Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

Sentiment Analysis for Literary Texts: Hemingway as a Case-study

Yuri Bizzoni[1] **and Pascale Feldkamp**[1]

[1]Center for Humanities Computing Aarhus, Aarhus University, Denmark

Corresponding author: Yuri Bizzoni , yuri.bizzoni@cc.au.dk

Abstract

The literary domain continues to pose a challenge for Sentiment Analysis due to its complex, nuanced, and layered form of expression. Meanwhile, Sentiment Analysis is becoming increasingly central as a method in computational approaches to literary analysis for tracking the story arc and development or mood of narratives. This paper explores the adequacy of different Sentiment Analysis tools – from dictionary to transformer-based approaches – for capturing valence and modeling sentiment arcs. We take Ernest Hemingway's novel *The Old Man and the Sea* as a case study to address challenges inherent to literary language, and compare system scores with human annotations to shed light on the complexities of analyzing sentiment in narrative texts.

Going beyond simple comparison, we probe sentences where humans and systems diverge significantly, seeking to relate these disagreements to certain textual features that have been perceived central to the implicit way of expressing sentiment in literary texts. We find that sentences where humans detected significant sentiment – but where models did not – often employ language with lower arousal and higher levels of concreteness.

Keywords

computational narratology; sentiment analysis; implicitness; objective correlative; literariness

I INTRODUCTION

Recent years have seen a significant general increase in the methods available for Sentiment Analysis (SA). While dictionary-based approaches like VADER (Hutto and Gilbert [2014]) seem to perform well (Ribeiro et al. [2016]), they still struggle when applied to some domains (Elsahar and Gallé [2019], Ohana et al. [2012], Bowers and Dombrowski [2021]), in much the same way as more state-of-the-art transformer-based models do, despite providing a much richer semantic representation of texts (Tabinda Kokab et al. [2022], Öhman [2021]). Moreover, while these tools are commonly used to analyze emotive language in contexts like social media (Alantari et al. [2022]), their suitability for literary texts remains relatively under-explored.

Sentiment Analysis (SA) has become an increasingly central method for computational literary studies research (Rebora [2023]). A popular application of SA has been as a tool to explore the narrative development (such as happy endings)(Zehe et al. [2016]) or to gauge the "sentiment arcs" of novels (i.e., the consecutive highs and lows of sentiment throughout a narrative)(Jockers [2014], Reagan et al. [2016a]), which have also been used to explore, for example, the connection between narrative dynamics and reader appreciation (Bizzoni et al. [2023]).

Still, the relation between sentiment arcs extracted with SA tools and actual reader experience or human annotation of stories remains under-studied. Popular tools like the *Syuzhet* package have garnered critique (Swafford [2015]), and though extensive computational studies of literature like that of Elkins [2022] go some way in comparing various approach to SA, the question

of how to validate – against (which) experts/readers judgements, and at what level (i.a., the story, the sentence) – is still very present in the field. Moreover, literary language is a particularly intriguing case for testing SA tools because it often aims to evoke rather than explicitly communicate, operating at multiple narrative levels (Jakobson [1981], Rosenblatt [1982], Booth [1983]). The present study seeks to fill this gap by comparing model and human scores in one literary case study against human annotations. Here, we use *The Old Man and the Sea*, often considered the masterpiece of Ernest Hemingway and exemplary of his philosophy of writing, as a benchmark for testing both rule- and transformer-based SA systems.[1] Building on the literary analysis tradition that seeks to model sentiment arcs of narratives (Jockers [2014], Maharjan et al. [2018], Elkins [2022]), we apply various methods for sentiment annotation to the sentences of the novel and compare them to a benchmark of human annotations, both as raw values and as detrended values – seeing that some method of detrending is often part of SA workflows when doing literary analysis.

As a second step, we examine instances in which models and human annotators diverge, seeing these cases as insightful for developing SA methods for literary texts. While divergences between human and model SA scores are generally taken to indicate shortcomings in SA tools themselves, instances of divergence are informative both for model improvement and for gaining a deeper understanding of sentiment expression in literary texts, if we try to test whether certain textual features characterize them.

