The fractality of sentiment arcs for literary quality assessment: The case of Nobel laureates

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

Kristoffer Nielbo
Aarhus University
kln@cas.au.dk

Telma Peura
University of Helsinki
telma.peura@gmail.com

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

Abstract

In the few works that have used NLP to study literary quality, sentiment and emotion analysis have often been considered valuable sources of information. At the same time, the idea that the nature and polarity of the sentiments expressed by a novel might have something to do with its perceived quality seems limited at best. In this paper, we argue that the fractality of narratives, specifically the long-term memory of their sentiment arcs, rather than their simple shape or average valence, might play an important role in the perception of literary quality by a human audience. In particular, we argue that such measure can help distinguish Nobel-winning writers from control groups in a recent corpus of English language novels. To test this hypothesis, we present the results from two studies: (i) a probability distribution test, where we compute the probability of seeing a title from a Nobel laureate at different levels of arc fractality; (ii) a classification test, where we use several machine learning algorithms to measure the predictive power of both sentiment arcs and their fractality measure. Our findings seem to indicate that despite the competitive and complex nature of the task, the populations of Nobel and non-Nobel laureates seem to behave differently and can to some extent be told apart by a classifier.

1 Introduction

The question of what defines the perception of quality in literature is probably as old as narrative itself, but the ability to process and analyze large quantities of literary texts, and to perform complex statistical experiments on them (Moretti, 2013), has recently made new ways of studying this question possible. This does not mean that the riddle has become easy at all: first of all, studying literary quality with methods from corpus linguistics means that one has to create a dataset of "high quality" texts, usually to contrast against "lower quality" texts; second, while it is possible to analyze a larger number of texts in a shorter time, we need to know where to look to find possible, non-obvious correlations with the perception of quality. Recently, a series of studies have looked into the possibility of correlating some fractal properties of a text - the degree of fractality of its sentences' length, sentiment arc, or succession of topics - with its literary quality. These studies have been using as a proxy to define the quality of a text either canons defined by a single scholar, or majority-vote measures taken by large reader platforms, where the aggregated score given by a large number of readers is used as the value of the book, often with a threshold to transform it into a binary problem. Other similar works have used the number of sales of a book to approximate its "quality".

In this work, we try to use a perhaps more daring, less explored metric to define quality: we apply an already tested measure: the fractality of the sentiment arc of a text, which is the curve that represents the changes in sentiment throughout the text. We compute this metric for a group of texts written by authors who won the Nobel Prize for Literature, and we ask whether this simple measure can help tell such texts from a highly competitive control group.

Despite the difficulty of the task - in the best cases, Nobel Prizes are assigned to only one among many valid competitors, which means that several high quality writers will fall in the negative class - our results seem to indicate that a weak but reliable signal is present, and that it can be exploited by classic machine learning algorithms to predict whether a narrative's arc belongs to a Nobel laureate or not.

The paper is organized as follows: in Section 2, we describe some of the most relevant related works in sentiment analysis and fractal theory for

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 31–41
November 20, 2022. ©2022 Association for Computational Linguistics

studies in literary quality. In Section 3 we present the corpus and discuss the idea of using Nobel Prize winners; Section 4 gives a detailed overview of the concept of series fractality for sentiment arcs. Finally, Section 5 details the settings of our experiments and Section 6 presents our main results. In Section 7 we discuss our conclusions and possible future works.

2 Related Works

2.1 Sentiment Arcs of Narratives

Drawing the sentiment arc of a story is one of the simplest methods to abstract a narrative's shape. At the same time the sentiment or emotional aspect of communication is often regarded as one of the most relevant in narrative, especially "artistic" narrative (Drobot, 2013), as it is linked with the central and somewhat unique property of literary texts of evoking, rather than describing, experiences and inner states. As Hu et al. (2021) argues, readers have to emotionally engage with the evolution of the story, and a sentiment arc is an index of those engagement "prompts". For this reason, sentiment analysis models (Alm, 2008; Jain et al., 2017), at the word (Mohammad, 2018), sentence (Mäntylä et al., 2018) or paragraph (Li et al., 2019) level, have often been employed in computational literary studies (Cambria et al., 2017; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017). Sentiment analysis usually draws its scores from human annotations of single words (Mohammad and Turney, 2013) or from lexicons induced from labelled documents (Islam et al., 2020). Several studies have tried to complement the essentiality of sentiment analysis with algorithms for textual emotion detection (Alm et al., 2005), or by developing more complex SA tools (Xu et al., 2020). Scholars usually analyse sentiment arcs in terms of their overall shape (Reagan et al., 2016), but recent developments have looked for more complex mathematical properties (Gao et al., 2016).

