Dimensions of Quality: Contrasting Stylistic vs. Semantic Features for Modelling Literary Quality in 9,000 Novels

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

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

Abstract

In computational literary studies, the challenging task of predicting quality or reader appreciation of narrative texts is confounded by volatile definitions of quality and the vast feature space that may be considered in modeling. In this paper, we explore two different types of feature sets: stylistic features on one hand, and semantic and sentiment features on the other. We conduct experiments on a corpus of 9,089 English language literary novels published in the 19th and 20th century, using GoodReads' ratings as a proxy for reader appreciation. Examining the potential of both approaches, we find that some types of books are more predictable in one model than in the other, which may indicate that texts have different prominent characteristics (i.a., stylistic complexity, narrative progression at the sentiment-level).

1 Introduction

Defining literary quality or reader appreciation is a complex challenge for quantitative literary studies due to the the heterogeneous nature of narrative texts, and the complexity of mechanisms of judgements and standards in the literary field. While recent studies demonstrate that literary quality appears above chance at the scale of large numbers, and that both text-extrinsic and text-intrinsic features systematically impact sales figures and reader judgements (Wang et al., 2019; Lassen et al., 2022; Koolen et al., 2020; Bizzoni et al., 2022a; Maharjan et al., 2017), the question of how these features interact, and what metrics can be used to validate them, remains open. The challenge lies not merely in modeling literary quality, but in selecting which features to include in a model, while ensuring a degree of interpretability. In this study, we examine two different sets of textual features for modelling literary quality: stylistic and syntactic characteristics vs. narrative and semantic features based on sentiment analysis and word-category profiling.

2 Related works

Generally, we may distinguish two types of feature-sets used to model literary quality: stylistic features (the "how" of writing) and those that capture deeper structures and content (the "what" of writing). Previous studies of literary quality have predominantly relied on stylistic features, such as sentence-length, lexical richness or redundancy (Koolen et al., 2020; Maharjan et al., 2017), syntactic complexity (Zedelius et al., 2019), or n-gram frequencies (Koolen et al., 2020).

More recent works have tested the effect of alternative features, such as sentiment analysis on reader experience (Drobot, 2013; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017; Reagan et al., 2016). Studies relying on sentiment analysis usually draw scores from lexica (Islam et al., 2020) or human annotations (Mohammad and Turney, 2013), to outline the sentiment arcs of narrative texts (Jockers, 2017), and have shown a correlation between reader appreciation and sentiment (Maharjan et al., 2017, 2018). Hu et al. (2021) and Bizzoni et al. (2022b) modelled persistence, coherence, and predictability of sentiment arcs using fractal analysis, a method to study the dynamics of complex systems (Hu et al., 2009; Gao and Xu, 2021), finding correlations with reader appreciation (Bizzoni et al., 2021). In summary, simple or more complex approaches methodologically based on sentiment-annotation show a predictive power for reader appreciation.

Beyond sentiment analysis, other approaches to modelling literary quality have focused on the semantic content of texts. Using topic modeling, Jautze et al. (2016) found that novels with a higher topic diversity elicited higher ratings, and less topically diverse works like genre fiction were perceived as less prestigious, while van Cranenburgh et al. (2019) found that the specific topics in texts

Proceedings of Recent Advances in Natural Language Processing, pages 739–747
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_080

also indicate higher or lower literary quality - topics linked to intimate and familiar relations, for example, seem to indicate lower ratings, which can be linked to the hypothesis that specific genres, especially those in which women authors are dominant, are perceived less literary (Koolen, 2018). While topic modelling or resources like Linguistic Inquiry and Word Count (LIWC)[1] are often used to model semantics (Luoto and van Cranenburgh, 2021; Naber and Boot, 2019), Jannatus Saba et al. (2021) have shown that the Roget thesaurus outperforms them in modeling literary quality.

