EmotionArcs: Emotion Arcs for 9,000 Literary Texts

Emily Öhman[†]
School of International Liberal Studies
Waseda University, Japan
ohman@waseda.jp

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

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

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

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. We evaluate emotion arcs manually by comparing them to human annotation and against other similar emotion modeling systems to show that our system produces coherent emotion arcs that correspond to human interpretation. We present and make this resource available for further studies of a large collection of emotion arcs and present one application, exploring these arcs for modeling reader appreciation. Using 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

Sentiment analysis and emotion detection are subjective in nature, 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, 2021b) and how the task of emotion detection to some degree constructs the phenomena it is trying to measure (Laaksonen et al., 2023). 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, the emotions touched upon in texts – in both explicit *and* evocative ways – determine essential aspects of the reader's experience at the structural and stylistic level (Mar et al., 2011). However, while this relation between emotions in literary texts and reader experience can seem relatively intuitive, it needs to be more obvious to test or quantify. This presents us with a few difficulties. The first difficulty is the modeling of "emotions in the text" – defining what we mean by that, deciding which emotions to define, and how to measure the emotional content of any given textual unit - word, phrase, sentence, or paragraph. Due to the complexity of human readers' interpretations and experiences of texts, this is a difficult task to model. The second difficulty is quantifying the relation between emotions in text and their reception or perceived quality of a literary narrative. In this paper, we introduce a new resource, 'EmotionArcs', to explore the relationship between these 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 impacts a novel's literary quality, measured both through library holding numbers and GoodReads ratings.

2 Related Work

Computational literary studies (CLS) is an active field of research affiliated with Digital Humanities

[†]These authors contributed equally to this work

Proceedings of LaTeCH-CLfL 2024, pages 51–66

and applied Natural Language Processing. Sentiments, emotions, and affect are all common research topics within CLS and include work in emotion classification, genre classification, story-type clustering, sentiment tracking, and character analysis (Kim and Klinger, 2018).

2.1 Emotion Analysis

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 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), narrative mood (Öhman and Rossi, 2023), and reader preferences and perceived quality (Bizzoni et al., 2022a). 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).

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, some recent papers have shown that lexicon-based methods can produce accuracies comparable to machine learning and transformer-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 (Teodorescu and Mohammad, 2023; Elkins, 2022; Öhman, 2021b).

2.2 Literary Quality

Studies that aim to forecast the perception of literary quality by relying on textual features[1] have mostly depended on stylistic features. This includes factors like sentence length and readability (Maharjan et al., 2017; Bizzoni et al., 2023a), the proportion of different classes of words (Koolen et al., 2020; Bizzoni et al., 2023c), and the frequency of word pairs (n-grams) (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 model as 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), but without taking into account essential aspects like plot variability or the progression of the narrative. Once the arcs are computed, it is possible to cluster them based on similarities (Reagan et al., 2016). For example, a simple sentiment arc clustering approach by identified six fundamental narrative arcs that they speculated might form the basis of narrative construction. 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 arcsrelated to reader appreciation (Bizzoni et al., 2021, 2022b, 2023b). 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) modeled the "flow of emotions" in literary texts using the NRC lexicon, showing that the shape of emotion-specific arcs had an effect on predicting whether books were successful (based on GoodReads ratings). The distribution of emotions seemed particularly telling for the "success" of a work, as Maharjan et al. (2018) found emotion intensity 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), it is particularly interesting for us to continue this line of assessing the importance of the shapes of emotion-specific arcs on quality perception, in our case examining novels only.

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

[1] In contrast to the study of *extra-textual* features (Verdaasdonk, 1983; Lassen et al., 2022)

