Sentimental Matters
Predicting Literary Quality with Sentiment Analysis and Stylistic Features

Yuri Bizzoni
Aarhus University, Denmark
yuri.bizzoni@cc.au.dk

Pascale Feldkamp Moreira
Aarhus University, Denmark

Mads Rosendahl Thomsen
Aarhus University, Denmark
madsrt@cc.au.dk

Kristoffer L. Nielbo
Aarhus University, Denmark
kln@cas.au.dk

Abstract

The task of predicting reader appreciation or literary quality has been the object of several studies. It remains, however, a challenging problem in quantitative literary analyses and computational linguistics alike, as its definition can vary a lot depending on the genre of literary texts considered, the features adopted, and the annotation system employed. This paper attempts to evaluate the impact on reader appreciation, defined as online users' ratings, of sentiment range and sentiment arc patterns versus traditional stylometric features. We run our experiments on a corpus of English-language literary fiction, showing that stylometric features alone are helpful in modelling literary quality, but can be outperformed by analysing the novels' sentimental profile.

1 Introduction

The question of what literary quality "is" is as complex as it is old. It may be argued that "literary quality" is an empty concept, since individual tastes of narrative and literature can differ widely among readers. Yet it is possible that a set of textual and narrative characteristics tend to improve or damage the appreciation of a literary piece independently from genre expectations and preferences. This persistent intuition, while controversial, has been amply discussed through the history of literary criticism, and also stands at the foundation of most rhetorical or writing advice. The idea of an intersubjective agreement on literary quality may be also sustained by the convergence of large numbers of readers (and when considering canons, generations of readers) on certain titles rather than others (Koolen et al., 2020a; Walsh and Antoniak, 2021b). In the quest of defining principles of literary quality, quantitative analyses ask two questions: whether it is possible to define literary quality at all; and whether it is possible to individuate textual patterns that contribute to make a text more appreciated. In this paper we aim to explore the interplay of the sentiment and stylometric characteristics of narrative texts and their role in the perception of literary quality.

2 Related works

Traditionally, quantitative studies of literary quality have relied on texts' stylometric properties, ranging from the percentage of adverbs (Koolen et al., 2020b) to the count of the most frequent n-grams in a text (van Cranenburgh and Koolen, 2020), to model the success or quality of literary works. More recent works, nonetheless, have emphasized the potential of sentiment analysis (Alm, 2008; Jain et al., 2017), at the word (Mohammad, 2018), sentence (Mäntylä et al., 2018) or paragraph (Li et al., 2019) level, to uncover meaningful mechanisms in the reading experience (Drobot, 2013; Cambria et al., 2017; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017), usually by drawing scores from human annotations (Mohammad and Turney, 2013) or induced lexica (Islam et al., 2020).While most studies have focused on the valence of sentiment arcs, Hu et al. (2021) and Bizzoni et al. (2022a) have tried to model the persistence, coherence, and predictability of novels' sentiment arcs, using fractal analysis (Mandelbrot and Ness, 1968; Mandelbrot, 1982, 1997; Beran, 1994; Eke et al., 2002; Kuznetsov et al., 2013), a method of studying patterns in complex systems, exploring the degree of predictability or self-similarity of narratives – a method that appears to capture meaningful patterns impacting reading experience. Naturally, beyond which features to consider, another great challenge of studying literary quality is that of finding "oracles" of quality. Measures of quality have been approximated by looking at readers' ratings on plat-