3 Methods

3.1 Quantifying quality

For practical reasons, computational studies tend to rely on a single proxy of literary quality, even if it may conflate types of literary evaluations (e.g. genre-specific evaluation) reducing them to a mono-dimensional scale. Various proxies have been used, such as readers' ratings on platforms like GoodReads (Kousha et al., 2017), or a text's presence in established literary canons (Wilkens, 2012). Still, different quality-standards may display significant convergences (Walsh and Antoniak, 2021). For the present study, we employed the average ratings and rating count (number of user-ratings) of books on **GoodReads**, a popular online literary platform.[2] While GoodReads as a proxy for reader appreciation does have the obvious limitations mentioned, it is a practical starting point for quantifying literary quality across a wide range of readers, genres, and authors. With more than 90 million users, GoodReads may be particularly valuable for giving an insight into reading culture "in the wild" (Nakamura, 2013), deriving both its listed books and ratings from a heterogeneous pool of readers in terms of background, nationality, gender, age, and reading preferences (Kousha et al., 2017). Note that while GoodReads average rating ranges from 0 to 5, it does display a positivity bias, with titles having a high mean rating overall (Fig. 1).

3.2 Data

We used the Chicago Corpus dataset of more than 9,000 English-language published in English between 1880 and 2000.[3] Novels were selected for this corpus based on the number of copies extant in libraries worldwide, resulting in a diverse collection of genres, from popular fiction genres to Nobel Prize laureates works (Bizzoni et al., 2022c), with a large subsection of texts featured in canonical collections such as the Penguin Classics book-series,[4] the GoodReads' Classics list,[5] the Norton Anthology (Shesgreen, 2009).[6] It should be noted that the corpus has a cultural and geographical tilt toward Anglophone authors.

	Titles	**Authors**
Number	9089 (727)	3150 (173)
Avg. rating	3.74	3.69
Avg. rating count	14246.36	12816.83

Table 1: Above: number of titles and authors in the corpus and in the canonical subset of the corpus (in parenthesis). Below: the average GoodReads' rating and average number of ratings per book and author.

3.3 Features

The task of predicting literary quality is inherently complex due to the large set of features that could be considered, but also because these seem to pertain to different levels of narrative texts. As noted previously, stylistic features are frequently used in this line of studies, while those pertaining to the sentiment and semantic profiles of narratives have been less explored. While recent studies have sought to assess the effect of adding sentiment features to a model based on stylistic features (Bizzoni et al., 2023b), and of adding semantic profiles (Roget categories) to a model based on sentiment features (Bizzoni et al., 2023a), it is still difficult to assess these two different levels of narrative against each other: the purely textual and stylistic features against those pertaining to more underlying narrative content and dynamics. To compare these two different types of features sets both in terms of effect and what aspects of texts they seem to capture, we train two models on each set, basing our selection of features on what has previously been used in studies on predicting literary quality. We call these two models the stylistic and the narrative model.

For *the stylistic model*, we chose stylistics features that have been applied in previous studies (Koolen et al., 2020; Maharjan et al., 2017; van Cranenburgh and Bod, 2017; van Cranenburgh et al.,

[1] https://www.liwc.app
[2] https://www.goodreads.com
[3] https://textual-optics-lab.uchicago.edu
[4] https://www.penguin.com/penguin-classics-overview/
[5] https://www.goodreads.com/shelf/show/classics
[6] https://www.norton.com/books/

	Whole (9089)		rated>130 (5827)	
Model	r2	MSE	r2	MSE
Baseline	-0.69	0.37	-0.47	0.09
Stylistic and syntactic features	0.37	0.14	0.16	0.05
Sentiment and semantic features	0.48	0.13	0.21	0.05

Table 2: Model performance comparison against a baseline (trained only on mean sentiment), showing the performance of the models when trained on the whole corpus and on the corpus subset (rated>130 times). In parenthesis the number of titles in each subset.