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 libraries holding numbers worldwide, with a preference for more circulated works. It 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 reviewed on tertiary platforms such as GoodReads, which facilitates the examination of correlations between public reception and novels' affective content. 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 of just 200-300 tokens can beat state-of-the-art machine learning models in accuracy (Teodorescu and Mohammad, 2023; **?**; Öhman and Rossi, 2023) Using too large bin sizes, on the other hand, could misrepresent and muddle the emotion arcs. We determined 500 tokens, roughly corresponding to text subsets that are 1-2 paragraphs in length, to be suitable in order to strike a balance between theory, interpretability, and practice. 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 with quotation marks, 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) for emotion labels as it is the most extensive emotion intensity lexicon we are aware of. Moreover, both it and its sister lexicon EmoLex (Mohammad and Turney, 2013) have been used in countless emotion detection tasks and have proven their accuracy and usability in a variety of tasks. This lexicon was created with the help of human annotators using best-worst scaling. It contains 9,829 lexemes 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*. As this lexicon is not specific to the domain of 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 to make it more domain-specific. Cosine similarity is a commonly used measurement to determine the similarity between two objects, in this case, lexemes, represented as vectors. For vectors a and b we can represent cosine similarity as follows: $\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{ab}}{\|\mathbf{a}\|\|\mathbf{b}\|}$ 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. Following this procedure, we created *emotion intensities* for the whole novels (e.g., see Fig. 1) as well as for each 500-token bin. For the former, the results were normalized by word count; for the latter, the results were simply sums of the word-emotion association intensities. These intensity calculations are available publicly[4].

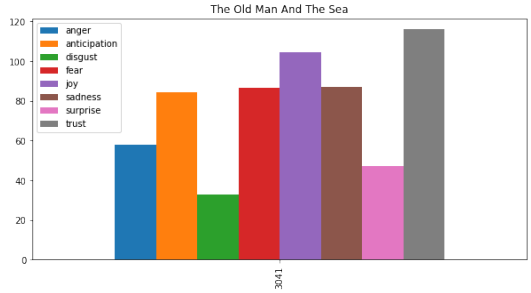


Figure 1: Emotion intensities for Hemingway's *The Old Man and The Sea*. For instance, the prevalence of trust might mirror the Santiago-Manolin relation and be a proxy for the protagonist's endurance.

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

[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]https://github.com/yuri-bizzoni/EmoArc

4 Agreement and Validation

As the approach used in this project does not allow for traditional accuracy measures often used in machine learning (Öhman, 2021b), 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, 2023; Öhman and Rossi, 2023; Koljonen et al., 2022). We validated the EmotionArcs resource in three different ways:

(i) Two literary scholars inspected the emotion arcs of select novels, relating style and narrative events to the shapes of emotion arcs. One example of a manual annotation of the correspondence of our emotion arcs with narrative events is shown in Figure 4 (see another in the Appendix). Note that while at first sight, the co-occurrence of peaks in fear and joy (especially from chunk 80 on) may appear puzzling, it 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. 4), which is also a token of the protagonist's 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. 4).

(ii) We randomly selected 11 passages from one novel: *The Old Man and the Sea*, asking 20 non-expert volunteer annotators to indicate, for each passage, which emotions were present from a predefined set and at what intensity on a 0-1 scale (for the agreement between model and annotator scores, see the Appendix). All passages received 3 to 6 annotations. After a first independent round, annotators were provided with the EmotionArcs scores for the same passage and asked whether they thought the model scores were present in the text, and whether they should be lower or higher (Moreira et al., 2023; Bizzoni and Feldkamp, 2023) Our annotators vastly agreed with the model's categorical choice (see Table 1), while agreement on their intensity varied. In 225 over 232 cases, annotators assessed that the model chose the correct emotions for the text. In 122 of these cases, the annotators also agreed with the intensity score assigned. Of the remaining 110 cases, 43 were given the assessment "could be higher" and 60 the assessment "could be lower".

	Agree			**Disagree**
	Higher	Lower	Correct	
Count	43	60	122	7
Total	225			7

Table 1: Annotators' agreement with EmotionArc's scores

Joy was the emotion that elicited most "could be lower" responses. We believe this is because in Plutchik's eight core emotions (Plutchik, 1980), *joy* is the only genuinely positive one.

(iii) Lastly, two novels were selected for close-reading evaluation. We evaluated the EmotionArcs by comparing their scores to valence scores produced by an independent (RoBERTa, fine-tuned for sentiment analysis on tweets (Barbieri et al., 2022)[5]) and average human annotations for valence of the same books.[6] As emotion annotation has been shown to correlate with valence scores – most notably joy and fear with positive and negative valence (Moreira et al., 2023) – we combined the emotion scores of joy and fear of our method to model arcs of novels, comparing them against the SA and human annotation of the same novel. An example can be seen in Figure 2, where human evaluation closely follows that of our model's *joy* values minus *fear* value as well as that of additional validation produced with the help of RoBERTa scores. In other words, by combining the most prevalent positive emotion and the most prevalent negative emotion, with a positive and negative sign respectively, it's possible to reproduce a novel's sentiment trendline [7]

[5]Note that to convert RoBERTa's categorical output we used the confidence score of labels as a proxy for sentiment intensity. If the model classifies a sentence as *positive* with a confidence of, for example, 0.89, we interpret it as a valence score of +0.89, and so on. Scores of the *neutral* category were converted to a score of 0.0. For further details SA with Transformers, see Bizzoni and Feldkamp (2023).

