Emotional Complexity as a Measure for Literary Reception

Anonymous ACL submission

Abstract

We introduce 'EmotionArcs', a dataset comprising emotional arcs from over 9,000 English novels, assembled to understand the dynamics of emotions represented in text and how these emotions may influence a novel's reception and perceived quality. Through the paper, we discuss the challenges of emotion annotation, suggesting improvements based on theory and case studies to redefine how emotions are modeled in literary narratives. Finally, we use information-theoretic measures to analyze the impact of emotions on literary quality. We find that emotional entropy, as well as the skewness and steepness of emotion arcs correlate with two proxies of literary reception. Our findings may offer insights into how quality assessments relate to emotional complexity and could help with the study of affect in literary novels.

1 Introduction

The importance of a literary text's emotional profile for its overall quality ("performance", reception) is hard to overestimate (Bal and Van Boheemen, 2009). While literary narratives are far from being only matters of emotions, emotions touched upon in texts - in both explicit *and* evocative ways - determine essential aspects of the overall plot formation, at the structural level, and of the reader's experience, at the stylistic level (Mar et al., 2011). Due to the complexity of human readers' interpretations and experiences of texts, the modeling of "emotions in texts" - defining what we mean by that, deciding which emotions to define, and how to measure the emotional content of any given textual unit is neither simple nor straightforward. Neither is quantifying the reception or perceived quality of a literary narrative, nor measuring the subsequent relation between "a texts emotions" and reader appreciation. In this paper, we introduce a new resource, 'EmotionArcs', to explore the relationship between emotion arcs and literary quality complete with some early analyses. 'EmotionArcs', is a dataset that comprises emotional arcs constructed from over 9,000 English novels through a novel approach that utilizes emotion intensity lexicons enhanced by word embeddings fine-tuned for the domain of literature to construct emotion arcs. We use the dataset to analyze and measure how affective language relates to a novel's literary quality, as measured through literary awards and GoodReads' ratings.

2 Related Work

2.1 Emotion Analysis

Sentiment analysis and emotion detection are by nature subjective as not even humans can typically agree on which emotions any specific text contains (Campbell, 2004; Bayerl and Paul, 2011). There are also crucial distinctions between whether we are measuring the evocation or association of emotions and whether we are doing this from the reader's or the writer's perspective (Mohammad, 2016). Approaches to sentiment analysis garner critique both for inherent problems in, for example, word-based annotation (Swafford, 2015), but also for being overly focused on evaluation metrics over applicability to downstream tasks (Öhman, 2021) and how the task of emotion detection to some degree constructs the phenomena it is trying to measure (Laaksonen et al., 2023).

Previous work has tested the potential of sentiment analysis (Alm, 2008; Jain et al., 2017) at the word (Mohammad, 2018a), sentence (Mäntylä et al., 2018), or paragraph level (Li et al., 2019), for capturing meaningful aspects of literary texts and the reading experience (Drobot, 2013; Cambria et al., 2017; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017; Reagan et al., 2016). Sentiment arcs have been used in multiple studies to model and evaluate narratives in terms of literary genre (Kim et al., 2017), plot archetypes (Reagan et al., 2016), dynamic properties (Hu et al., 2021), and reader preferences and perceived quality (Bizzoni et al., 2022a). To model the sentiment arcs of novels (Jockers, 2017) – that is, detrended arcs based on raw valence or emotion scores at the word or sentence level – studies have annotated sentiment in narratives, usually drawing scores of words or sentences from induced lexica (Islam et al., 2020), or human annotations combined with machine learning (Mohammad and Turney, 2013). Challenges inhere to each approach (Da, 2019; Öhman, 2021; Teodorescu and Mohammad, 2023), and new methods for estimating both the valence and emotions of texts are developing quickly.

The nature and origin of emotions is an active field of research with many competing schools of thought. Many approaches are based on the theory of universal emotions by Ekman (1971), including Plutchik's wheel of emotions (Plutchik, 1980), and SenticNet (Cambria et al., 2012). More recently Cowen and Keltner (2021) have tried to expand on these models using computational methods with promising results. Still, current resources are predominantly based on the theories of Plutchik (1980) including the NRC Emotion intensity lexicon (Mohammad, 2018b) used in this paper.

Because literary texts have additional layers of affective meaning (cf. the distinction between tone and mood) at more narrative levels, (narrator, character, style, etc.) than other texts, additional challenges accompany annotating emotions in them. However, several recent papers have shown that lexicon-based methods produce accuracies comparable to machine learning-based methods using chunks or bin sizes (a set number of tokens) of only a few hundred tokens with the additional benefit of transparency and human interpretability (Öhman, 2021; Teodorescu and Mohammad, 2023).

2.2 Literary Quality

Studies that aim to forecast the perception of literary quality by relying on textual features[1] have mostly depended on stylometry (the study of variations in literary style). This includes factors like sentence length and readability (Maharjan et al., 2017; Koolen et al., 2020), the proportion of different classes of words (Koolen et al., 2020), and the frequency of certain word pairs (n-grams) in texts (van Cranenburgh and Koolen, 2020). Other recent studies have explored the use of alternative textual or narrative elements such as sentiment analysis (Alm, 2008; Jain et al., 2017), to capture a significant aspect of the reading experience (Drobot, 2013; Cambria et al., 2017; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017; Reagan et al., 2016). This strand of research predominantly focuses on sentiment valence with the aim of roughly modeling the sentiment arcs – the ups and downs – of novels (Jockers, 2017). Once arcs are computed, it is possible to cluster them based on similarities (Reagan et al., 2016). More recently, Hu et al. (2021) and Bizzoni et al. (2022a) applied fractal analysis, a technique to study complex systems' dynamics (Hu et al., 2009), to model the persistence, coherence, and predictability of sentiment arcs. This approach aimed to assess the predictability and self-similarity of arcs and relate it to reader appreciation (Bizzoni et al., 2021, 2022b). Systems to distinguish between different emotions have also been applied to study narratives (Somasundaran et al., 2020) and the aesthetics of literary works (Haider et al., 2020). Maharjan et al. (2018) modelled the "flow of emotions" in literary texts using the NRC lexicon, showing that the shape of emotion-specific arcs had an effect for predicting whether books were successful (based on their GoodReads rating). The distribution of emotions seemed particularly telling for the "success" of a work, as Maharjan et al. (2018) found emotion concentration and variation (std. deviation) higher for successful than for unsuccessful works. As it has been shown that emotion distribution and levels may vary across genres (Mohammad, 2011) – attesting also to the potential of going beyond simple valence annotation – it is particularly interesting for us continue this line of assessing the importance of the shapes of emotion-specific arcs on quality perception.