2019; Crosbie et al., 2013; Ganjigunte Ashok et al., 2013; Algee-Hewitt et al., 2016; Zedelius et al., 2019). These are **sentence length**; **lexical diversity** (Torruella and Capsada, 2013); ratio of text-**compressibility**, indicating redundancy or formulaicity (Benedetto et al., 2002); **entropy** of words and bi-grams, the unpredictability or information present in a collection of words or pairs of consecutive words (Shannon, 1948); five classic indices of **readability**,and several **syntactic features**: frequencies of parts of speech and selected syntagms such as subjects, passive auxiliaries and relative clauses (see the full list of features in appendix).

For *the narrative model*, we similarly selected measures from previous studies (Maharjan et al., 2017; Mohseni et al., 2022, 2021; Bizzoni et al., 2022a; Jannatus Saba et al., 2021). With a simple approach to sentiment analysis, we extracted compound sentiment scores of all sentences in novels (tokenizing with NLTK[7]) with the VADER lexicon (Hutto and Gilbert, 2014). From these values, we also computed and detrended sentiment arcs of the novels [8]. Thus, we based our model on **mean sentiment** valence and **standard deviation**, as well as two measures of arc dynamics based on the detrended arcs: **Hurst** exponent, and **Approximate Entropy**, which is a measure of the complexity or irregularity of a time series (Delgado-Bonal and Marshak, 2019). Beyond sentiment-features, we calculated the frequency of 1044 **Roget "paragraphs"** (i.e., topics in each of subcategory) of *Roget's Thesaurus of English Words* (Roget, 1997; Liddy et al., 1990) indicating the topical interplay of semantically based word-categories in our novels (see example in appendix, fig.2).

3.4 Model

For our prediction task we employed a Random Forest regressor, a robust and well-regarded machine learning technique (Breiman, 2001) that combines multiple decision trees to deliver more accurate and stable predictions. As a non-parametric method, it is well-suited to complex tasks where the relationship between predictors and outcome is not easily approximated by a simple function. The Random Forest algorithm offers two key advantages for our study: first, the method is capable of handling high-dimensional data; second, by aggregating the results of many decision trees, each trained on a slightly different set of data, this approach mitigates the risk of overfitting, making it apt for relatively small, highly complex datasets like the one we are using. Regarding our model training and testing protocol, we opted for a standard split of our dataset - we partitioned the corpus into two subsets: 80 % of the data was used for training our models, while the remaining 20 % was reserved for testing. We chose not to stratify authors, i.e., we did not make sure that titles of the same author appeared in the training and test set, as we seek to assess the reader appreciation of individual titles and since the perceived quality and GoodRead's average rating may vary a lot between titles of the same author.

4 Results

4.1 Baseline

As it can be difficult to assess model performance, we included a baseline model for comparison, which is only trained on a single feature (mean sentiment of a novel), and naturally exhibits poor performance (Table 2). This baseline is naturally undemanding and more complex models could have been used to assess model performance. However, our interest is not in assessing the performance of our two models against the state of the art, but rather to examine the difference between them to gain a better insight into the behaviour of the two types of feature sets. The baseline is, as such, only included as a reference to evaluate the effect when comparing the two models.

[7] https://www.nltk.org/

[8] See Hu et al. (2021) for details on this method

Best predicted				Worst predicted			
Error	**Title**	**Author**	**Rating count**	**Error**	**Title**	**Author**	**Rating count**
0.0013	*Children Of Dune*	Frank Herbert	149561	1.4385	*The Color Purple*	Alice Walker	628511
0.0031	*The Heart Is A Lonely Hunter*	Carson McCullers	102550	0.3280	*The Screwtape Letters*	C.S. Lewis	394394
0.0037	*The Black Echo*	Michael Connelly	179372	0.3176	*Animal Farm*	George Orwell	3967590
0.0043	*To The Lighthouse*	Virginia Woolf	159757	0.2832	*Anne Of Windy Poplars*	L.M. Montgomery	103599
0.0054	*The Fountainhead*	Ayn Rand	312146	0.2819	*Giovanni's Room*	James Baldwin	102685
0.0067	*Dolores Claiborne*	Stephen King	140124	0.2771	*The Green Mile*	Stephen King	286816
0.0079	*A Portrait Of The Artist*	James Joyce	141170	0.2414	*The Wayward Bus*	John Steinbeck	486536
0.0079	*The Maltese Falcon*	Dashiell Hammett	99733	0.2397	*Fight Club*	Chuck Palahniuk	547786
0.0102	*Catch-22*	Joseph Heller	788426	0.2375	*The Velveteen Rabbit*	Margery W. Bianco	246379
0.0112	*The Virgin Suicides*	Jeffrey Eugenides	273576	0.2248	*The Red Tent*	Anita Diamant	565946