2.2 Fractality

The study of fractals (Mandelbrot and Ness, 1968; Mandelbrot, 1982, 1997), especially applied to long series (Beran, 1994; Eke et al., 2002; Kuznetsov et al., 2013) offers a new way of looking into the properties of narrative and literary texts, exploring their degree of predictability or self-similarity (Cordeiro et al., 2015), following links with fractal properties already found in visual arts and musics. Recently, Mohseni et al. (2021) have looked into the degree of fractality of canonical and non-canonical literary texts using a series of classical stylometric features such as sentence length, type-token ratio and part of speech ratio, while Hu et al. (2021) applied fractal analysis to a novel's sentiment arc. Bizzoni et al. (2022) explore this possibility further, showing that sentiment arcs' fractality appears to correlate with the perceived quality of literary fairy tales. Nonetheless, not all studies on literary quality have relied on sentiments or fractality: important results have also been obtained with much simpler measures such as bigram frequency (van Cranenburgh and Koolen, 2015).

2.3 Quality

The idea that readers' perception of what is pleasant or engaging could be found in complex statistical patterns has given rise to a series of attempts to approach literary quality using quantitative models (Moretti, 2013). While it is hardly meaningful to define an absolute measure for something like the apperception of quality, this line of research has had to define strategies to approximate a value of quality for a dataset of texts. To "measure quality", most works to this date have looked for large scale collections of readers' preferences, from books' sales to average scores on reading platforms such as GoodReads (Kousha et al., 2017), while a smaller number of work has instead tried to rely on established literary canons (Wilkens, 2012). Although these two concepts of quality are distinct and often retrieve different collections of titles, Walsh and Antoniak (2021) have observed that their overlap might be much larger than expected. In both cases, the possibility of comparing different canons and different aggregations of readers' preferences has opened the possibility of expanding the scope and reliability of aesthetic studies of literature (Underwood, 2019; Wilkens, 2012).

3 A dataset of Nobel literature

The first problem in determining the relationship between sentiment arcs and literary quality is finding a metric for literary quality itself; and it could be argued that the problem of finding a reliable source of quality judgments is the same that every individual reader has when faced with an amount of literature too large to read and evaluate alone (Underwood, 2019) - it's one of the main reasons why literary awards exist at all. While several previ-

	N. Authors	N. Titles
Whole corpus	7000	9089
Nobel group	18	85
Control group	738	1312

Table 1: Overall titles and authors in the corpus, number of Nobel laureates and dimensions of the control group.

ous works attempting to classify or measure some aspect of literary quality have relied on quantitative metrics such as number of copies sold or average reviews on large scale platforms - or even in newspapers - few attempts have been made, to the best of our knowledge, to use high prestige literary awards as a central metrics to approximate the quality of a work. In this paper we try to use arguably the most prestigious international literary award, the Nobel Prize for Literature, as our sole guide to select literary quality. Naturally, this setting is a deliberate extremization: no literary award can possibly be considered the unique indicator of what literary quality is, and questions on the sensibility of the Nobel committee's choices, both in terms of who got the prize and in terms of who did not receive it, rise almost every year. At the same time, high level literary prizes can work as imperfect guidelines for one kind of quality, and it would be interesting to find out whether, on a larger scale than single books or authors, a "signal" telling Nobel-winning texts from a control population can be found.

Unfortunately, no comprehensive corpus of Nobel-winning authors exists to date. To carry out our experiment, we used a recent corpus of literary texts, the Chicago Corpus, compiled by Hoyt Long and Richard Jean So, composed of 9089 novels published in the US between 1880 and 2000. The corpus contains key works of US Nobel laureates, seminal works from mainstream literature as well as relevant works in genres such as Mystery and Science Fiction (Long and Roland, 2016). [1]

The US Nobel laureates in the corpus make the relative majority of the group of Nobel laureates, e.g. John Galsworthy, Sinclair Lewis, William Faulkner, Ernest Hemingway, John Steinbeck, Saul Bellow and Toni Morrison. Works by non-US writers like for example Knut Hamsun, Samuel Beckett and Nadine Gordimer are represented with a more limited selection of their work.

As noted, the corpus is highly curated and contains high quality fiction from authors who have received other prizes, like the National Book Award, e.g. Don DeLillo, Joyce Carol Oates, and Philip Roth. Our expectation is therefore not that Nobel laureates will be completely different from the rest of the corpus, also in terms of literary quality.