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 11–18

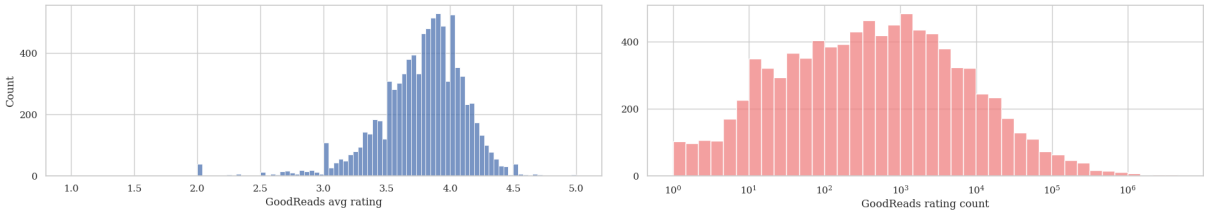

Figure 1: Distribution of GoodReads' ratings and number of ratings in our corpus. Note that the latter is logarithmically scaled.

forms such as GoodReads (Kousha et al., 2017), or by relying on established literary canons (Wilkens, 2012). Despite their diversity, different concepts of quality display large overlaps (Walsh and Antoniak, 2021a), thus to a degree allowing for the comparison across canons and preferences(Underwood, 2019; Wilkens, 2012).

3 Methods

3.1 Corpus

We use the Chicago corpus: over 9,000 English-language novels written in, or translated into English from 1880 to 2000, compiled based on the number of libraries that hold a copy, with a preference for more widely held titles. As such, the corpus is diverse, ranging from well-known genres of popular fiction to important works of "highbrow" literature, including novels from Nobel Prize winners (Bizzoni et al., 2022b) and other prestigious awards, as well as texts included in canonical collections like the Norton Anthology (Shesgreen, 2009). Yet, the corpus has an obvious cultural and geographic bias, with a strong over-representation of Anglophone authors. For this study, we used the whole corpus, as well as a subset of the corpus where 140 titles were filtered out because of their very low rating on GoodReads. We refer to this as the filtered corpus.

	Titles	Authors
Number	9089	3150
Number below 2.5 rating	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

3.2 Quality Measures

As a source of quality judgments we decided to opt for **GoodReads**' average ratings. [1] This metric has limitations: i.a., reducing very different reader preferences and backgrounds to one single score (ratings or "stars"), conflating underlying motivations and important differences among readers. Still, the resource has a uniquely large number of users, facilitating an unprecedented amount of data for quantitative literary analysis, where popular titles are graded by hundreds of thousands of users (Kousha et al., 2017). The advantage of GoodReads is its wide audience, not only in terms of numbers, but because it reaches across genres and curricula (Walsh and Antoniak, 2021a), deriving its scores from a particularly diverse pool of readers, as the platform is accessed from several countries, by users of different genders, ages, etc.

3.3 Stylometric Features

Considering traditional stylometric features, we examine texts' adjusted **lexical diversity** as a measure of proven stylistic importance with obvious cognitive effects on the readers (Torruella and Capsada, 2013); the texts' ratio of **compressibility**, a measure of redundancy and formulaicity (Benedetto et al., 2002; van Cranenburgh and Bod, 2017); five different measures of textual **readability**[2], (based on, i.a., sentence length, word length, and number of syllables),and several grammatical and **syntactic** features, such as the frequency of parts of speech and of a selection of syntagms such as subjects, passive auxiliaries and relative clauses (see Appendix).

3.4 Sentiment Analysis

To build the sentiment arcs of each novel we opted for a simple and "classic" sentiment analysis algorithm: the **VADER** model (Hutto and Gilbert,

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

[2] The Flesch Reading Ease, the Flesch-Kincaid Grade Level, the SMOG Readability Formula, the Automated Readability Index, and the New Dale–Chall Readability Formula.

2014), applied at the sentence level. We chose this method because it is transparent, being based on a lexicon and a set of rules. It is widely employed and shows good performance and consistency across various domains (Ribeiro et al., 2016; Reagan et al., 2016), which is an ideal feature when dealing with narrative, as it enables the comparison across genres, while its origins in social media analysis do not appear to hinder the annotation of literary texts (Bizzoni et al., 2022b). Moreover, plotted arcs appear comparable to the **Syuzet-package** (Elkins and Chun, 2019), one specifically developed for narrative texts (Jockers, 2017), while side-stepping some of the problems of this package (Swafford, 2015), such as those inherent to word-based annotation. To assure the validity of the method, we manually inspected a selection of novels at global and local level (fig. 2, 3). As fig. 2 and 3 show, the high and dips appear to adequately correspond to narrative events, and performance is also good on the sentence-level when looking at the VADER annotation of, for example, the first lines and the corresponding text.[3]