Table 3: Top 10 best and worst predicted titles, using **stylistic features** only, and trained on all titles, but showing only titles rated >90,000 times. Titles in red are the same worst predicted titles in both of our models, stylistic and narrative (cf. Table 4).

Best predicted				Worst predicted			
Error	**Title**	**Author**	**Rating count**	**Error**	**Title**	**Author**	**Rating count**
0.0005	*Hatchet*	Gary Paulsen	356112	1.0477	*The Color Purple*	Alice Walker	628511
0.0007	*House Of Sand And Fog*	Andre Dubus III	129687	0.3056	*The Screwtape Letters*	C. S. Lewis	394394
0.0008	*Midnight'S Children*	Salman Rushdie	114828	0.2761	*Giovanni's Room*	James Baldwin	102685
0.0015	*The Sound And The Fury*	William Faulkner	171316	0.2580	*Fight Club*	Chuck Palahniuk	547786
0.0023	*The Grapes Of Wrath*	John Steinbeck	840278	0.2502	*The Wayward Bus*	John Steinbeck	486536
0.0029	*American Psycho*	Bret Easton Ellis	274920	0.2466	*2001: A Space Odyssey*	Arthur C. Clarke	290785
0.0040	*Lord Of Chaos*	Robert Jordan	155112	0.2404	*The Green Mile*	Stephen King	286816
0.0042	*The Fires Of Heaven*	Robert Jordan	167184	0.2353	*The Dispossessed*	Ursula K. Le Guin	107350
0.0051	*The Pilot's Wife*	Anita Shreve	94753	0.233	*Animal Farm*	George Orwell	3967590
0.0054	*Firestarter*	Stephen King	211794	0.232	*Murder on the Orient Express*	Agatha Christie	517455

Table 4: Top 10 best and worst predicted titles, using **narrative features** only, and trained on all titles, but showing only titles rated >90,000 times. Titles in red are the same worst predicted titles in both of our models, stylistic and narrative (cf. Table 3).

4.2 Stylistic vs narrative model

As we show in Table 2, we observe a differential performance between the stylistic and narrative models. Although the stylistic model does exceed the pre-established baseline, it is surpassed in performance by the narrative model. In both cases, the performances of the models are quite robust given the intricacy of the task, but as shown by the relatively high Mean Square Error (MSE), it might be that some subgroups of titles are particularly well predicted, inflating the models' overall score.

4.3 Rating count threshold

We also applied a threshold for the number of times a book is rated, as the average rating titles with very low numbers of ratings are sensitive to arbitrariness of opinion of very few and do not reflect a consensus among readers. We set an arbitrary threshold at 130 ratings (0.000001 of all ratings in our corpus), and filtering out books with>130 ratings, 5827 titles remained. When training our models on these titles, their performance is significantly lower, yet the MSE is also evidently reduced. Despite this lowered performance, it is worth noting that both models still perform significantly above chance level. This suggests that, while the rating count threshold has an impact, the models retain some predictive ability in both settings, as is also evident when we visualize the real and predicted values of each model (Fig. 3).