3 Dataset Construction

3.1 Selecting and Curating Novels

Our data comes from the "Chicago Corpus". This corpus consists of 9,089 novels published in the US between 1880 and 2000, making it an unusual collection for both size and modernity, as it contains both more and more recent novels than the works available on most other platforms[2]. The corpus was compiled based on the number of li-

[1] As opposed to the study of extra-textual features contributing to the perception of quality (Verdaasdonk, 1983; Lassen et al., 2022)

[2] On average, studies on literary quality and success tend to rely on collections of tens to hundreds of novels (Ashok et al., 2013).

braries worldwide that hold each novel, with a preference for more circulated works, and features works by Nobel laureates (i.e., Ernest Hemingway, Tony Morrison), widely popular works, and "genre literature", from Mystery to Science Fiction (e.g., from Agatha Christie to Philip K. Dick) (Long and Roland, 2016).[3] The use of more commonly available or "popular" books also means that the novels are more likely to be reviewed on tertiary platforms such as GoodReads, allowing us to better examine correlations between public reception of novels and their affective content.

All in all the dataset consists of 1,108,108,457 tokens, ranging from 246 tokens to 723,804 tokens per book with an average of 121,918 tokens per book. For parts of our analysis, we split the books into bins each containing 500 tokens, which means there are on average 244 bins per book. We chose a 500-bin size for both practical and theoretical reasons. Multiple studies have shown that using bin sizes as small as 200-300 tokens can beat state-of-the-art machine learning models in accuracy (Teodorescu and Mohammad, 2022, 2023; Öhman, 2021; Öhman and Rossi, 2023) therefore we did not want to use bin sizes of fewer than 300 tokens. Using too large bin sizes, on the other hand, could mean that we would miss out on shorter emotion arcs so we determined 500 to be suitable in order to strike a balance between theory, interpretability, and practice as that would roughly correspond to text subsets that are 1-2 paragraphs in length. Note that the token count will be much higher than the word count of the same text. This is especially true for literary texts which tend to have dialogue marked by quotation marks, em dashes, and more punctuation marks all of which count as individual tokens.

3.2 Affective Word Embeddings

We utilize the NRC Affect Intensity Lexicon (Mohammad, 2018b) as our base emotion labels. We chose this lexicon since it is the most extensive emotion intensity lexicon we are aware of and both it and its sister lexicon (EmoLex) (Mohammad and Turney, 2013) have been used in countless emotion detection tasks successfully. It contains 9,829 entries with at least one emotion association and a value between 0 and 1 for each emotion to represent the intensity of the labeled emotion. The emotions included are *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust*. To increase the accuracy of the lexicon for literary texts, we used the novels in our dataset to create a semantic vector space model with Word2Vec (Mikolov et al., 2018) and then with the aid of cosine similarity measures expanded the lexicon and made it more domain-specific by evaluating frequency distributions of words that were not in the lexicon, but had a high cosine similarity with words that were. These additions were manually checked for accuracy.

As there has been some criticism of using cosine similarity for similarity measures of high-frequency words (Zhou et al., 2022), we also conducted manual evaluations of the newly added terms to ensure the appropriateness of the modifications. The lexicon was checked for unsubstantiated emotion associations and the lemmas in the novels for words that have an emotion association but were not in the lexicon. Both processes are iterative and were repeated three times.

Following this procedure, we created intensity measures for the whole text/novel as well as for each 500-token bin. For the former, the results were normalized by word count (Fig. 1); for the latter, the results were simply sums of the word-emotion association intensities. These intensity calculations are available publicly[4].

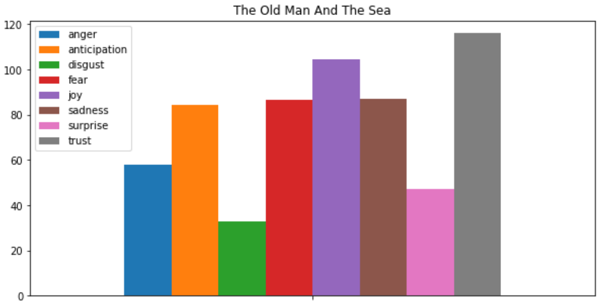


Figure 1: Emotion intensities for Hemingway's *The Old Man and The Sea*. For instance, the prevalence of trust might mirror the Santiago-Manolin relation and be a proxy for the protagonist's endurance and confidence in his abilities at sea.

4 Agreement and Validation

Typically sentiment analysis is considered a supervised classification task, and therefore the most common validation methods tend to use f1 scores

[3] Other quantitative studies are based on this corpus (Underwood et al., 2018; Cheng, 2020), which can be viewed at https://textual-optics-lab.uchicago.edu/us_novel_corpus.

[4] GITHUB link

for accuracy measurements of predicted labels versus actual labels.When sentiment analysis is not a classification task, such as when using simple emotion association, i.e., whether an emotion is present or not, it can be possible to use similar evaluation metrics for lexicon-based methods as well, however, even in such cases one is usually limited to comparing lexicon-based emotion scores with manual annotations of the target text. Humans are also notoriously bad at rating scales (Kiritchenko and Mohammad, 2016) and thus this approach is already problematic and unlikely to offer any true indication of the comparability of evaluations. The incompatibility of the traditional evaluation methods only increases when using emotion intensity over mere presence as this amplifies the difficulty of rating on a scale.