Finally it is worth noting that the whole corpus is heavily skewed towards the Anglosaxon literature, and that both the Nobel laureates and their control group are mainly constituted by Anglophone writers. This naturally moves the whole contest on the plain of a well refined "Anglo-centric" canon. While it does not damage our experiments per se, given that the same imbalance happens among the Nobel laureates as among the remaining writers, it is a distortion that we have to keep in mind.

4 Fractality of sentiment arcs

To estimate the long-term memory of sentiment arcs we combine non-linear adaptive filtering with fractal analysis, specifically adaptive fractal analysis (Gao et al., 2011; Tung et al., 2011). Non-linear adaptive filtering is used because of the inherent noisiness of story arcs. First, the signal is partitioned into segments (or windows) of length $w = 2n + 1$ points, where neighboring segments overlap by $n + 1$. The time scale is $n + 1$ points, which ensures symmetry. Then, for each segment, a polynomial of order D is fitted. Note that $D = 0$ means a piece-wise constant, and $D = 1$ a linear fit. The fitted polynomial for ith and $(i + 1)th$ is denoted as $y^{(i)}(l_1), y^{(i+1)}(l_2)$, where $l_1, l_2 = 1, 2, ..., 2n + 1$. Note the length of the last segment may be shorter than w. We use the following weights for the overlap of two segments.

$$y^{(c)}(l_1) = w_1 y^{(i)}(l + n) + w_2 y^{(i)}(l), \quad l = 1, 2, \ldots, n + 1 \quad (1)$$

where $w_1 = (1 - \frac{l-1}{n}), w_2 = 1 - w_1$ can be written as $(1 - \frac{d_j}{n}), j = 1, 2$, where d_j denotes the distance between the point of overlapping segments and the center of $y^{(i)}, y^{(i+1)}$. The weights decrease linearly with the distance between the point and center of the segment. This ensures that the filter is continuous everywhere, which ensures that non-boundary points are smooth.

We use the Hurst exponent to measure long-term memory. Assuming that stochastic process $X =$

[1] Several quantitative literary studies have used the corpus (Underwood et al., 2018; Cheng, 2020), which can be found at https://textual-optics-lab.uchicago.edu/us_novel_corpus.

$X_t : t = 0, 1, 2, ...$, with stable covariance, mean μ and σ^2, the process' autocorrelation function for $r(k), k \geq 0$ is:

$$r(k) = \frac{E\left[X(t)X(t+k)\right]}{E\left[X(t)^2\right]} \sim k^{2H-2}, \text{as} \quad k \to \infty \tag{2}$$

where H is called the Hurst exponent (Mandelbrot, 1982). For $0.5 < H < 1$ the story arc is characterized by persistent such that increments are followed by increases and decreases by further decreases. For $H = 0.5$ the story arc only has short-range correlations; and when $H < 0.5$ the story arc is anti-persistent such that increments are followed by decreases and decreases by increments. For the specific application domain (i.e., narratives) persistent story arcs are characteristic of coherent narratives, where the emotional intensity evolves at longer time scales. Story arcs' that only show short memory lack coherence and appear like a collection of short stories. Anti-persistent story arcs will appear bland and rigid narratives oscillating around an average emotional state (Hu et al., 2021).

Detrended fluctuation analysis (DFA) is the most widely used method for estimating the Hurst parameter, but DFA may involve discontinuities at the boundaries of adjacent segments. Such discontinuities can be detrimental when the data contain trends (Hu et al., 2001), non-stationarity (Kantelhardt et al., 2002), or nonlinear oscillatory components (Chen et al., 2005; Hu et al., 2009). Adaptive fractal analysis is a more robust alternative to DFA (Gao et al., 2011; Tung et al., 2011). AFA consists of the following steps: first, the original process is transformed to a random walk process through first-order integration $u(n) = \sum_{k=1}^{n}(x(k) - \overline{x}), n = 1, 2, 3, ..., N$, where $\overline{x}$ is the mean of $x(k)$. Second, we extract the global trend $(v(i), i = 1, 2, 3, ..., N)$ through the nonlinear adaptive filtering. The residuals $(u(i) - v(i))$ reflect the fluctuations around a global trend. We obtain the Hurst parameter by estimating the slope of the linear fit between the residuals' standard deviation $F^{(2)}(w)$ and w window size as follows:

$$F^{(2)}(w) = \left[\frac{1}{N}\sum_{i=1}^{N}(u(i) - v(i))^2\right]^{\frac{1}{2}} \sim w^H \tag{3}$$