4.4 Individual titles

To examine the differences between the two models, we inspected their performance on individual titles. We show only the most highly rated books in the corpus (rated >90,000 times) for the purpose of displaying highly recognizable works (Table 3, 4). Since we were not interested in the models' predictive abilities *per se*, but to examine whether some groups of literary works were apter to be modelled through the semantic and sentimental rather than the stylistic feature set when optimising for reader appreciation, for this test we trained and tested both models on the whole corpus. As such, the errors reported in the Tables 3 and 4 are to be taken as merely comparative measures. A literary scholar manually inspected the 100 best and worst predicted individual titles (lowest and highest error, or the difference between actual and predicted value), finding that while the models might indeed be better capturing different aspects of text in terms of genre and type in their best predictions, they seem to often struggle with the same group of titles (Tab. 3, 4). The **worst predicted** titles in both models distinguish themselves by having some *extra-textual* strong point, such as the author

having a large fan-base (Lewis, Orwell), being important works with regard to contemporary issues, like sexuality and racism (*The Color Purple, Giovanni's Room*), or being popular movie adaptions (*Fight Club, The Green Mile*), which – we conjecture – are factors that influence the ratings of these titles beyond what can be substantiated from textual features alone. This observation is not trivial, since it would have been entirely possible that these works have gained their fame, e.g., were adapted into movies, *because* of their textual characteristics. However, it is still possible that these novels have characteristics that are not adequately captured by any of the features included in our models.

Looking at **best predicted** titles, we find that contemporary canonical fiction of the broad "literary novel" genre (such as novels by Hemingway, Fitzgerald, Joyce and Woolf) appear among the top predictions of the stylistics model more often than among those of the narrative model. To further estimate the performance of the models on canonical vs. non-canonical fiction in our corpus, we aggregated titles found in various standards of literary canonicity, marking all titles extant in our corpus by authors mentioned in a series of lists indicating canonicity.[9] Here, we find that both models are slightly better at predicting canonical than non-canonical works, although for the narrative model, the difference is almost insignificant (p-value 0.049). Finally if we compare their errors when trained on titles>130 rating count, the narrative model does not show any difference in predicting canon vs. noncanon works, while the stylistics model is better at predicting canonical works in this setting (Table 5).

Especially considering that canonical works tend to belong to the more vague genre of "literary fiction", where more acclaimed works tend to be acclaimed for their style while dealing with a broad array of topics, it is possible that the stylistic model is simply better at predicting novels that stand out in terms of style. Consider the stylistic experimentation of works like *A Portrait of the Artist as a Young Man* and *To the Lighthouse*, which appear at the top of best predicted titles in the stylistic model (3). On the other hand, it is possible that the narrative model picks up on characteristics of novels' semantic and sentiment profile that may be more prevalent in genre-fiction, and of which fewer novels become canonical than of the "literary fiction" category. As such, it may be that these two sets of features, the stylistic and the narrative, underlie different types of reader judgements, and capture characteristics of quality in more high-brow vs. more low-brow fiction, which are not necessarily evaluated in the same way, and which, in turn, the GoodRead's average rating conflates.

[9]The Norton Anthology, the Penguin Classics series, GoodReads' Classics list, and the top 1000 most frequent titles in the English literature syllabi collected by the OpenSyllabus project.

Training on the whole corpus		
	Stylistic	**Semantic**
Canon error	0.086	0.084
Non-canon error	0.096	0.091
T-statistic	-2.198	-1.967
P-value	0.028	0.049
Training with a threshold of 130 Ratings		
	Stylistic	**Semantic**
Canon error	0.292	0.082
Non-canon error	0.351	0.085
T-statistic	-3.041	-1.020
P-value	0.002	0.308

Table 5: Difference between the mean error of canonical and non-canonical titles in the whole corpus estimated via t-tests. Note that the p-value for the narrative model tends to be insignificant.