Nonetheless, evaluation is a crucial part of any quantitative methodology and is required for the demonstration of the validity of the results. We, therefore, examine a selection of novels by comparing the emotion arcs with human judgment to validate the accuracy of our methods (Jockers, 2016; Öhman and Rossi, 2023).

4.1 Human Validation

As the approach used in this project does not allow for traditional accuracy measures as typically used in machine learning (Öhman, 2021), we focus our validation efforts on comparing human interpretations with those generated by our lexicon-based model, which has shown to be accurate in multiple prior studies (Teodorescu and Mohammad, 2022; Öhman and Rossi, 2022; Koljonen et al., 2022). One example of a manual annotation of the correspondence of our emotion arcs with narrative events is shown in Figure 3. Note that while at first sight, the co-occurrence of peaks in fear and joy (especially from chunk 80) may appear puzzling, it actually illustrates an important aspect of Hemingway's style in describing complex emotions and reflects the themes of the story overall: in moments of crisis and violence, Hemingway's protagonist still reflects on the natural beauty and his love for the sea. This creates a mix of complex feelings in key scenes, love and hatred, fear and admiration, so that intensities in these feelings co-occur (see, e.g., box 7 in Fig. 3), which is also a token of the protagonist's character: his endurance and optimistic outlook on life. The slope and generally high levels of trust in the story also follow the progression of narrative events (see, e.g., box 5 in Fig. 3).

4.2 Agreement in Emotions

Certain emotions are more likely to co-occur than others. This can lead to lower accuracy scores in multilabel machine learning models when the features of correlated emotions are muddled, but increased detail in lexicon-based models when we can differentiate better between closely related associations. Figure 2 shows the correlation of emotions in the entire 'EmotionArcs' corpus. The negative emotions *anger*, *disgust*, *fear*, and *sadness* show a high rate of co-occurence as expected, while *joy* is negatively correlated with both *anger* and *fear* and positively correlated with *anticipation* and *trust*.

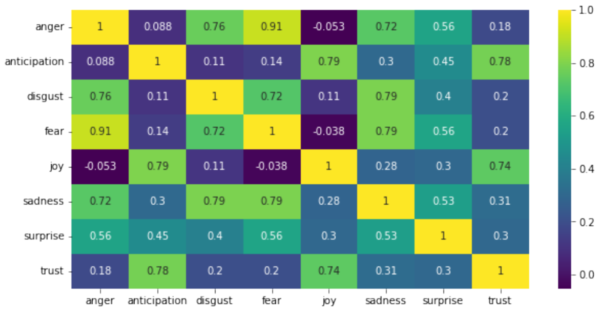


Figure 2: Correlation between emotions in all emotion arcs

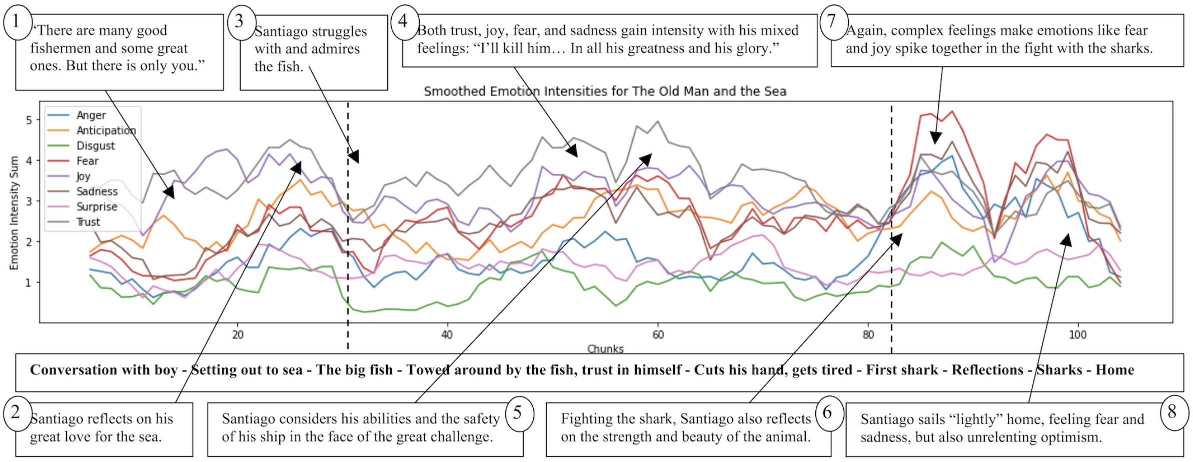
5 Quality Proxies

5.1 Rationale

The idea that the distribution and dynamics of the emotions expressed in a text are related to the reception of that text is widespread, and several studies have used both sentiment analysis and emotion detection to capture meaningful aspects of the reading experience (Drobot, 2013; Cambria et al., 2017; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017; Öhman and Rossi, 2023). In this work, we tried several different resources that approximate the reception of a novel – specifically, its perceived overall quality – by either a large number of lay readers (crowd-based resources) or a small number of expert readers (expert-based resources).

5.2 Expert-based and crowd-based resources

Expert-based judgments of literary works originate from a limited group of expert readers, such as editors, publishers (Karlyn and Keymer; Vulture, 2018), individual literary scholars (Bloom, 1995),



"There are many good fishermen and some great ones. But there is only you."

Santiago struggles with and admires the fish.

Both trust, joy, fear, and sadness gain intensity with his mixed feelings: "I'll kill him… In all his greatness and his glory."

Again, complex feelings make emotions like fear and joy spike together in the fight with the sharks.

Conversation with boy - Setting out to sea - The big fish - Towed around by the fish, trust in himself - Cuts his hand, gets tired - First shark - Reflections - Sharks - Home

Santiago reflects on his great love for the sea.

Santiago considers his abilities and the safety of his ship in the face of the great challenge.

Fighting the shark, Santiago also reflects on the strength and beauty of the animal.

Santiago sails "lightly" home, feeling fear and sadness, but also unrelenting optimism.