[6]Human annotators (n=2) read from beginning to end and scored sentences on a 1 to 10 valence scale.

[7]We focus our comparison on *joy* and *fear* as they are among the most frequent in text and we see them as the purest representatives of unambiguous valence in the available categories and highly representative of overall valence due to the overall overlap of emotions with *fear* and *joy* (Bizzoni and Feldkamp, 2023; Öhman, 2020a,b).

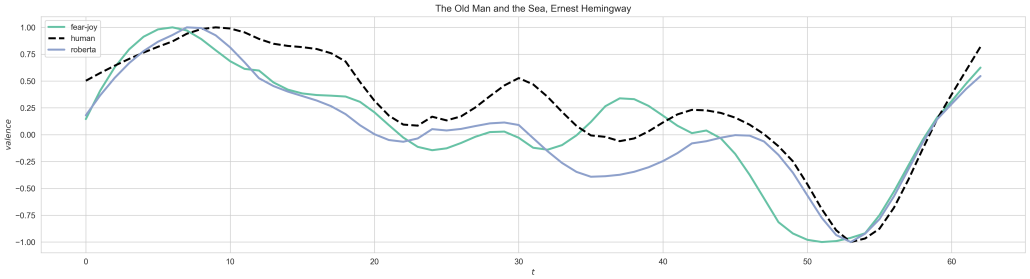


Figure 2: *The Old Man and the Sea*, manual evaluations, EmotionArcs (fear minus joy), and RoBERTa. Arcs were smoothed using adaptive filtering (Jianbo Gao et al., 2010).

4.1 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 3 shows the correlation of emotions in the entire 'EmotionArcs' corpus. The negative emotions *anger*, *disgust*, *fear*, and *sadness* show a high rate of co-occurrence as expected, while *joy* is negatively correlated with both *anger* and *fear* and positively so with *anticipation* and *trust*. *Anticipation* strongly correlates not only with *joy*, but also with *trust* [8], an emotion of more ambiguous valence. *Anger* correlates significantly also with *surprise*. It stands to reason that a passage expressing *anger* can be framed as sudden, surprising and, even cathartic.[9].

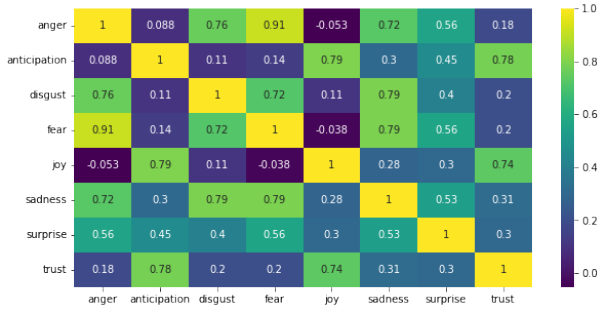


Figure 3: Correlation between emotions in all emotion arcs

[8] *trust* is commonly associated with its negative counterpart *distrust*, which is not a label in Plutchik

[9] Plutchik considers *anger* a positive emotion, counter to how it is used in most NLP models, and it is not immediately clear whether the valence in a literary setting should be reversed from its psychological roots as is standard practice (Plutchik, 1980; Öhman, 2021a).

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 proxies) or a small number of expert readers (expert-based proxies).

5.2 Expert-based and crowd-based proxies

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), 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. 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