5 Conclusions and future works

We find that novels' stylistic and syntactic features, as well as the characteristics of their overall emotional tone, the dynamics of their sentiment arcs, and the semantic categories they cover appear to be indicative of their appeal to readers and their perceived overall quality. Moreover, while a model based on the selected sentiment and semantic features clearly outperforms a model based on selected stylistic and syntactic features, each model might be best at modelling different types of literary texts, where the stylistic model is better at predicting canonical from non-canonical titles. Interestingly, the models converge on struggling to predict some titles that are perhaps popular because of extra-textual factors. Naturally the subject of predicting reader appreiciation of literar texts is complex. In the future we aim to repeat the experiment looking at various quality proxies beyond GoodReads ratings to study convergences between different perceptions of quality, as well as using a larger set of features. We may also attempt more sophisticated models, as long as some interpretability remains, as the main objective is not to effectively predict a score, but to understand more about how literary texts affect readers at various narrative levels.

References

Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Franco Moretti, Ryan Heuser, and Hannah Walser. 2016. *Canon/Archive. Large-scale Dynamics in the Literary Field*. Pamphlets of the Stanford Literary Lab.

Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. 2002. Language Trees and Zipping. *Physical Review Letters*, 88(4):1–5.

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, Pascale Moreira, Kristoffer Nielbo, Ida Marie Lassen, and Mads Thomsen. 2023a. Modeling readers' appreciation of literary narratives through sentiment arcs and semantic profiles. In *Proceedings of the The 5th Workshop on Narrative Understanding*, pages 25–35, Toronto, Canada. Association for Computational Linguistics.

Yuri Bizzoni, Pascale Moreira, Kristoffer Nielbo, and Mads Thomsen. 2023b. Sentimental matters: Predicting literary quality with sentiment analysis and stylistic features. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 11—-18, Toronto, Canada. Association for Computational Linguistics.

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*, pages 1–15.

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).

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.

Andreas van Cranenburgh and Rens Bod. 2017. A data-oriented model of literary language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1228–1238, Valencia, Spain. Association for Computational Linguistics.

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.

Tess Crosbie, Tim French, and Marc Conrad. 2013. Towards a model for replicating aesthetic literary appreciation. In *Proceedings of the Fifth Workshop on Semantic Web Information Management*, SWIM '13, pages 1–4, New York, NY, USA. Association for Computing Machinery.

Alfonso Delgado-Bonal and Alexander Marshak. 2019. Approximate Entropy and Sample Entropy: A Comprehensive Tutorial. *Entropy*, 21(6):541.

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

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1764, Seattle, Washington, USA. Association for Computational Linguistics.

Jianbo Gao and Bo Xu. 2021. Complex systems, emergence, and Multiscale Analysis: A tutorial and brief survey. *Applied Sciences*, 11(12):5736.

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.

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.

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. Package 'syuzhet'. *URL: https://cran. r-project. org/web/packages/syuzhet*.

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. 2020. Literary quality in the eye of the Dutch reader: The national reader survey. *Poetics*, 79:1–13.

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. *Journal of the Association for Information Science and Technology*, 68(8):2004–2016.

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. 2021. 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.

Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio. 2018. Letting emotions flow: Success prediction by modeling the flow of emotions in books. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Volume 2, Short Papers*, pages 259–265, New Orleans, Louisiana. Association for Computational Linguistics.

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. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJData Science*, 5(1):1–12.

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

C. E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Sean Shesgreen. 2009. Canonizing the canonizer: A short history of The Norton Anthology of English Literature. *Critical Inquiry*, 35(2):293–318.

Joan Torruella and Ramon Capsada. 2013. Lexical statistics and tipological structures: A measure of lexical richness. *Procedia - Social and Behavioral Sciences*, 95:447–454.

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.

Claire M. Zedelius, Caitlin Mills, and Jonathan W. Schooler. 2019. Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods*, 51(2):879–894.

