Roget categories enhances the predictive power of regression models for perceived literary quality – as measured through average GoodReads' scores – across thousands of novels from the 19th and 20th century. Our findings indicate that by accounting for a diverse set of psycho-semantic features in combination with measures that consider both the dynamics and valences of the novels sentiment arcs, we can obtain a performance that is better than that of any of the latter two approaches in isolation. A surprising finding was that the worst-rated titles seem to exhibit a particular predictability, possessing a more distinguishable profile in comparison to other titles, which might have contributed to an artificial inflation of our model's performance. It suggests that these particular titles may share specific sentiment or topical features that make them stand out from the rest, by which our model can identify them more easily. Our results also highlight that the sheer magnitude of readers' ratings consistently enhances model performance. This observation supports the idea that certain aspects of literary quality tap into aesthetic preferences that are shared among large numbers of readers, at least widely enough to make predictions based on text profiling more reliable with a larger pool of evaluators.

Moreover, the predictive capacity of Roget categories and sentiment arcs for literary quality perception indicate that there exists a underlying structure in how readers perceive and evaluate literary works. Roget categories enable us to capture a coarse representation of the semantic content within texts, offering insights into themes, motifs, and granular references to topics that might resonate with readers. Our related measures of sentiment arcs, in contrast, capture the emotional dynamics of the narratives, allowing us to examine the progression of feelings and the level of consistency and predictability of the story as it unfolds. This aspect is crucial because it highlights the role of sentiments in shaping the reader's engagement and overall impression of a text. By combining these two dimensions — semantic content and sentimental dynamics — we can delve deeper into the complex interplay between emotional patterns and thematic elements which impacts the perception of literary quality. This holistic approach enables us to gain a more nuanced understanding of the factors that contribute to the appreciation of literary works and the ways in which readers discern quality in literature. Additionally, this combined analysis might potentially unveil commonalities and differences among various genres, styles, and time periods, further enriching our understanding of the multifaceted nature of literary quality.

Our approach still has a large number of limitations that need to be acknowledged. First, our approach relies on a reductive representation of the narrative texts, overlooking all traditional stylometric measures. The perception of literary quality is an intricate concept that relies on numerous factors, ranging from the stylistics, characters, plot development and pace, to cultural contexts. By reducing each narrative text to a subset of chosen features, our approach inevitably discards much of the richness and subtlety of works, while the narrow range facilitated by GoodReads' scores forces the models to discern nuanced differences in perceived quality among texts that may be considered generally good by readers. This clearly limits our understanding of literary quality, especially when it comes to the more linguistically or stylistically virtuous titles. Secondly, the reliance on GoodReads scores as the sole metric of quality introduces biases, as these scores are inevitably influenced by factors such as genre preferences and reader demographics. Finally, the analysis is based on a limited sample of English-language texts from the 19th and 20th centuries, potentially limiting the generalizability of our findings to other periods, languages, or contexts. For the same reason, our study cannot consider the potential impact of translation and its effect on the reception of the texts. At the same time, given the inherent complexity of these constraints and the subjective nature of literary evaluation, the performances achieved by our models in terms of r2 scores and mean squared errors, which would be modest for easier tasks, can be considered rather promising.

Naturally, there is much that can be done from here. In the future, we intend to compile an even larger data set, in terms of both texts and features. Integrating stylometric and syntactic features, for instance, could provide additional insights into the complex nature of literary quality. Furthermore, we plan to investigate genre-specific patterns, as observing the performance of our models across different genres may reveal unique patterns and relationships that are specific to particular types of literature. Finally, we intend to use more diverse and sophisticated metrics than GoodReads: exploring alternative sources such as anthologies, awards, and canon lists. Leveraging a richer set of indicators for literary quality/qualities, we hope to gain clearer insights into the complex interplay of factors that contribute to the perception of literary quality.