[3]Corresponding text: "He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him. But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally salão, which is the worst form of unlucky, and the boy had gone at their orders in another boat which caught three good fish the first week. It made the boy sad to see the old man come in each day with his skiff empty and he always went down to help him carry either the coiled lines or the gaff and harpoon and the sail that was furled around the mast. The sail was patched with flour sacks and, furled, it looked like the flag of permanent defeat. The old man was thin and gaunt with deep wrinkles in the back of his neck. The brown blotches of the benevolent skin cancer the sun brings from its reflection on the tropic sea were on his cheeks. The blotches ran well down the sides of his face and his hands had the deep-creased scars from handling heavy fish on the cords. But none of these scars were fresh. They were as old as erosions in a fishless desert. Everything about him was old except his eyes and they were the same color as the sea and were cheerful and undefeated."

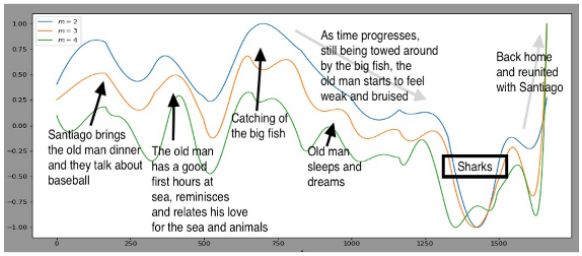

Figure 2: Sentiment arc of Hemingway's *The Old Man and the Sea* with different polynomial fits (m = polynomial degree). Y-axis values represent compound sentiment score (VADER). Values on the x-axis represent the narrative progression by number of sentences.

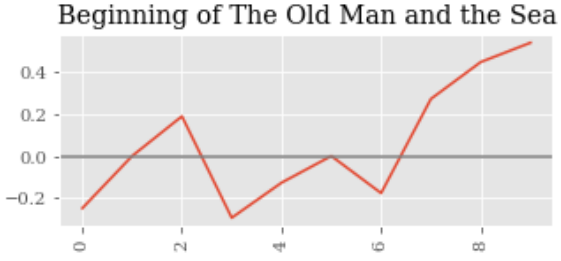

Figure 3: First sentences of Hemingway's *The Old Man and the Sea*, annotated with VADER.

From the annotated arcs, we extracted simple sentiment-arc features: **mean sentiment**, its **standard deviation**, the **mean sentiment of the ending 10 percent of each arc**, the **mean sentiment of the beginning 10 percent of each arc**, as well as the **difference between the main part of the arc and the ending (10 percent)**. Moreover, we computed two more complex measures of arc coherence: their **Hurst** exponent, based on the detrended version of arcs, which is a measure of the long-term memory or persistence of a time series, and their **Approximate Entropy**, which is a measure of the complexity or irregularity of a time series, quantifying the likelihood that patterns will repeat at a later time. These measures of arcs' dynamics have recently proved promising for literary quality modelling (Hu et al., 2021; Bizzoni et al., 2022b).

3.5 Models

As we are particularly interested in the combinations of features that can more accurately predict ratings, we prefer relatively simple and interpretable regression models, using a small set of "classic" algorithms such as **Linear Regression**, **Lasso** and **Bayesian Ridge** (see the complete list in Appendix). Our interest in identifying combinations of features that can accurately predict ratings goes beyond simply achieving high prediction accuracy; we also prioritize interpretability of our model, making explicit the relationships between predictors and outcomes. Simple and interpretable regression models, such as Linear Regression, Lasso, and Bayesian Ridge, provide a number of benefits in this context. First of all, these models allow for direct and straightforward interpretations of feature influences. For example, the coefficients in linear regression quantify the change in response variable for a unit change in the predictors. This is especially useful in our case as we aim to under-