All our sentiment arcs are sentence based, extracted using the VADER model (Hutto and Gilbert, 2014) in NLTK's implementation (Bird, 2006). While VADER is not the most recent Sentiment Analysis model, we chose it for its transparency, since it is possible to reconstruct the reasons of its judgments based on its systems of rules, as well as its popularity, as its underlying dictionary and set of rules has proven the weapon of choice for a large number of previous works. The sentiment arc is obtained by first computing the sentiment of each word in the text, and then by computing the average sentiment of each sentence. The sentiment of a word is in turn obtained using an ad-hoc lexicon, which links a sentiment score to each word and takes care of morphological variations. The sentiment of a sentence is then computed as the average of the sentiment scores of all the words in that sentence, by taking care of tricky structures like negations, intensifiers and so forth.

5 Experiments

We present the results for two experiments:

1. Without directly testing the predictive power of narrative sentiment arcs and their Hurst exponent, we analyzed its distribution in both Nobel-winning and non-Nobel-winning populations, to test whether the two populations might differ in their average score;

2. To directly test the predictive power of our Hurst exponent, we ran a series of classifiers to check whether sentiment arcs and their Hurst score can provide a degree of predictive power on telling whether or not a given text is likely to belong to a Nobel-winning author.

In both cases, we decided to design the non-Nobel-winning class (or control group) in order to be as contextual to the Nobel population as possible: for each book belonging to an author who won the Nobel prize, we took all novels published between one year before and one year after its publication date, and we considered them as the "control group" for that book. All the control groups for all books of one author work as the control group for that author, and all control groups together combine into the overall control group for the Nobel prize population. We did this also to mimic as much as possible the logic of the prize itself, that selects between contemporary candidates. A detailed summary of this selection process can be seen in Table 2.

Nobel	N. titles	Control
S. Beckett	1	32
S. Bellow	5	228
W. Churchill	4	125
W. Faulkner	15	332
J. Galsworthy	9	105
W. Golding	2	6
N. Gordimer	2	3
K. Hamsun	1	1
E. Hemingway	7	170
R. Kipling	3	19
D. Lessing	3	34
S. Lewis	8	137
T. Morrison	5	192
A. Munro	1	2
J. Steinbeck	15	81
R. Tagore	1	19
S. Undset	2	32
P. White	1	0
Total	85	1518

Table 2: Number of titles per Nobel and control group. Notice that the control group's total number is higher than the one reported in Table 1 since one title can figure in more than a subgroup.

5.1 Probability distribution

In the first experiment, we simply focused on the possibility that the Nobel-winning population might have a different Hurst score distribution than the control group, and that such difference might be statistically significant on the large scale. To further test this idea, we divided our corpus in Hurst classes (e.g. all titles having a Hurst score of 0.51, 0.52, etc.) and we looked at the probability of seeing a title from a Nobel laureate in each of these classes. To deal with the problem of having a heavily imbalanced dataset, since the control authors are much more numerous in any class than Nobel winning authors, we computed the probabilities on a sub-sampled portion of the control group as large as the Nobel group, so that both populations sum up to the exact same amount. Finally, in order to avoid relying on random lucky or unlucky sub-samplings from the majority class, and in general to increase the representativity of our comparison, we repeated the random majority class sub-sampling 100 times and drew the average probability for each Hurst class. The result is that for each class of Hurst values, we compute the probability of seeing a Nobel author's title and the average probability of seeing a non-Nobel author's title as computed over several subsamples.[2]

5.2 Classification

In the second experiment, we trained four different classifiers:

- **Quadratic Discriminant Analysis** classifier (Bose et al., 2015): a generative model that is particularly apt to classify data when the decision boundaries are non-linear;
- **Gaussian Naive Bayes** classifier (Chan et al., 1982): we chose this model particularly for its ability to handle small and complex training data;
- **Random Forest** classifier (Ho, 1995): this algorithm is well suited to make fine-grained predictions on data that are not necessarily linearly divisible;
- **Decision Tree** classifier, which has the benefits of being simple and able to handle relatively small datasets (Swain and Hauska, 1977).

As features, we used the Hurst score and a condensed version of the sentiment arc for each novel.