References

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in text and speech*. University of Illinois at Urbana-Champaign.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Yuri Bizzoni, Ida Marie Lassen, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2022a. Predicting Literary Quality How Perspectivist Should We Be? In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 20–25, Marseille, France. European Language Resources Association.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022b. Fractal sentiments and fairy tales-fractal scaling of narrative arcs as predictor of the perceived quality of andersen's fairy tales. *Journal of Data Mining & Digital Humanities*.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022c. Fractality of sentiment arcs for literary quality assessment: The case of nobel laureates. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 31–41, Taipei, Taiwan. Association for Computational Linguistics.

Yuri Bizzoni, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2021. Sentiment dynamics of success: Fractal scaling of story arcs predicts reader preferences. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 1–6, NIT Silchar, India. NLP Association of India (NLPAI).

Peter Boot and Marijn Koolen. 2020. Captivating, splendid or instructive?: Assessing the impact of reading in online book reviews:. *Scientific Study of Literature*, 10(1):35–63. Publisher: John Benjamins Publishing Company.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. Gutentag: an nlp-driven tool for digital humanities research in the project gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47.

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. 2017. Affective computing and sentiment analysis. In *A practical guide to sentiment analysis*, pages 1–10. Springer.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Irina-Ana Drobot. 2013. Affective narratology. the emotional structure of stories. *Philologica Jassyensia*, 9(2):338.

Katherine Elkins and Jon Chun. 2019. Can Sentiment Analysis Reveal Structure in a Plotless Novel? ArXiv:1910.01441 [cs].

Jianbo Gao and Bo Xu. 2021. Complex Systems, Emergence, and Multiscale Analysis: A Tutorial and Brief Survey. *Applied Sciences*, 11(12):5736.

Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Jing Hu, Jianbo Gao, and Xingsong Wang. 2009. Multifractal analysis of sunspot time series: the effects of the 11-year cycle and Fourier truncation. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(02):P02066.

Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2021. Dynamic evolution of sentiments in never let me go: Insights from multifractal theory and its implications for literary analysis. *Digital Scholarship in the Humanities*, 36(2):322–332.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

SM Mazharul Islam, Xin Luna Dong, and Gerard de Melo. 2020. Domain-specific sentiment lexicons induced from labeled documents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6576–6587.

Swapnil Jain, Shrikant Malviya, Rohit Mishra, and Uma Shanker Tiwary. 2017. Sentiment analysis: An empirical comparative study of various machine learning approaches. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 112–121, Kolkata, India. NLP Association of India.

Syeda Jannatus Saba, Biddut Sarker Bijoy, Henry Gorelick, Sabir Ismail, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. A Study on Using Semantic Word Associations to Predict the Success of a Novel. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 38–51, Online. Association for Computational Linguistics.

Mario Jarmasz. 2012. Roget's thesaurus as a lexical resource for natural language processing.

Kim Jautze, Andreas van Cranenburgh, and Corina Koolen. 2016. Topic modeling literary quality. In *Digital Humanities 2016: Conference Abstracts*, pages 233–237.

Matthew Jockers. 2017. Syuzhet: Extracts sentiment and sentiment-derived plot arcs from text (version 1.0. 1).

George H John, Ron Kohavi, and Karl Pfleger. 1994. Irrelevant features and the subset selection problem. In *Machine learning proceedings 1994*, pages 121–129. Elsevier.

Evgeny Kim and Roman Klinger. 2018. A survey on sentiment and emotion analysis for computational literary studies. *arXiv preprint arXiv:1808.03137*.

Corina Koolen, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020a. Literary quality in the eye of the Dutch reader: The national reader survey. *Poetics*, 79:1–13.

Corina Koolen, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020b. Literary quality in the eye of the dutch reader: The national reader survey. *Poetics*, 79:101439.

Cornelia Wilhelmina Koolen. 2018. *Reading beyond the female: the relationship between perception of author gender and literary quality*. Number DS-2018-03 in ILLC dissertation series. Institute for Logic, Language and Computation, Universiteit van Amsterdam, Amsterdam.