	baseline	Linear	Ridge	Lasso	ElasticNet	BayesRidge	Huber	Polynomial	TheilSen
r2	-1.07	0.23 (0.21)	0.23 (0.21)	0.04 (0.03)	0.05 (0.04)	**0.24** (0.21)	0.13 (0.11)	-0.02 (0.16)	0.22 (0.23)
neg_rmse	0.72	-0.14 (-0.15)	-0.14 (-0.15)	-0.22 (-0.22)	-0.20 (-0.20)	**-0.14** (-0.15)	-0.16 (-0.16)	-0.28 (-0.15)	-0.15 (-0.15)
r2 (filtered)	-0.944	0.061 (0.04)	0.063 (0.05)	0.04 (-0.02)	0.04 (-0.02)	**0.07** (0.05)	-0.46 (0.05)	-0.40 (-0.01)	-0.18 (-0.01)
neg_rmse (filtered)	0.445	-0.10 (-0.1)	-0.09 (-0.1)	-0.10 (-0.11)	-0.10 (-0.11)	**-0.09** (-0.11)	-0.15 (-0.1)	-0.15 (-0.11)	-0.12 (-0.11)

Table 2: Performance (r2 and negative MSE) comparison of regression models using 5-fold cross-validation for the whole (upper) and filtered (lower) corpus, with and without sentiment features (in parenthesis). Lasso and ElasticNet underperform on the larger data-set due to coefficient shrinkage, while Polynomial Regression likely overfits. The best-performing model is Bayesian Ridge. A random baseline is included for comparison.

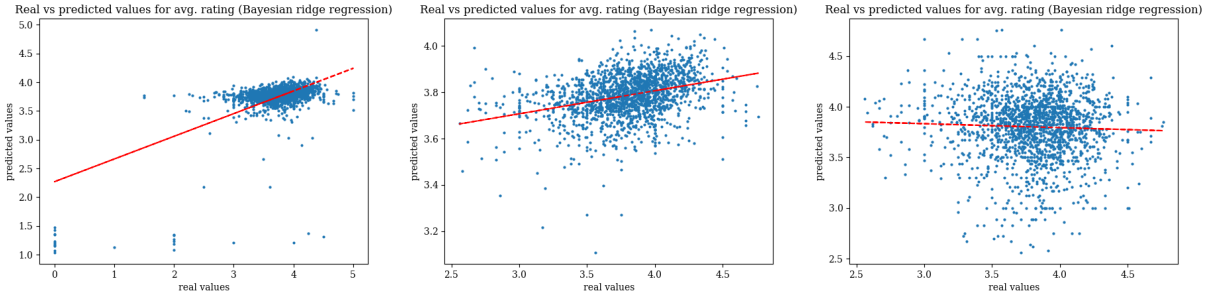

Figure 4: Distribution of real and predicted avg. rating values using Bayesian ridge regression. From left to right: 1) Whole corpus. Notice how ratings under 2.5 appear particularly predictable, despite their scarcity. 2) Filtered corpus. Even in the narrower interval ratings are relatively predictable. 3) Distribution of real and predicted avg. rating values in a random baseline for comparison.

stand not just how well we can predict the ratings, but how each individual feature influences these predictions. Secondly, these models are less prone to overfitting compared to deeper machine learning approaches. While deeper models can potentially yield higher predictive performance, they can also lead to models that are too complex, fitting the noise in our data rather than the underlying relationships. This would reduce the generalizability of our findings and potentially make them less reliable. Finally, using simpler models decreases the computational cost, which can be significant for more complex machine learning algorithms. This efficiency allows for more extensive model tuning and repeated testing, increasing the robustness of our results.