Figure 3: Arcs of *The Old Man and the Sea* annotated for narrative events.

and award committees like the Nobel Prize. Crowd-based judgments, on the other hand, are formed by a large number of readers without a given literary expertise, and offer more inclusivity and statistical robustness, because of the scale. GoodReads, a social readership platform with over 90 million users, provides insight into such crowd-based judgments (Maharjan et al., 2017; Bizzoni et al., 2021; Jannatus Saba et al., 2021; Porter, 2018) and especially into reading culture "in the wild" (Nakamura, 2013), as it catalogs books from different genres and derives ratings from a heterogeneous pool of readers (Kousha et al., 2017). There are various issues with using GoodReads' ratings as a metric, among others, how this heterogeneity is conflated into one single score (0-5) that takes no account of differential rating behavior, for example across genres. Beyond the rating or "stars" on GoodReads, another option is to use the rating count itself as a proxy of quality perception, supposing that more frequently rated titles are also more popular and liked. There are also less clear-cut, more nuanced measures of literary reception. For example, a conceptually hybrid measure between crowd- and expert-based is the number of libraries holding a given title worldwide, as indicated on Worldcat (Bennett et al., 2003). Expert choice and user-demand may influence what titles are acquired by libraries, and since the libraries are many, the compound nature of all title-selection approximates crowd-based judgment.

In this work, we selected the latter two proxies: for each book we collected the number of ratings of GoodReads (as of December 2022) and the number of libraries holding the title. In our corpus, library holdings and rating count are correlated with a coefficient of 0.50 ($p<0.01$) using a simple Spearman correlation.

6 Data Analysis

6.1 Emotion Distribution

To examine the association between the emotional content of novels and their perceived quality, we examined the **overall intensity** of the eight emotions in each novel. As noted, intensity values were length-normalized to ensure comparability across texts of different sizes. To understand the variation in emotions in each novel, we computed the **entropy** of their emotion intensity distribution. In our context, the concept of entropy serves as a measure of the uncertainty of emotional intensities in novels: a low entropy value indicates that one emotion may dominate the text, being reliably more intense than other emotions. Conversely, high entropy indicates a more diverse emotional profile, where each emotion is represented with comparable intensity. In Fig. 1 the emotional profile of *The Old Man and the Sea* appears to have medium-high entropy.

6.2 Emotion Trends

To relate the linear shapes of the eight emotion arcs to quality perceptions, we computed the **skewness** and **slope** steepness of each emotion arc, both as a score for each emotion separately and as the average score of all eight emotions per novel. The slope value for each emotion is computed with linear regression and represents the development in intensity of that particular emotion across the narrative: if the joy arc increases or decreases linearly across a novel, the slope of its joy arc will be relatively steep (Su et al., 2012). Skewness captures the symmetry of an emotion arc: an emotion arc having few large values or intensities but many small values is positively skewed, while an arc with an even distribution of large and small values has a skewness approximating 0 (Kokoska and Zwillinger, 2000).

6.3 Overall novel emotion

We correlated overall emotion intensities of the novels with our quality proxies: rating count and library holdings, finding slight correlations of certain features with library holdings (Table 1). Moreover, we also correlated emotion arc shape, slope, skewness and entropy, with the quality proxies. A simple Spearman correlation shows a slight negative value for the relation between library holdings and emotional entropy, as well as slope in sadness, showing no correlation with rating counts (Table 1).

Emotion	Coefficient
Fear (sum)	0.14
Sadness (sum)	0.14
Anger (sum)	0.14
Disgust (sum)	0.13
Anticipation (sum)	0.13
Surprise (sum)	0.13
Joy (sum)	0.12
Entropy (all emotions)	-0.12

Table 1: Emotion features holding a correlation of $>$ .10 with library holdings (Spearman). All correlations are statistically significant ($p < 0.01$).

As illustrated in Figure 4, the traditional metrics for correlation do not suffice in capturing the dynamics between emotion entropy and our chosen quality proxies. The data distribution is markedly non-linear, showing separate clusters with different characteristics. Specifically, we find two distinct groups: (i) Titles with low rating counts and low library holdings that manifest across the entire spectrum of emotion entropy; (ii) Titles with high rating counts and high library holdings that populate a specific subset of the emotional entropy range. To dissect this relationship further, we segment the data into two broad categories based on rating count and library holdings: low rating and library holdings (<100) and high rating and library holdings (>500). While the thresholds 100 and 500 are somewhat arbitrary, they demarcate regions in the data space where observed trends remain largely consistent, lending our approach a level of replicability. The implications of different upper thresholds are shown more exhaustively in the Appendix. With this separation of marginally "successful" and "unsuccessful" groups of titles, the relation between emotion entropy and quality perception is more evident than before: negative correlations of emotion entropy and the quality proxies continue only up to a certain entropy value, before which there is even a positive correlation between entropy and library holdings; and when looking at rating count, the correlation is almost completely positive. In general, it seems that titles with higher entropy of emotions receive a higher number of ratings and, up to a tipping point, are held in more libraries (Figure 5).

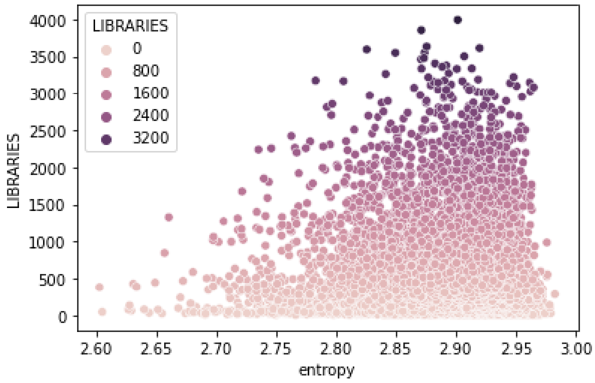


Figure 4: Distribution of library holdings with respect to emotion entropy.