The large difference in our classes' sizes represents an additional difficulty. The sparsity of Nobel titles makes training on the dataset as is a seemingly meaningless task, since classifiers systematically ignore or misrepresent the minority class. To contrast that dataset's imbalance, we tried three resampling techniques:

- **Random** subsampling: this is the easiest resampling technique, and it simply means that we randomly drew from the majority class a number of data points equal to the size of the minority class, as we did in Section 5.1;
- **Near Miss** subsampling (Mani and Zhang, 2003; Bao et al., 2016), specifically the so called *Near-Miss 1* method: this is a more sophisticated undersampling technique based on the distance between items from the majority and items from the minority class, where the elements from the majority class with the smallest average distance to three minority class examples are selected for comparison.

[2] This naturally means that the probabilities do not necessarily sum up to 1.

	Score	**p-value**
T-test	2.57	0.01
Anova	6.63	0.01
Mann-Whitney U	55106	0.023
Kruskal-Wallis	5.166	0.023

Table 3: Difference between Nobel laureates and control group as tested by four significance measures (the first two assume that the populations have a normal distribution, the last two do not make such assumption). In all cases, the difference in Hurst score distributions is statistically significant.

In this way, the algorithm selects datapoints that are closest to the decision boundary;

- **SMOTE** upsampling (Chawla et al., 2002), a upsampling technique widespread in machine learning, often used in cases of severely imbalanced datasets (Liu et al., 2019; Rustogi and Prasad, 2019). SMOTE has the considerable benefit of creating not simple duplicates of the observed datapoints, but rather slightly different synthetic datapoints, increasing the ability of a classifier of modeling a minority class.

6 Results

6.1 Probability distribution

The difference between the distributions of Hurst scores for the Nobel and the control group is statistically significant according to several measures, as can be seen in Table 3.

The probability of seeing a text from a Nobel laureate peaks at a different point than the probability of seeing a text from the control group (see Figure 1). The distribution of the two groups reinforces the hypothesis, laid by Hu et al. (2021), that high literary quality might lie in a specific area on the Hurst continuum - in other words, that there might be a specific interval of Hurst values where high quality narrative texts are most likely to fall. Naturally we should not ignore the fact that the two probability distributions have a considerable overlap; that the statistical significance, while being strong, does not mean that the two groups are completely separable; and that the number of control titles is higher than the number of titles from Nobel-winning authors for any Hurst interval. In other words, any text has a lower probability of belonging to a Nobel laureate than of belonging to an author that did not win the Nobel prize - after all it's possible to award the Nobel prize to just one person every year. At the same time, if we take equal-sized classes for the two groups, texts having a Hurst score ranging approximately between 0.53 and 0.61 seem to have a higher probability of belonging to a Nobel laureate than of belonging to a control author, while texts falling outside of this range have a higher probability of belonging to a control author than of belonging to a Nobel laureate: again, the Nobel population and the control population display statistically different behaviours on the Hurst continuum. Figure 1 offers a visualization of our results.

A cursory qualitative examination of the results for different authors proved that these results often (but not always) correspond to what we might expect from a given title or author. For example John Steinbeck, one of the best represented writers in the corpus with 15 novels, has an average Hurst exponent of 0.598, and thus differs insignificantly from the 90 works in its control group, that score an average of 0.606, but with a more significant standard deviation (0.41 vs. 0.25). While Steinbeck's novels Hurst scores range from 0.56 to 0.64, the two novels that get by far the highest average grades on GoodReads (*Mice and Men* and *The Grapes of Wrath* with *Cannery Row* as a very distant third) both have a Hurst exponent of exactly 0.58, at the apex of the probability curve for Nobel titles. Similar observations can be made for the works for other popular Nobel laureates, such as Hemingway, with his most renowned titles (such as for example *The Old Man and the Sea* or *For whom the bell tolls*) roughly falling within what we considered a fuzzy Goldilocks interval for literary quality, while less acclaimed texts such as *To have and have not* are clearly out of it (Figure 1). Many other factors play into the success of these prominent novels, but their location in the middle of what seems to be a "Goldilocks"-zone for variability is significant, also when studied on the level of the individual authorship.