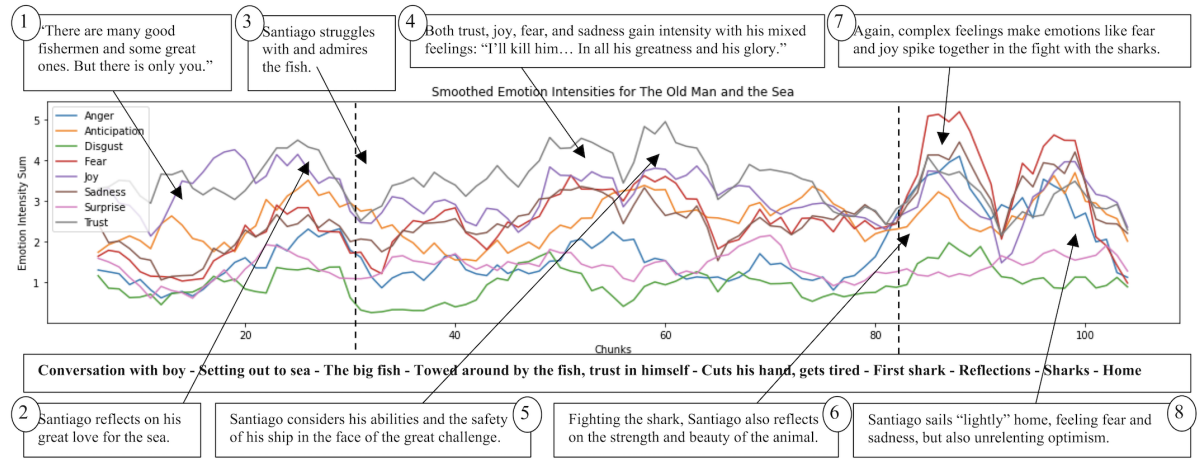


Figure 4: Arcs of *The Old Man and the Sea* annotated for narrative events.

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 selections 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 libraries holdings of the title.[10]

6 Data Analysis

6.1 Emotion Distribution

Building on previous work (Maharjan et al., 2018), we 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

Building on work examining the shape and dynamics of narrative arcs (Bizzoni et al., 2021; Öhman and Rossi, 2023; Moreira et al., 2023), we relate the linear shapes of the eight emotion arcs to quality perceptions, computing the **skewness** and **slope** steepness of each emotion arc; 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 retrieved by 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 arc with 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

The intensity of most emotions appears to hold a correlation with the number of library holdings, but a weak one. There is also a weak negative correlation between library holdings and the overall entropy of the emotional values of a text (Table 2).

Yet it seems that the distribution of the data is unfit for standard correlation, as the relation be-

[10]Note that in our corpus, library holdings and rating count are correlated with a coefficient of 0.50 ($p<0.01$) using a simple Spearman correlation.

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 2: Emotion intensities correlation with library holdings (Spearmann). For all correlations, $p < 0.01$.

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 3: Correlations of emotion arc features with reception proxies (Spearman correlation). *$p < 0.05$, **$p < 0.01$

tween emotion entropy and library holdings and GoodReads rating count is not linear. Different populations have different distributions: one group of titles with relatively low rating count and low library holdings is present at almost every level of entropy, while a group of titles with increasingly high rating count and library holdings cluster in a subset of the space.

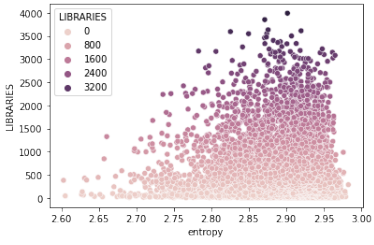


Figure 5: Distribution of library holdings with respect to emotion entropy.

To account for this hill-like distribution, we divided our data into two groups: one with low and the other with high rating counts (RC) and library holdings (LH), setting a threshold of ratings and library holdings at below 100 or above 500.[11] While these thresholds of 100 and 500 are somewhat arbitrary, they represent relatively robust trends in the data that can be reproduced with different cutoff points (see Fig 6 for the effect of different upper thresholds).

[11]Number of books in each goup: RC<100 = 2978, RC>500 = 4340; LH<100 = 2206, LH>500 = 3464

With this separation of marginally more "successful" and "unsuccessful" groups of titles, the relation between emotion entropy and quality perception is more evident: 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 point – are held in more libraries (see Fig.7).

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. Again, by grouping before correlating, we find a correlation between quality proxies and arc shape. It seems that more novels with more sloping arcs are rated less often and are held in fewer libraries; and where novels with more skewed emotions seem to have more ratings and library holdings (Table 3). We similarly correlated slopes of each emotion in a novel, as we might suppose that titles (or even genres) exhibit a steep slope for one emotion (not for others), making the mean unrepresentative. Here, we find that the average patterns represented in Table 3 hold for 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 library holdings, while the opposite appears true for skewness (Table 4).