4 Results

Most models tested show predictive power, i.e., perform better than random. Their performance is reported in Table 2. This is our first important result since it would have been entirely possible that none of the chosen features had anything to do with large-scale reader appreciation. The behaviour of our models shows that combinations of some of the selected textual and narrative features can predict novels' average ratings on GoodReads. A second important finding is that sentiment measures improve the performance of almost all models: while a combination of syntactic, readability, and redundancy measures is already enough to partly model ratings, the novels' average sentiment, variation in sentiment intensity, and the overall predictability and persistence of the sentiment arcs increase our ability to predict perceived quality. When looking at the distribution of most models' predictions, we find an evident split: not only does the vast majority of GoodReads' ratings (in our corpus) fall between 3 and 5, with few low scores, but the distinction between very low-rated and the rest of the novels appears to be very easy to model: low rating titles have a distinctive textual and sentiment profile. To make sure we are not incurring in inflated scores due to the special predictability of this "low-rating group", we repeated the experiment with only the novels with a higher rating than 2.5 (still the majority, ca. 8900 titles). Also in this case, the models performed better than random: able to predict the "quality slope" better than chance (see fig. 4 for a visualization of model performance). Given the relative tightness of the scale and the potential volatility of the scores themselves, we find the models' performance far from obvious. We finally looked at the most predictive features. When modelling the whole corpus, readers' judgments of quality appear inversely related to punctuation, text compressibility, reading ease, verb, pronoun and adverb frequency, and directly

	coefficient
Whole corpus	
Punctuation freq.	-3.261
Text compressibility	-2.841
Flesch reading ease	-2.205
Stopword freq.	-2.100
Verb freq.	-1.502
Pronoun freq.	-1.502
Flesch-Kincaid grade level	1.380
Adverb freq.	-1.004
Noun freq.	0.941
Lexical richness	0.697
Filtered corpus	
Pronoun freq.	-1.419
Nominal subject freq.	-0.761
Lexical richness	0.602
Adjective freq.	-0.436
New Dale–Chall readability formula	-0.351
Stopword freq.	-0.323
Relative clause modifier freq.	-0.263
Text compressibility	-0.231

Table 3: Most important non-sentiment features for the best performing model (Bayesian Ridge) in the whole (upper) and filtered corpus (lower).

	coefficient
Whole corpus	
Approximate entropy	-1.500
Mean sentiment	-1.352
Difference between main and ending	1.152
Beginning sentiment	-0.935
Ending sentiment	0.861
Hurst	-0.649
Std. deviation sentiment	0.295
Filtered corpus	
Hurst	0.576
Std. deviation sentiment	0.214
Beginning sentiment	-0.169
Approximate entropy	0.148
Mean sentiment	0.082

Table 4: Most important sentiment features for the best performing model (Bayesian Ridge) in the whole (upper) and filtered corpus (lower).

related to lexical richness and reading difficulty. A simplistic style combined with many verbs, adverbs and pronouns is linked to lower ratings. The most important sentiment measures were, negatively, approximate entropy, Hurst and mean sentiment, and positively, the difference between the arc's mean and the ending's sentiment, and the ending sentiment. In other words, texts that have particularly chaotic and unpredictable arcs receive low scores, while higher average sentiment and endings with more positive values receive higher scores. When filtering out the "low-rating few", the landscape changes. Novels have a higher perceived quality if they tend towards fewer pronouns, explicit subjects, adjectives, stopwords, relative clauses and repetitions, a higher lexical richness, more nouns and a slightly easier vocabulary. These features suggest a style that is more sophisticated, diverse in vocabulary, and concise, with simpler or more direct sentences, and less reliant on nominal subjects and adjectives. At the sentiment level, the Hurst exponent is the strongest predictor: GoodReads users favour novels that have more persistent sentiment arcs without being too flat nor repetitive in their sentimental palette (having a higher standard deviation and slightly higher approximate entropy). Literary quality appears associated with novels that have strong, coherent, and dynamic emotional progressions and a broader range of sentiment, with more intricate and nuanced changes. They may also start in the low end of sentiment and maintain a slightly more positive tone throughout. Overall, these measures seem to point to an equilibrium between simplicity and diversity, both at the stylistic and at the sentiment level.