6.2 Classification

Among the three techniques we adopted to resample our dataset, we found that randomly undersampling the majority class does not yield particularly strong results, while Near Miss understampling and SMOTE oversampling both bring the models to better performances (see Figure 2). The reason for this lies probably in the fact that the difference between

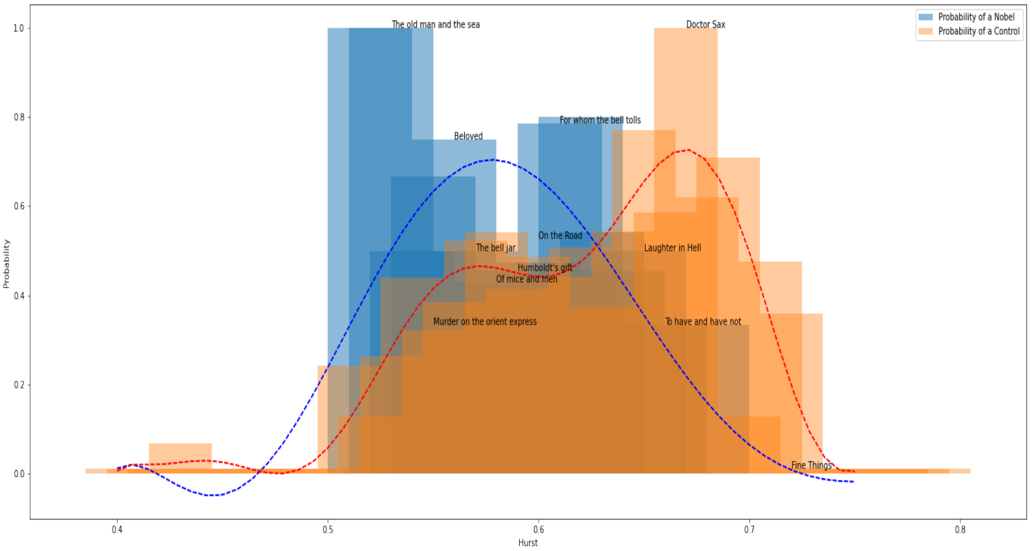

Figure 1: Probability distribution of the Nobel group and of the control group. The control population's probabilities are averaged over 100 different selections. We added some titles for reference. Not all works from Nobel laureates fall in the Hurst "sweet spot": for example, *The Old Man and The Sea* has a Hurst score of 0.53, while the less acclaimed *To Have and Have Not* from the same author has a Hurst score of 0.69.

the two populations, while present, is quite difficult to pick up even when we control for size: after all, we are using a corpus with a large number of high quality authors that did not win a Nobel prize, so the control group is both much larger than the Nobel group and bound to have several elements similar to its members. Just randomly subsampling from the majority class to create a small group of non-Nobels to learn from makes the task very difficult, while an algorithm like Near Miss, that selects data with the least distance to the negative classe's samples, essentially selecting learning cases that is most fruitful for the classifier to model, brings significantly better results. Finally, it's worth noting how SMOTE upsampling brings about the highest performances of the group (excluding the "All dataset" case): while this technique does not create completely dependable results, since it relies on the synthetic generation of new data points for the minority class, its effectiveness can make us more confident in postulating that a difference between the Nobel and the control populations does indeed exist.

In Table 4 we provide a summary of the performances, adding in parenthesis the performance of the classifiers when they are only fed information from the sentiment arcs, without accessing the Hurst exponent. The comparison seems to us quite interesting: the sentiment arcs seem to suffice in bringing about better-than-chance performances, and in some cases even quite high scores; on the other hand, all classifiers trained on a feature set enriched by the single dimension of the arcs' Hurst exponent perform better than when they do not have access to such information, with no exception, and in some cases the single presence of the Hurst exponent increases the F scores significantly.

7 Discussion and Conclusions

In this paper we have tried to use a measure of fractality for sentiment arcs to distinguish Nobel-winning writers in a corpus of selected literary texts in the English language, as a case for the relevance of this metric in literary quality evaluation. We are not interested in the overall valence of a literary work as such, but in its patterns of variation and repetition throughout the narrative arc, although the underlying argument for using sentiment analysis (and not just, for example, PoS tagging) is that it can be linked to the evocation of emotions in the work. Even if it is far from catching the expressions of emotions perfectly, as there are many way to express them, also through words with a neutral sentiment, we believe it remains a strong

	Original dataset	**Random Subs.**	**Near Miss**	**SMOTE Ups.**
Quadratic Discr. An.	0.90 (0.90)	0.55 (0.51)	0.56 (0.51)	0.57 (0.50)
Gaussian Naive Bayes	0.91 (0.90)	0.52 (0.49)	0.80 (0.67)	0.67 (0.53)
Decision Tree Cl.	0.88 (0.88)	0.57 (0.52)	0.69 (0.60)	0.87 (0.82)
Random Forest	0.91 (0.90)	0.53 (0.51)	0.79 (0.62)	**0.90 (0.86)**
Average				