6.4 Emotion arcs

Using the same groupings of high and low rating and library holdings titles, we examined the correlation between our quality proxies and the average slope intensity, as well as the average skewness of arcs, averaging the values across slopes and skewness of each of the eight emotion for each title. This added layer offers an insightful correlation with our quality proxies: titles with pronounced emotional slopes tend to have fewer ratings and less representation in libraries, while, contrarily, a greater emotional skewness correlates positively with both library holdings and rating counts. We similarly correlated the slopes of each emotion in a novel, as we might suppose that titles (or even genres) may exhibit a steep slope for one emotion, but not for another, making the mean unrepresentative. Here, we find that the average patterns represented in Table 2 hold when looking at almost any single emotion: titles above 500 ratings and library holdings correlate negatively with slopes, and the reverse is true for titles below 100 ratings and li-

Variable	rating count	libraries
mean skewness > 500	0.60*	0.50*
mean skewness < 100	-0.55*	-0.41*
mean slope inclination > 500	-0.81**	-0.71*
mean slope inclination < 100	0.83**	0.69*
mean entropy > 500	0.76*	0.29
mean entropy < 100	-0.69*	-0.63*

Table 2: Correlations of emotion arc features with quality proxies (Spearman correlation). *p < 0.05, **p < 0.01

	Joy	Anger	Sadness	Fear	Disgust	Surprise	Trust	Ant.
Rating count >500	-0.656**	-0.861**	-0.560*	-0.694**	-0.686**	-0.809**	-0.764**	-0.667**
Rating count <100	0.652**	0.886**	0.776**	0.721**	0.772**	0.765**	0.737**	0.589 **
Holdings >500	-0.938**	-0.953**	-0.913**	-0.885**	-0.875**	-0.835**	-0.839**	-0.794**
Holdings <100	0.935**	0.930**	0.749**	0.885**	0.782**	0.617*	0.757**	0.725**
Rating count >500	0.272*	0.068	0.288**	0.019	0.309**	0.453**	0.548**	-0.774*
Rating count <100	-0.272*	0.020	-0.199*	-0.020	-0.151	-0.516**	-0.550**	0.662*
Holdings >500	0.035	0.136	0.347**	0.247*	0.308**	0.427**	0.332**	0.92*
Holdings <100	0.047	-0.188*	-0.324**	-0.138	-0.233**	-0.527**	-0.477**	0.93*

Table 3: Correlation of the emotion arcs' slopes (rows 1-4) and their skewness (rows 5-8) with Rating Count and libraries' holdings for both >500 and <100 values. Asterisks denote the significance of the p-value: * p<0.05, ** p<0.01.

brary holdings, while the opposite appears true for skewness (Table 3).

7 Concluding Discussion

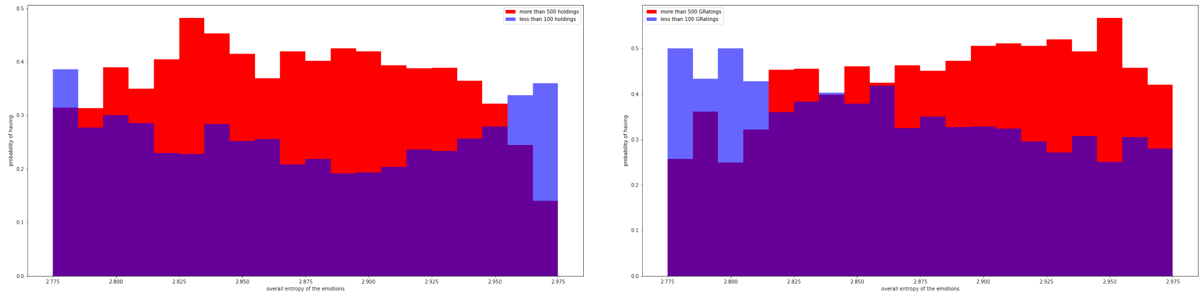
With 'EmotionArcs' we have presented a new resource for the study of emotions in literary novels that we hope will enable many other researchers to investigate how affect in literary works is intertwined with other aspects of literature. We have shown that our method produces reliable, useful, and easily interpretable emotion arcs that can help more traditional literary scholars compare larger corpora of literary works that are possible using only qualitative methods. It seems that overall emotional entropy, the slopes of emotion arcs and their level of skewness hold some relation with the reception of the novels as measured via rating count and library holdings.

1. **Entropy**. A novel with higher emotional entropy will have an overall higher probability of being rated more than five hundred times on GoodReads. The same holds for its likelihood of being held in a large number of libraries – up to a point: "too much entropy" is related to lower circulation in libraries.

2. **Slope**. A novel with steeper overall emotion arcs will have overall a lower probability of being rated more than five hundred times on GoodReads or being held by more than five hundred libraries; conversely, it will have an increased probability of being rated less than 100 times and held by less than 100 libraries.

3. **Skewness**. A novel with a low level of overall emotion skewness will have overall a lower probability of being rated more than five hundred times on GoodReads or being held in more than five hundred libraries; conversely it will have a increased probability of being rated less than 100 times and being held by less than 100 libraries.

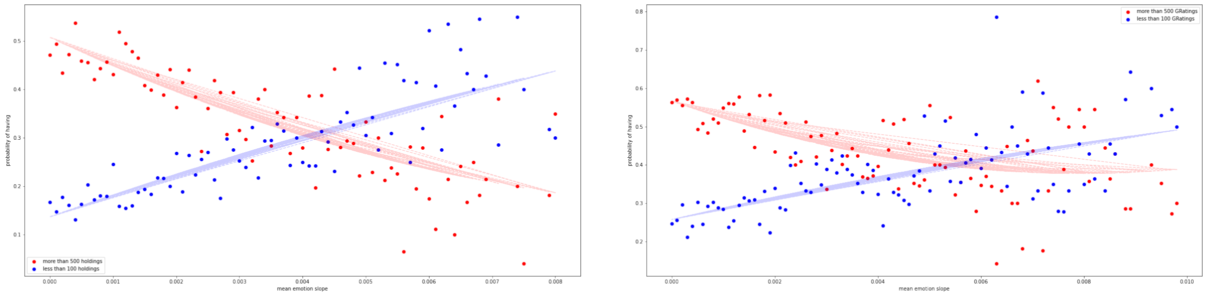
Our results on entropy might bear a relation to with Jautze et al. (2016) about topics: novels with relatively few, dominating topics are perceived as being less good than novels that use a larger topical palette. It is possible there is a similar effect at the level of the emotions represented in a text. It's important to remember that we are talking about fine-grained emotions: a novel with a high level of fear does not necessarily correspond to a narrative where characters are constantly scared. Rather, because of its selection of certain events, a text may be more likely to sample from an emotional vocabulary of fear than from that of another emotion. Something similar might be inferred from the slopes' steepness and skewness: excessively predictable and smooth emotion arcs might not create as effective a reader experience. This interpretation is corroborated by studies that have found that readers tend to prefer fractal story arcs but only with moderate level of coherence (Hu et al., 2021; Bizzoni et al., 2021). Story arcs that



(a) Titles below 100 or above 500 holdings.