5 Conclusion and future works

We have tried a new set of experiments in the highly challenging task of modelling literary quality, represented as the online average ratings of readers, from a small set of textual and sentiment features. While a similar attempts have been made before (on smaller corpora), to the best of our knowledge, we are the first to show that the *addition* of several sentiment-related features improves the predictive power of most models. The sentiment features considered here were of two kinds: a global kind, such as the mean sentiment of a novel; and a dynamic kind, such as the level of entropy and fractality of the sentiment arcs. We have also found that the bottom 2% of titles elicit distinctly lower ratings, and that their appreciation is partly predictable through the textual features we have included. Finally, we analysed the features needed to predict perceived literary quality, noting that a balance between simplicity and diversity seems to characterize more appreciated titles. Naturally this is a study on a complex subject. In the future we aim to repeat the experiment optimizing for quality proxies beyond GoodReads ratings to study convergences between ways of defining quality, and use a larger set of features. We may also set it as a classification problem, and attempt more sophisticated models, as long as some interpretability remains.

References

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

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

Jan Beran. 1994. *Statistics for Long-Memory Processes*, 1 edition. Chapman and Hall/CRC, New York.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022a. 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. 2022b. 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.

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.

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

A. Eke, P. Herman, L. Kocsis, and L. R. Kozak. 2002. Fractal characterization of complexity in temporal physiological signals. *Physiological Measurement*, 23(1):R1.

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

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.

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

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.

Kayvan Kousha, Mike Thelwall, and Mahshid Abdoli. 2017. Goodreads reviews to assess the wider impacts of books. 68(8):2004–2016.

Nikita Kuznetsov, Scott Bonnette, Jianbo Gao, and Michael A. Riley. 2013. Adaptive Fractal Analysis Reveals Limits to Fractal Scaling in Center of Pressure Trajectories. *Annals of Biomedical Engineering*, 41(8):1646–1660.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

Benoit Mandelbrot. 1982. *The Fractal Geometry of Nature*. Times Books, San Francisco.

Benoit B. Mandelbrot. 1997. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk. Selecta Volume E*, 1997 edition edition. Springer, New York.

Benoit B. Mandelbrot and John W. Van Ness. 1968. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review*, 10(4):422–437.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

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

Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. 27:16–32.

Andrew J. Reagan, Brian Tivnan, Jake Ryland Williams, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. 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.

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.

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

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press. Publication Title: Distant Horizons.

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

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

Melanie Walsh and Maria Antoniak. 2021b. The Goodreads "Classics": A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism. *Post45: Peer Reviewed*.

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

Readability measures
Flesch reading ease
Flesch-Kincaid Grade Level
SMOG Readability Formula
Automated Readability Index
New Dale–Chall Readability Formula
Stylometric features
Lexical diversity
Text compressibility
Sentence length
Syntactic features
Verb frequency
Noun frequency
Adjective frequency
Adverb frequency
Pronoun frequency
Punctuation frequency
Stopword frequency
Nominal subject frequency
Auxilary frequency
Passive auxiliary frequency
Relative clause modifier frequency
Negation modifier frequency
Simple sentiment arc features
Mean sentiment
Std. deviation sentiment
Ending sentiment
Beginning sentiment
Difference between main and ending
Sentiment arc measures
Hurst
Approximate entropy

Table 6: Textual and arc-features considered

Linear regression
Ridge regression
Lasso
Elastic net regularization
Bayes ridge regression
Huber loss
Polynomial
PLS
TheilSen

Table 5: Complete list of models

	Linear	Ridge	Lasso	ElasticNet	BayesRidge	Huber	Polynomial	TheilSen
Std. Deviation	0.052	0.045	0.03	0.03	0.049	0.037	0.14	0.121

Table 7: Standard deviation per model on the filtered corpus, not considering sentiment-features.