Table 4: Weighted F scores, averaged from a 10-fold cross-validation, for four classifiers trained on different versions of the dataset. Notice how the results on the "all dataset" column are effects of the majority class being overwhelmingly larger than the minority class. In parenthesis, we add the performances when only using information from the sentiment arcs. The other three columns, reporting results based on resampled versions of the dataset, do not resent of the distortion.

indicator of the work's rhetoric appeal structure. Overall, the best attitude towards this kind of metric is probably similar to the attitude we can have towards the aesthetic properties of fractals in music or visual arts: it is never *necessary* for a work of art to contain anything fractal, but on the large scale we could expect fractal patterns to hold a correlation with the perception of beauty. In the same way, we should not imagine a systematic relationship between quality and a given range of Hurst exponents: first of all because there is no single way to measure literary quality, and second because a "good" Hurst exponent can hardly be the single factor in high quality textual narrative. Nonetheless, we have found that the distribution of Hurst exponents, as computed on the sentiment arcs of whole novels, for the titles of authors who won a Literature Nobel Prize is different from the distribution of Hurst exponents for the titles of the control group. This is particularly relevant considering that the control group still included several high-level writers, from Nabokov to Woolf, who can be said to rival the Nobel population in terms of both fame and critical acclaim. What this difference in distribution seems to indicate is that there might be a "sweet spot" of self-similarity in sentiment arcs, roughly between 0.53 and 0.61, where the probability of seeing a text from a Nobel laureate grows and the probability of seeing a title from a non-Nobel laureate decreases. Following on this finding, we tried to create a classifier that would tell whether a text came from a Nobel laureate or not based on its Hurst exponent and a representation of its sentiment arc only. What we found is that when we control for data imbalance by using Near Miss subsampling or SMOTE upsampling, classifiers appear to perform well above chance, while if we subsample randomly their performance suffers considerably. We consider this a indication that a "signal" for Nobel laureates exists, despite the highly competitive control group, and that it falls in line with previous studies on the Hurst exponent for sentiment arcs.

8 Future Work

Given the scope and complexity of the concept of literary quality, there are several interesting directions this research can take. A sensible next step would be to increase the size of our corpus to include more texts, in order to see if the signal for Nobel laureates becomes more pronounced. Specifically, we aim at increasing the number of titles in the minority class, both by looking at other prestigious awards and by including not only the winners, but also the list of nominees. Being pre-selected for a prestigious award, nominees could help creating a larger "quality class" and might even temper the random or political factors playing in the choice of a single individual winner. The Chicago corpus does not offer such information in its metadata, but it is still possible and even relatively easy to access it for the Nobel prize. Other large English language prizes like the Pulitzer Prize would also be of great interest to create a larger subset. Another goal worth striving for, albeit on a longer time scale, is to include a more diverse range of titles. The Chicago corpus is constituted mainly of Anglophone writers - both the Nobel group and its control are heavily skewed towards the Anglo-Saxon literature. Finally, the internal imbalance in the amount of titles that different Nobel laureates hold in our selection might play a role in the be-

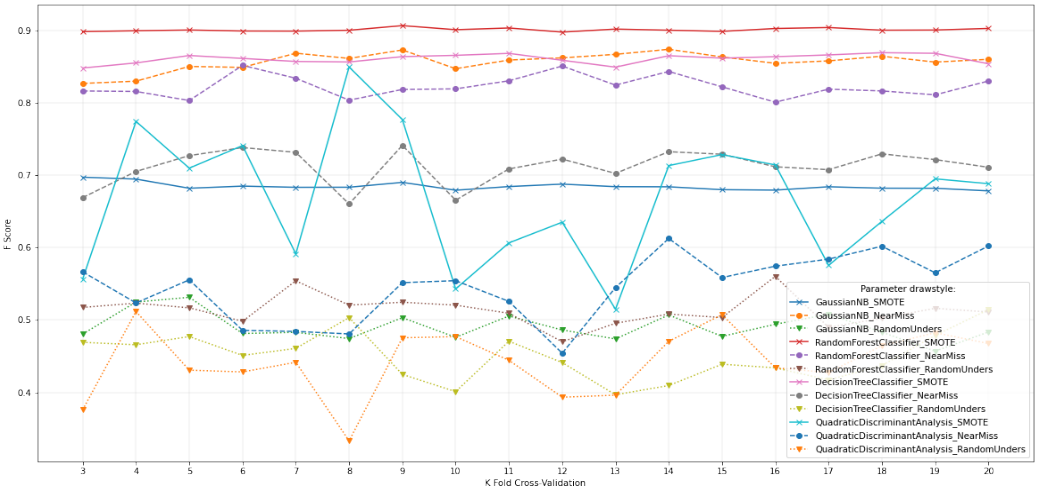

Figure 2: Classification results for our 4 classifiers under three different assumptions: random undersampling, Near Miss undersampling and SMOTE upsampling, with increasing number of folds in a K-folds cross-validation.