Kayvan Kousha, Mike Thelwall, and Mahshid Abdoli. 2017. Goodreads reviews to assess the wider impacts of books. 68(8):2004–2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23805.

Ida Marie Schytt Lassen, Yuri Bizzoni, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Laigaard Nielbo. 2022. Reviewer Preferences and Gender Disparities in Aesthetic Judgments. In *CEUR Workshop Proceedings*, pages 280–290, Antwerp, Belgium. ArXiv:2206.08697 [cs].

Elizabeth D. Liddy, Caroline A. Hert, and Philip Doty. 1990. Roget's International Thesaurus: Conceptual Issues and Potential Applications. *Advances in Classification Research Online*, pages 95–100.

Severi Luoto and Andreas van Cranenburgh. 2021a. Psycholinguistic dataset on language use in 1145 novels published in English and Dutch. *Data in Brief*, 34:106655.

Severi Luoto and Andreas van Cranenburgh. 2021b. Psycholinguistic dataset on language use in 1145 novels published in english and dutch. *Data in brief*, 34:106655.

Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A. González, and Thamar Solorio. 2017. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1217–1227, Valencia, Spain. Association for Computational Linguistics.

David S. Miall and Don Kuiken. 1998. The form of reading: Empirical studies of literariness. *Poetics*, 25(6):327–341.

Saif Mohammad and Peter Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:1–234.

Mahdi Mohseni, Volker Gast, and Christoph Redies. 2021. Fractality and variability in canonical and non-canonical english fiction and in non-fictional texts. 12.

Mahdi Mohseni, Christoph Redies, and Volker Gast. 2022. Approximate entropy in canonical and non-canonical fiction. *Entropy*, 24(2):278.

Floor Naber and Peter Boot. 2019. Exploring the features of naturalist prose using LIWC in Nederlab. *Journal of Dutch Literature*, 10(1). Number: 1.

Lisa Nakamura. 2013. "Words with friends": Socially networked reading on Goodreads. *PMLA*, 128(1):238–243.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016a. The emotional arcs of stories are dominated by six basic shapes. 5(1):1–12.

Andrew J. Reagan, Brian Tivnan, Jake Ryland Williams, Christopher M. Danforth, and Peter Sheridan Dodds. 2016b. Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. ArXiv:1512.00531 [cs].

Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29. Number: 1 Publisher: SpringerOpen.

Peter Mark Roget. 1997. *Roget's II: the new thesaurus*. Taylor & Francis.

Robert Roget. 1962. Introduction [1852]. In Peter Mark, editor, *The Original Roget's Thesaurus of English Words and Phrases.*, pages 25–43. St. Martin's Press, New York.

Syeda Jannatus Saba, Biddut Sarker Bijoy, Henry Gorelick, Sabir Ismail, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. A study on using semantic word associations to predict the success of a novel. In *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 38–51.

Sean Shesgreen. 2009. Canonizing the canonizer: A short history of the norton anthology of english literature. *Critical Inquiry*, 35(2):293–318.

Annie Swafford. 2015. Problems with the Syuzhet Package.

Andreas van Cranenburgh and Corina Koolen. 2020. Results of a single blind literary taste test with short anonymized novel fragments. *arXiv preprint arXiv:2011.01624*.

Andreas van Cranenburgh, Karina van Dalen-Oskam, and Joris van Zundert. 2019. Vector space explorations of literary language. *Language Resources and Evaluation*, 53(4):625–650.

Melanie Walsh and Maria Antoniak. 2021. The goodreads 'classics': A computational study of readers, amazon, and crowdsourced amateur criticism. *Journal of Cultural Analytics*, 4:243–287.

Xindi Wang, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, and Albert-László Barabási. 2019. Success in books: Predicting book sales before publication. *EPJ Data Science*, 8(1):31.

Matthew Wilkens. 2012. Canons, close reading, and the evolution of method. *Debates in the digital humanities*, pages 249–58.