(b) Titles below 100 or above 500 ratings.

Figure 5: Probability of having high/low number of library holdings or goodreads ratings (below 100/above 500) at different values of emotional entropy. All probabilities were computed on populations of at least 10 different titles. The relation with the number of libraries' holdings might point to a "sweet spot".



(a) Titles below 100 or above 500 holdings.

(b) Titles below 100 or above 500 ratings.

Figure 6: Probability of having high/low library holdings or high/low number of GoodReads' ratings (below 100/above 500) at different levels of average slope steepness. All probabilities were computed on populations of at least 10 different titles.

monotonically focus on one emotion or have a very steep slope will either be overly predictable, and by extension overly coherent, or, at some point, too unpredictable and locally incoherent.

Finally, in addition to a novel resource, the methods used in this study offer simple and robust guidelines that should be a part of any lexicon-based emotion projects. We strongly believe our methodology of fine-tuning existing lexicons to be more domain- and period-specific with the help of affective word embeddings should be the first step in any sentiment analysis or emotion detection task that utilizes lexicons as it not only makes the lexicons more attuned to the specific domain at hand but also increases precision and recall in general and can even negate some of the effects of semantic shifts in language. Furthermore, exactly because the method is simplistic, even non-computational experts can easily replicate the steps for their own data. Although our focus is on the relationship between emotional complexity and literary reception, the applications of empirical findings on literary quality and novel reception are manifold; from writer-oriented interfaces (both for professional and lay writers) to systems designed to help editors. A new possible area of application is also that of creating better/more sophisticated fiction-writing LLMs.

In the future, we aim to continue with similar projects further improving and enhancing the lexicon and extending the use cases of emotion arcs to, e.g., the exploration of narrative structure and differences in affective language used by individual authors and across genres. We also aim at experimenting with different proxies for perceived literary quality and reception, including more expert-based resources such as canon lists and prestigious awards. Finally, we intend to combine our emotional arcs with more sophisticated modeling techniques for fractal analysis and time series forecasting in order to have a more complex view of the relation between the textual representation of emotions and overall reader experience.

Limitations

As emotion annotation is a notoriously difficult task, this study has attempted to make the process as robust as possible, regardless, emotions are always subjective and difficult to measure. Emotions are also partly constructed by the measuring process itself and therefore always a reflection of the methods used (Laaksonen et al., 2023). Methodologically, the choice of lemmatization, and to a lesser extent other preprocessing steps, affects how the semantic vector space is constructed and how words match the affective space. Although English is a comparatively easy language to lemmatize, there were instances of lexemes in the data that could have been further broken down.

Word embeddings are inherently contextual, however, they are not immune to polysemy, particularly when used with a hybrid lexicon-based approach. We reduced the effect of polysemous words and other similar artifacts with our iterative approach, however, it is unlikely we were completely able to remove the effects of semantic shifts or cultural biases that occur in language and stem from the original annotations of the NRC lexicon as well as the diverse nature of the data. Ultimately, unlimited iterations are possible, and we made a balanced choice between feasibility, time, cost, and practicality.

One important limitation of our corpus of novels is its strong Anglophone and American tilt: there are few non-American and non-Anglophone authors, which inevitably situates the entire analysis within the context of an "Anglocentric" literary field.

Regarding the proxies of reader appreciation used in this study, it is hard to control the demographics of each proxy for literary quality and reception. Generally, sources like GoodReads are more diverse and represent a more comprehensive demographic selection than awards committees or anthologies' editorial boards. Yet it should be noted that the majority of GoodReads users from the beginning of GoodReads in 2007 were anglophone. The number of library holdings as a proxy reflects a complex interaction of user demand and expert choice, where demographics is difficult to gauge.

It is also likely that there is a correlation between reviews on GoodReads and quality, but as with any proxy measurement, it is difficult to concretely distinguish popularity, success, and quality.

Ethics Statement

We strongly believe in reproducible and replicable science and are therefore making all data and code freely available where possible. We adhere to best practice guidelines in both the creation and publication of the datasets as suggested by Gebru et al. (2021) and Mohammad (2022). We have assessed the lexicon's suitability for the task at hand and tried to mitigate any inherent biases with our lexicon-enhancement process, however, it is possible we have missed some details and welcome feedback.

References

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

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.

Mieke Bal and Christine Van Boheemen. 2009. *Narratology: Introduction to the theory of narrative*. University of Toronto Press.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Rick Bennett, Brian F Lavoie, and Edward T O'Neill. 2003. The concept of a work in worldcat: an application of frbr. *Library collections, acquisitions, and technical services*, 27(1):45–59.

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.

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).

Harold Bloom. 1995. *The Western Canon: The Books and School of the Ages*, first riverhead edition edition. Riverhead Books, New York, NY.

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.

Erik Cambria, Andrew Livingstone, and Amir Hussain. 2012. The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer.

Nick Campbell. 2004. Perception of affect in speech-towards an automatic processing of paralinguistic information in spoken conversation. In *Eighth International Conference on Spoken Language Processing*.

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

Alan S Cowen and Dacher Keltner. 2021. Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences*, 25(2):124–136.

Nan Z Da. 2019. The computational case against computational literary studies. *Critical inquiry*, 45(3):601–639.