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 re-

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

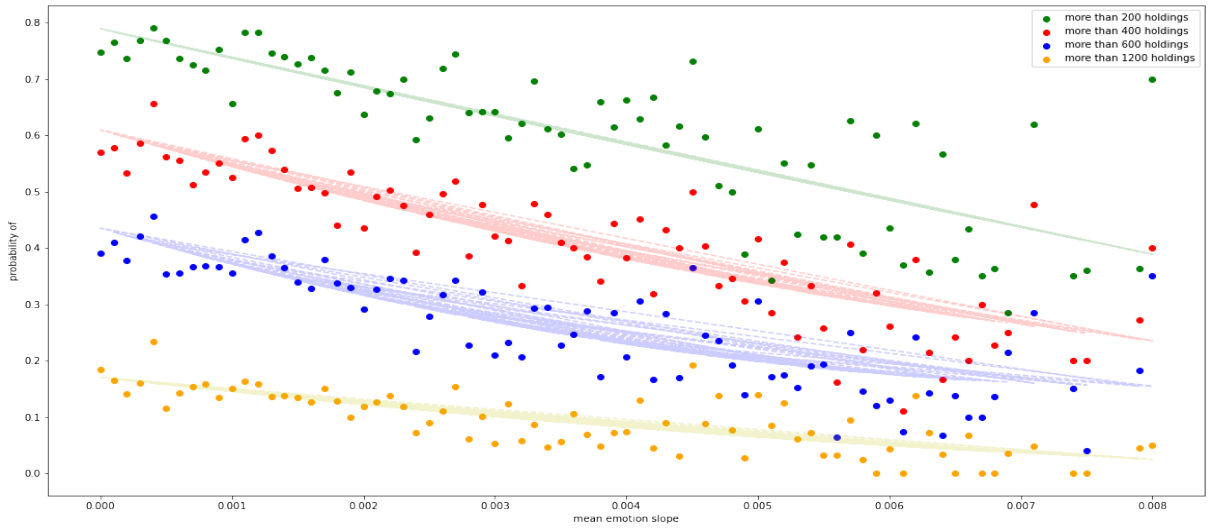


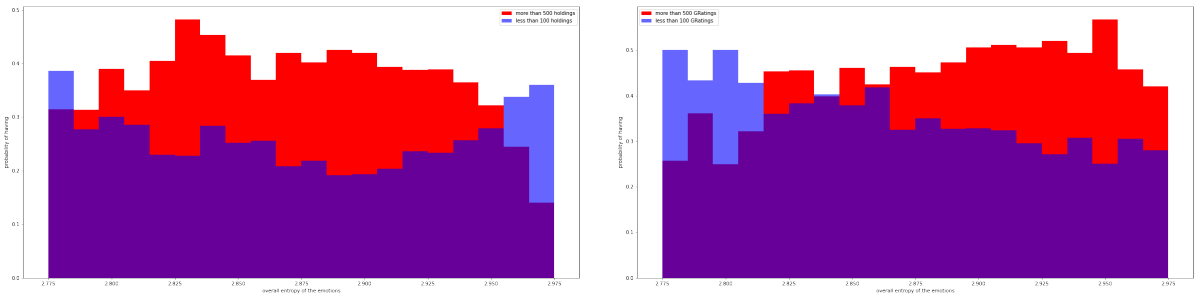
Figure 6: Trends in the probability of being in the high- or low-rating group at different cutting points of emotion slope value. While 100 and 500 rating counts and library holdings are somewhat arbitrary thresholds, trends in our data are reproduced at different cutoff points.

ception of the novels as measured via rating count and library holdings.

(i) **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.

(ii) **Slope**. A novel with steeper overall emotion arcs will have an overall 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.