A Appendix

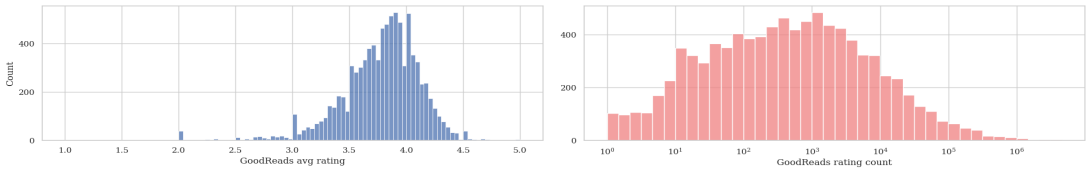


Figure 1: Histograms showing the distribution of average rating and rating count scores in our corpus (note that the latter histogram is logarithmically scaled).

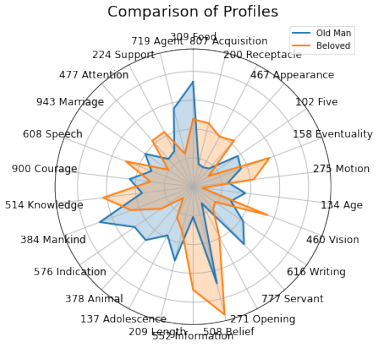
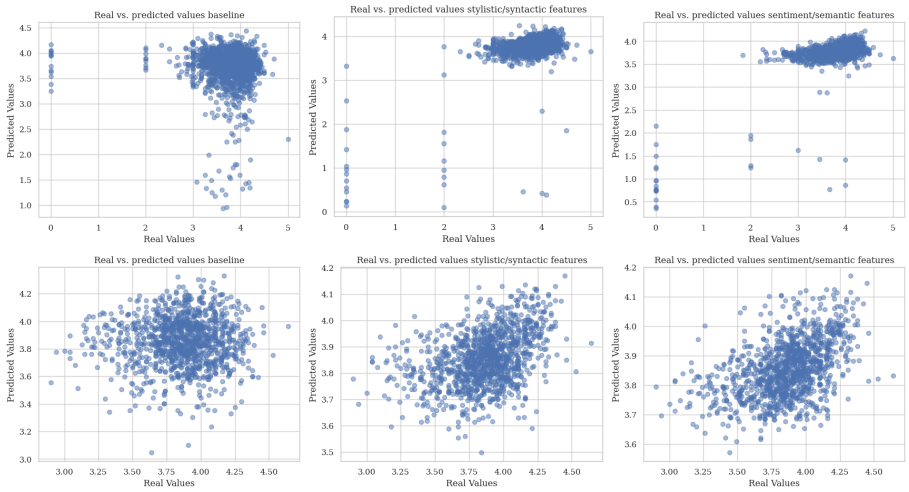


Figure 2: Roget profiles of Hemingway's *The Old Man and the Sea* and Morrison's *Beloved* along their most frequent "paragraphs".



1) Baseline 2) Stylistic model 3) Narrative model

Figure 3: Distribution of real and predicted avg. rating values, for models trained on the full corpus (above) and on titles rated >130 times (below).

Type	Feature	Count
Stylistic features		
Readability indices	Flesch Reading Ease Flesch-Kincaid Grade Level SMOG Readability Formula Automated Readability Index New Dale–Chall Readability Formula	5
Stylistic measures	Lexical diversity (MSTTR) Text compressibility (bzip compression) Word and bi-gram entropy Sentence length	4
Syntactic frequencies	Verb frequency Noun frequency Adjective frequency Adverb frequency Pronoun frequency Punctuation frequency Stopword frequency Nominal subject frequency Auxiliary frequency Passive auxiliary frequency Relative clause modifier frequency Negation modifier frequency	12
Narrative features		
Simple sentiment features	Mean sentiment Std. deviation of sentiment Sentiment of beginning (10%) Sentiment of ending (10%) Difference in mean sentiment (main/ending)	5
Complex sentiment measures	Hurst exponent Approximate entropy	2
Semantic features	Frequencies of Roget subcategories	1044

Table 6: Full feature-sets