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

Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. 2020. Po-emo: Conceptualization, annotation, and modeling of aesthetic emotions in german and english poetry. *arXiv preprint arXiv:2003.07723*.

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.

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

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

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

Kim Jautze, Andreas van Cranenburgh, and Corina Koolen. 2016. Topic Modeling Literary Quality. In *Digital Humanities 2016: Conference Abstracts.*, pages 233–237, Kraków.

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

Matthew L Jockers. 2016. More Syuzhet validation. *Matthew L. Jockers blog*.

Danny Karlyn and Tom Keymer. Chadwyck-Healey Literature Collection.

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

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the Relationship between Literary Genres and Emotional Plot Development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California. Association for Computational Linguistics.

Stephen Kokoska and Daniel Zwillinger. 2000. *CRC standard probability and statistics tables and formulae*. Crc Press.

Juha Koljonen, Emily Öhman, Pertti Ahonen, and Mikko Mattila. 2022. Strategic sentiments and emotions in post-second world war party manifestos in finland. *Journal of computational social science*, pages 1–26.

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*, 79:101439.

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.

Salla-Maaria Laaksonen, Juho Pääkkönen, and Emily Öhman. 2023. From hate speech recognition to happiness indexing: critical issues in datafication of emotion in text mining. In *Handbook of Critical Studies of Artificial Intelligence*. Edward Elgar.

Ida Marie Schytt Lassen, Yuri Bizzoni, Telam Peura, Mads Rosendahl Thomsen, and Kristoffer Laigaard Nielbo. 2022. Reviewer preferences and gender disparities in aesthetic judgments. *arXiv preprint arXiv:2206.08697*.

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*.

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

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.

Raymond A Mar, Keith Oatley, Maja Djikic, and Justin Mullin. 2011. Emotion and narrative fiction: Interactive influences before, during, and after reading. *Cognition & emotion*, 25(5):818–833.

Tomáš Mikolov, Édouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Saif Mohammad. 2011. From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *WASSA@ NAACL-HLT*, pages 174–179.

Saif M. Mohammad. 2018a. 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. 2018b. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

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

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.

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

Emily Öhman. 2021. The Validity of Lexicon-based Sentiment Analysis in Interdisciplinary Research. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 7–12, NIT Silchar, India. NLP Association of India (NLPAI).

Emily Öhman and Riikka Rossi. 2022. Computational Exploration of the Origin of Mood in Literary Texts. *Natural Language Processing for Digital Humanities@Asian Association for Computational Linguistics*, page 8.

Emily Öhman and Riikka Rossi. 2023. Affect as Proxy for Mood. *Journal of Data Mining and Digital Humanities*, Special Issue: Natural Language Processing for Digital Humanities.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.

J.D. Porter. 2018. Popularity/Prestige: A New Canon.

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.

Swapna Somasundaran, Xianyang Chen, and Michael Flor. 2020. Emotion arcs of student narratives. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 97–107, Online. Association for Computational Linguistics.

Xiaogang Su, Xin Yan, and Chih-Ling Tsai. 2012. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3):275–294.

Annie Swafford. 2015. Problems with the syuzhet package. *Anglophile in Academia: Annie Swafford's Blog*.

Daniela Teodorescu and Saif M Mohammad. 2022. Frustratingly easy sentiment analysis of text streams: Generating high-quality emotion arcs using emotion lexicons. *arXiv preprint arXiv:2210.07381*.

Daniela Teodorescu and Saif M Mohammad. 2023. Generating high-quality emotion arcs for low-resource languages using emotion lexicons. *arXiv preprint arXiv:2306.02213*.

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

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*.

Hugo Verdaasdonk. 1983. Social and economic factors in the attribution of literary quality. *Poetics*, 12(4-5):383–395.

editors Vulture. 2018. A Premature Attempt at the 21st Century Literary Canon.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–423, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

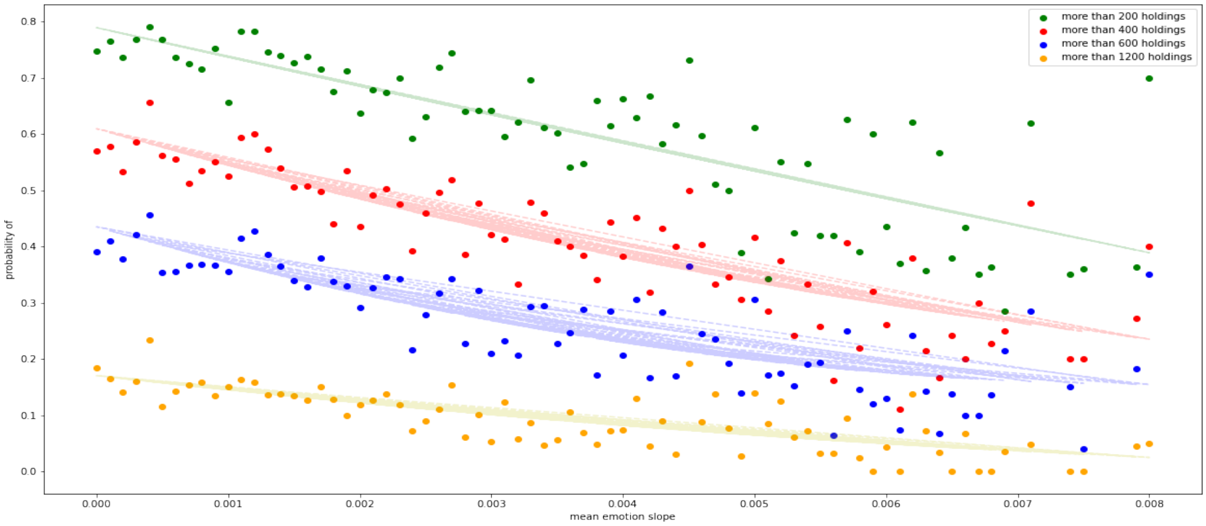


Figure 7: Trends in probability of being in the high- or low-rating group at different cutting points of emotion slope value. While 100 and 500 rating count and library holdings are somewhat arbitrary thresholds, trends in our data are reproduced at different cutoff points.