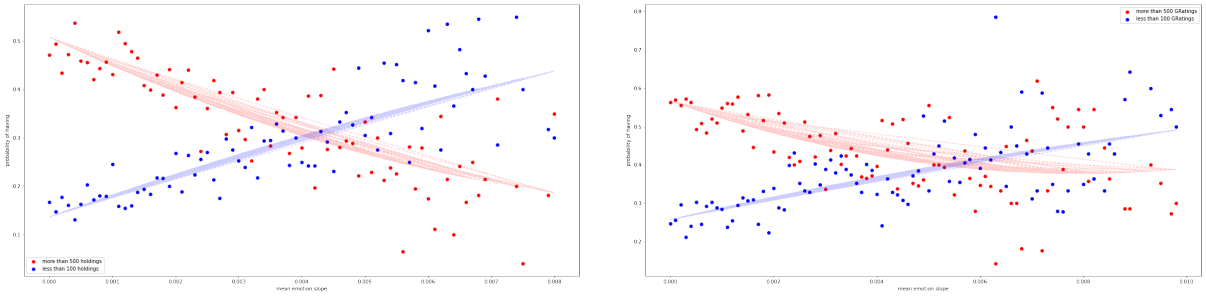
(iii) **Skewness**. A novel with a low level of overall emotion skewness will have an overall lower probability of being rated more than five hundred times on GoodReads or being held in more than five hundred libraries; conversely, it will have an increased probability of being rated less than 100 times and being held by less than 100 libraries. Our results on entropy might bear a relation to Jautze et al. (2016) regarding topics: novels with relatively few, dominating topics are perceived as being less good than novels that use a larger topical palette. There may be a similar effect at the level of the emotions represented in a text. It is 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 corrobo-



(a) Titles below 100 or above 500 holdings.

(b) Titles below 100 or above 500 ratings.

Figure 7: 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 8: 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.

rated by studies that have found that readers tend to prefer fractal story arcs but only with a moderate level of coherence (Hu et al., 2021; Bizzoni et al., 2021). Story arcs that 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 tools 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. 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. Emotional expectations are likely to vary greatly across genres and might yield further insight into the relation between affect, mood, and reception. We also aim to experiment with different proxies for perceived literary quality, 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 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 pro-

cess 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 are 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, we may have missed some details and welcome feedback.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22K18154.

References

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in* text and speech*. Ph.D. thesis, 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.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

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 and Pascale Feldkamp. 2023. Comparing transformer and dictionary-based sentiment models for literary texts: Hemingway as a case-study. In *Proceedings of the 3rd International Workshop on Natural Language Processing for Digital Humanities*, pages 219–226, Tokyo, Japan. Association for Computational Linguistics.

Yuri Bizzoni, Pascale Moreira, Nicole Dwenger, Ida Lassen, Mads Thomsen, and Kristoffer Nielbo. 2023a. Good Reads and Easy Novels: Readability and Literary Quality in a Corpus of US-published Fiction. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 42–51, Tórshavn, Faroe Islands. University of Tartu Library.

Yuri Bizzoni, Pascale Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023b. Sentimental Matters - Predicting Literary Quality by Sentiment Analysis and Stylometric Features. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 11–18, Toronto, Canada. Association for Computational Linguistics.

Yuri Bizzoni, Pascale Feldkamp Moreira, Kristoffer Nielbo, Ida Marie Lassen, and Mads Thomsen. 2023c. Modeling Readers' Appreciation of Literary Narratives Through Sentiment Arcs and Semantic Profiles. In *Proceedings of the The 5th Workshop on Narrative Understanding*, pages 25–35, Toronto, Canada. Association for Computational Linguistics.

Yuri Bizzoni, 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.

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.

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

Katherine Elkins. 2022. *The Shapes of Stories: Sentiment Analysis for Narrative*. Cambridge University 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.

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.

Jianbo Gao, H. Sultan, Jing Hu, and Wen-Wen Tung. 2010. Denoising Nonlinear Time Series by Adaptive Filtering and Wavelet Shrinkage: A Comparison. *IEEE Signal Processing Letters*, 17(3):237–240.

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

Matthew L Jockers. 2015. Some thoughts on Annie's thoughts . . . about Syuzhet. *M. 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. *arXiv preprint arXiv:1808.03137*.

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.

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

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, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Laigaard Nielbo. 2022. Reviewer Preferences and Gender Disparities in Aesthetic Judgments. In *CEUR Workshop Proceedings*, pages 280–290, Antwerp, Belgium.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. *W-NUT 2019*, page 34.

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.

Pascale Feldkamp Moreira, Yuri Bizzoni, Emily Öhman, and Kristoffer L. Nielbo. 2023. Not just Plot(ting): A Comparison of Two Approaches for Understanding Narrative Text Dynamics. In *Computational Humanities Research 2023*, pages 191–205, Paris, France. CEUR Workshop Proceedings.

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

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

Emily Öhman. 2020a. Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task. In *Digital Humanities in the Nordic Countries 2020*. CEUR Workshop Proceedings.

Emily Öhman. 2020b. Emotion Annotation: Rethinking Emotion Categorization. In *Digital Humanities in the Nordic Countries Post-Proceedings*, pages 134–144. CEUR WS.