haviour of the systems. While we are comforted by the fact that the same metrics have proved useful with completely different authors in previous works, in future we would like to design ablation experiments aimed at checking the performance of the machine learning models on the less represented names. Finally, it would be interesting to see if this signal is specific to English-language texts or if it appears in other languages as well.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586.

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

Lei Bao, Cao Juan, Jintao Li, and Yongdong Zhang. 2016. Boosted near-miss under-sampling on svm ensembles for concept detection in large-scale imbalanced datasets. *Neurocomputing*, 172:198–206.

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

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

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

Smarajit Bose, Amita Pal, Rita SahaRay, and Jitadeepa Nayak. 2015. Generalized quadratic discriminant analysis. *Pattern Recognition*, 48(8):2676–2684.

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.

Tony F Chan, Gene H Golub, and Randall J LeVeque. 1982. Updating formulae and a pairwise algorithm for computing sample variances.

In *COMPSTAT 1982 5th Symposium held at Toulouse 1982*, pages 30–41. Springer.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Zhi Chen, Kun Hu, Pedro Carpena, Pedro Bernaola-Galvan, H. Eugene Stanley, and Plamen Ch. Ivanov. 2005. Effect of nonlinear filters on detrended fluctuation analysis. *Phys. Rev. E*, 71(1):011104.

Jonathan Cheng. 2020. Fleshing out models of gender in english-language novels (1850–2000). *Journal of Cultural Analytics*, 5(1):11652.

João Cordeiro, Pedro R. M. Inácio, and Diogo A. B. Fernandes. 2015. Fractal beauty in text. In *Progress in Artificial Intelligence*, Lecture Notes in Computer Science, pages 796–802. Springer International Publishing.

Andreas van Cranenburgh and Corina Koolen. 2015. Identifying literary texts with bigrams. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 58–67.

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.

Jianbo Gao, Jing Hu, and Wen-wen Tung. 2011. Facilitating Joint Chaos and Fractal Analysis of Biosignals through Nonlinear Adaptive Filtering. *PLoS ONE*, 6(9):e24331.

Jianbo Gao, Matthew L Jockers, John Laudun, and Timothy Tangherlini. 2016. A multiscale theory for the dynamical evolution of sentiment in novels. In *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, pages 1–4. IEEE.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

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.

Kun Hu, Plamen Ch. Ivanov, Zhi Chen, Pedro Carpena, and H. Eugene Stanley. 2001. Effect of trends on detrended fluctuation analysis. *Physical Review E*, 64(1).

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).

Jan W. Kantelhardt, Stephan A. Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H. Eugene Stanley. 2002. Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications*, 316(1-4):87–114.

Evgeny Kim and Roman Klinger. 2018. A survey on sentiment and emotion analysis for computational literary studies.

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.

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. *arXiv preprint arXiv:1910.00883*.

Shiyu Liu, Ming Lun Ong, Kar Kin Mun, Jia Yao, and Mehul Motani. 2019. Early prediction of sepsis via smote upsampling and mutual information based downsampling. In *2019 Computing in Cardiology (CinC)*, pages Page–1. IEEE.

Hoyt Long and Teddy Roland. 2016. Us novel corpus. Technical report, Textual Optic Labs, University of Chicago.

Benoit Mandelbrot. 1982. *The Fractal Geometry of Nature*, updated ed. edition edition. 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.

Inderjeet Mani and I Zhang. 2003. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, pages 1–7. ICML.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2.

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

Franco Moretti. 2013. *Distant reading*. Verso Books.

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

Rishabh Rustogi and Ayush Prasad. 2019. Swift imbalance data classification using smote and extreme learning machine. In *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, pages 1–6. IEEE.

Philip H Swain and Hans Hauska. 1977. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3):142–147.

Wen-wen Tung, Jianbo Gao, Jing Hu, and Lei Yang. 2011. Detecting chaos in heavy-noise environments. *Physical Review E*, 83(4).

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

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction. *Journal of Cultural Analytics*, 3(2):11035.

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.

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

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2020. Dombert: Domain-oriented language model for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.13816*.