Pair	Coefficient	Type of Correlation
Anger, Fear	0.90	Strong Positive
Anticipation, Joy	0.77	Strong Positive
Disgust, Anger	0.77	Strong Positive
Disgust, Sadness	0.78	Strong Positive
Fear, Sadness	0.78	Strong Positive
Anticipation, Trust	0.76	Strong Positive
Joy, Trust	0.71	Strong Positive
Anger, Entropy	0.63	Moderate Positive
Entropy, Joy	-0.53	Moderate Negative
Entropy, Trust	-0.51	Moderate Negative

Table 4: Pairwise correlation of emotions

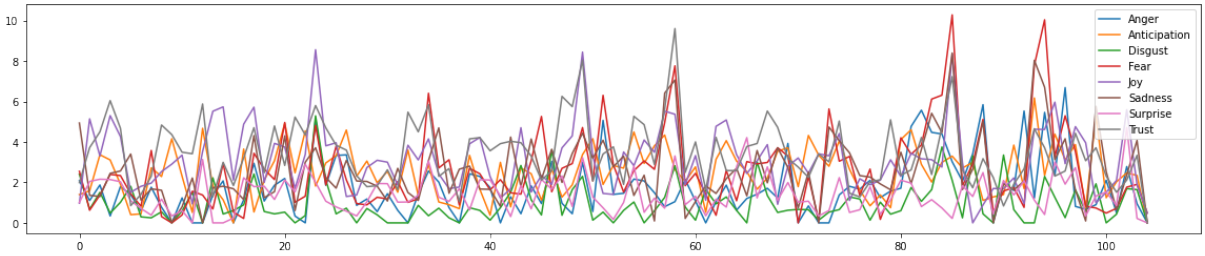


Figure 8: Unsmoothed emotion arcs for The Old Man and the Sea

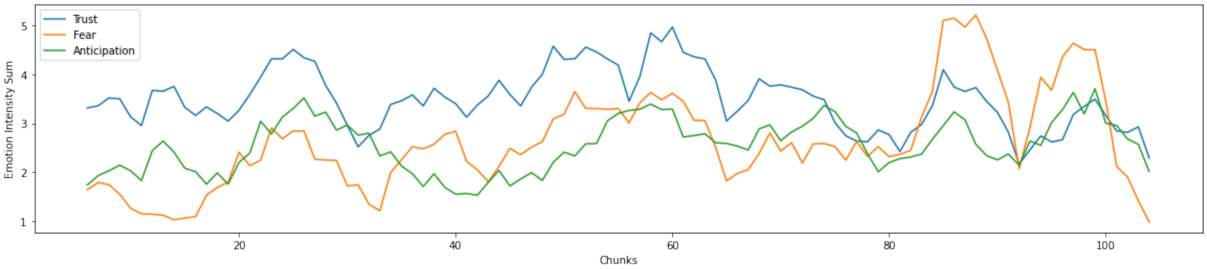


Figure 9: Smoothed arcs for trust, fear, and anticipation for The Old Man and the Sea

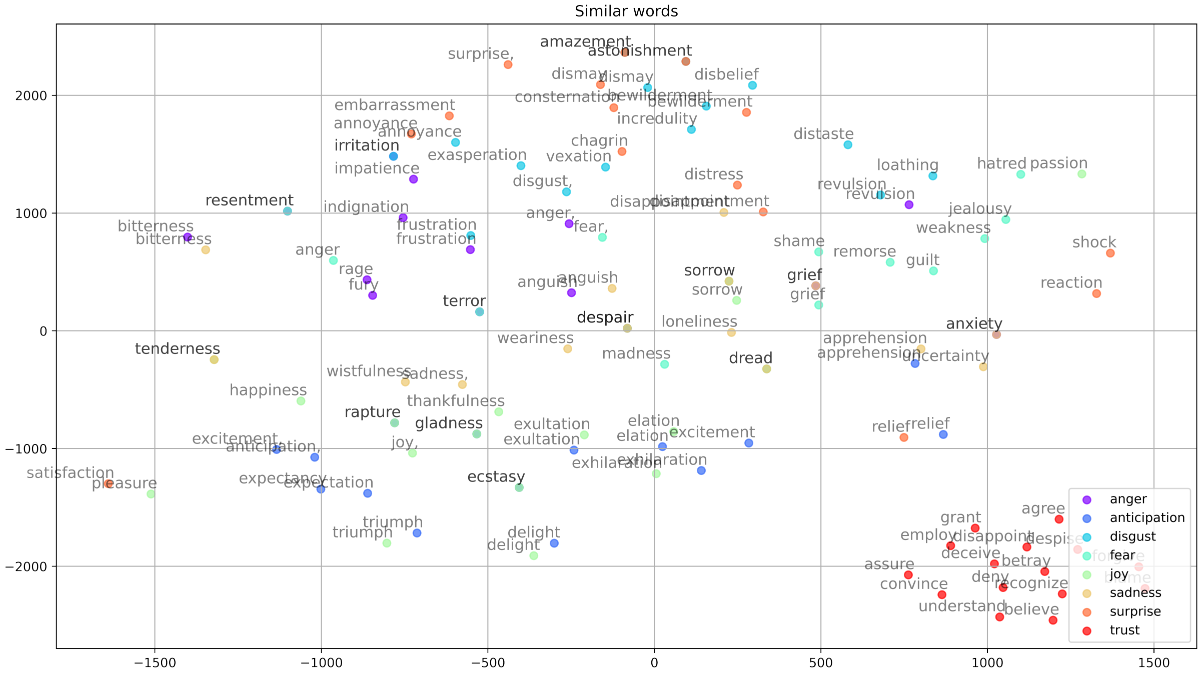


Figure 10: Word similarities for Plutchik's core emotions in the corpus in the affective semantic vector space as measured by cosine similarity. We can see that *trust*, although commonly co-occurring with both *joy* and *anticipation* does not overlap with these emotions. On the other hand, the negative emotions both overlap and co-occur.