Emily Öhman. 2021a. *The Language of Emotions: Building and Applying Computational Methods for Emotion Detection for English and Beyond*. Ph.D. thesis, University of Helsinki.

Emily Öhman. 2021b. 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. 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. *Stanford Literary Lab Pamphlet 17: Popularity/Prestige*. Stanford Literary Lab.

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. *EPJ Data Science*, 5(1):1–12.

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. 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. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 121–126.

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

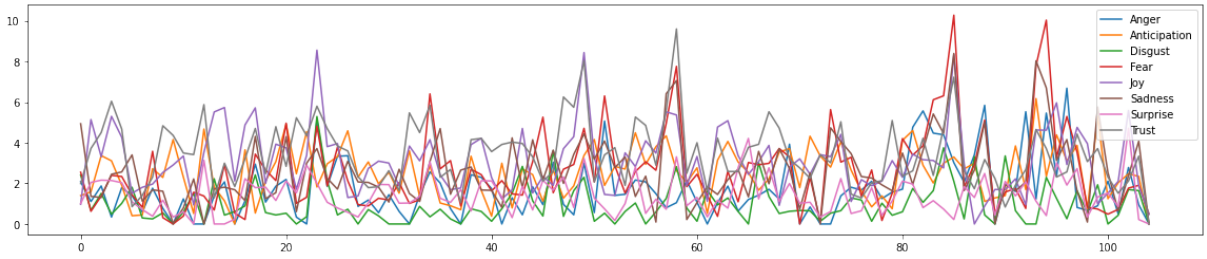


Figure 9: Unsmoothed emotion arcs for *The Old Man and the Sea*

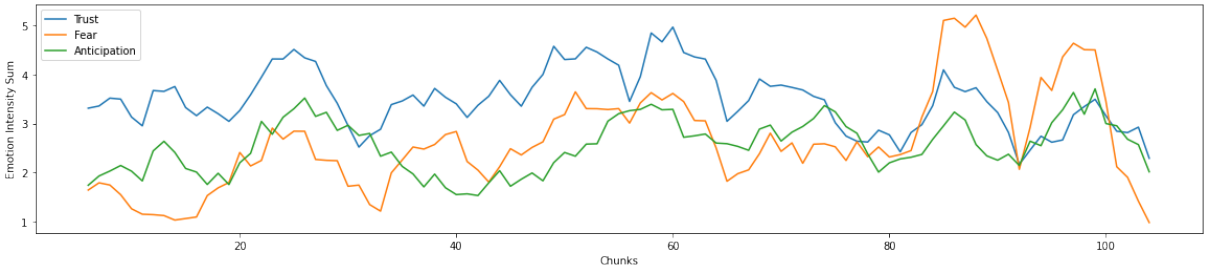


Figure 10: Smoothed arcs for trust, fear, and anticipation for *The Old Man and the Sea*

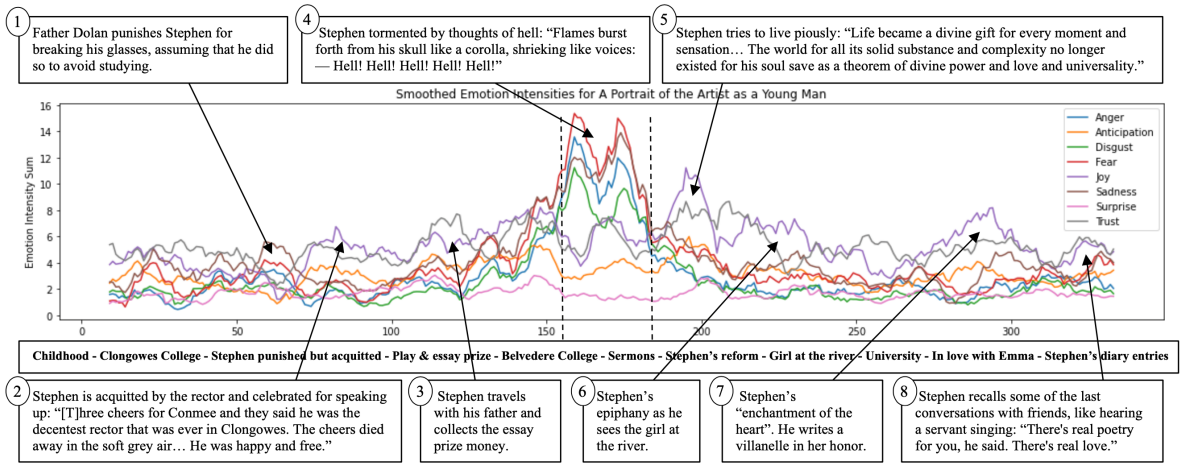


Figure 11: Another annotation of the novel *The Portrait of the Artist as a Young Man* by James Joyce. Note the peak of negative emotions in the center for this book, which Jockers (2015) has called a "man in a hole" narrative.

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 5: Pairwise correlation of emotions

Figure 12: 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.

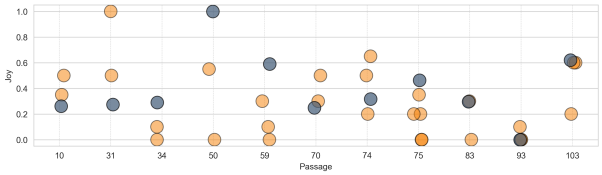


Figure 13: **Joy**, human and model scores.

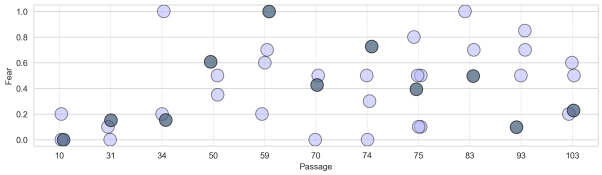


Figure 14: **Fear**, human and model scores.

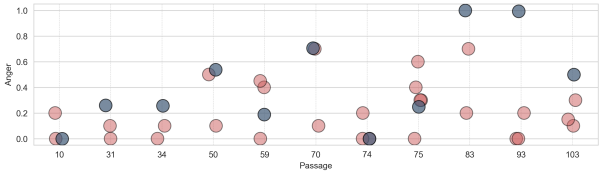


Figure 15: **Anger**, human and model scores.

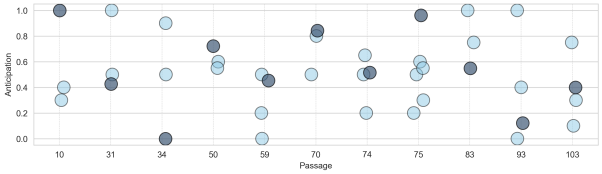


Figure 16: **Anticipation**, human and model scores.

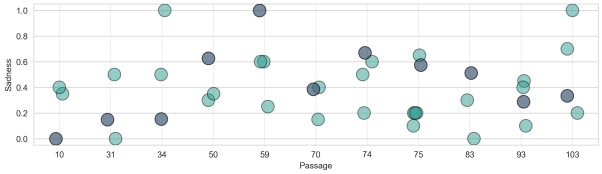


Figure 17: **Sadness**, human and model scores.

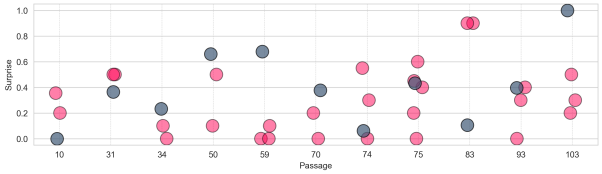


Figure 18: **Surprise**, human and model scores.

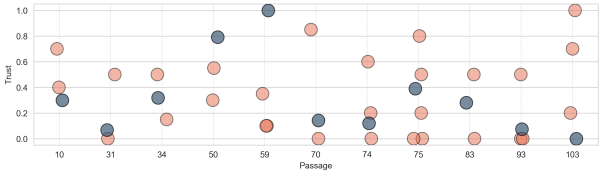


Figure 19: **Trust**, human and model scores.

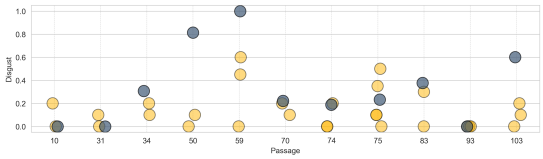


Figure 20: **Disgust**, human and model scores.

Annotator and model scores for 11 randomly selected passages per emotion. On the x-axis, each passage with scores arranged as increasing on the y-axis. For each passage, darker dots represent the EmotionArcs score for the emotion of the passage. Note that the number of annotators varies with respect to emotion and passage.