First, we seek to find sentences where human sentiment annotation diverges from that of models, the latter of which may not capture implicit or "omissive" sentiment as well as do human readers (Zhou et al. [2021], Li et al. [2021]). Then, we test whether these sentences of "implicit sentiment expression" can be told apart from other sentences by the specific textual features that characterize them. The choice of which features to look at was here informed by literary theory on evocation and descriptions of implicitness in Hemingway's style. They include: the mean valence,[2] arousal,[3] and dominance,[4] as well as their mean concreteness.[5]

By model comparison and by examining these textual features, the study assesses the complexities of analyzing sentiment in literary texts. We moreover assess the necessity for advanced SA methodologies tailored specifically for computational literary analysis, considering on the domain specificity of literature – nuanced characteristics which pose unique challenges that predominantly nonfiction-based SA tools may struggle to address effectively.

II RELATED WORKS

2.1 Sentiment in literary analysis

In literary studies, what is often called the "*affective* turn" (Armstrong [2014]) has led to a stronger focus on the sentiment and emotions expressed in narrative texts, whether certain emotions (Ngai [2007]) or affective states as linked to cultural codes (Ahmed [2014]). As such, recent applications of Sentiment Analysis to literature might be seen as both an extension of a focus that was already there, as well as an influx of methodology from Natural Language

[1] Link to the annotated text (human and automatic scores): https://github.com/PascaleFMoreira/Annotated_Hemingway

[2] The degree of positiveness or negativeness (/pleasure or displeasure) (Mohammad [2018a]).

[3] The degree to which a word prepares for action, captures or focuses attention (Borelli et al. [2018]).

[4] The degree of control evoked (Warriner et al. [2013]).

[5] The degree to which a word denotes a perceptible entity (Brysbaert et al. [2014]).

Processing. A popular application of SA in computational literary studies has been to profile texts and model the "shape of stories" (Reagan et al. [2016a]). To capture meaningful aspects of the reading experience, previous works have tested the potential of SA (Alm [2008], Jain et al. [2017]) at the word (Mohammad [2011, 2018b]), sentence (Mäntylä et al. [2018]), or paragraph level (Li et al. [2019]) to model narrative arcs (Kim and Klinger [2018], Reagan et al. [2016a], Jockers [2014]). Sentiment arcs have been used to evaluate literary texts in terms of shape or plot (Reagan et al. [2016a]), progression (Hu et al. [2020]), and mood (Öhman and Rossi [2022]). Specific shapes or arc dynamics have been connected to reader appreciation, considering both simple and more complex narratives (Bizzoni et al. [2022a, 2023]), and Bizzoni et al. [2023] have shown that sentiment features, such as measures of sentiment arc progression, have an effect even compared to the predominantly stylistic features usually employed for this type of task (Koolen et al. [2020], Maharjan et al. [2017]). As such, modeling sentiment arcs holds the potential for gaining a more in-depth understanding of how narratives, in their unfolding, affect readers.

However, both the validity of the dictionary-based approaches and the adequacy of methods for *detrending* arcs (Gao et al. [2016]) have been controversial in literary SA (Swafford [2015], Hammond [2017], Elkins [2022], Rebora [2023]). For example, dictionary-based methods seem to perform well even on so-called "nonlinear" narratives (Richardson [2000], Elkins and Chun [2019]), although they appear to do poorly on a word basis (Reagan et al. [2016b]). While more recent transformer-based approaches have been tested, they show potential and pitfalls in analyzing sentiment in literary texts (Chun [2021], Elkins [2022]). For literary texts Elkins [2022] has recently presented an exploration of methods for sentiment arc modeling, where it is suggested that ensembles of various models, dictionary-based methods and transformers alike, may help gauge narrative arcs by leveraging their different "perspectives" on a text.

2.2 Literary language and implicitness

Literary language may convey emotions in various ways beyond simply using words directly associated with emotional states (e.g., "happy"). While language on social media (which has had perhaps the largest applications of SA) may also rely on omission and subtlety, literary theorists have frequently claimed that literariness or the *poetic function of language* is distinct from its more communicative function;[6] and that it intentionally deviates from or distorts conventional language use (Mukařovský [1964], Attridge [1988]). Literary language may as such be perceived "un-communicative" in a manner that tweets are not, for example in the way they express stances or attitudes.

The concept of "implicit" expression is particularly relevant and complex in literary writing. Several theories of literary writing point to the importance of avoiding to present concepts or ideas (however this may be intended) in an explicit way. For example, the widely known precept of "Show Don't Tell" points at least partly in this direction. As is also made clear by Booth [1983], the distinction between types of narration (showing vs. telling) is not always adequate. However, critics continually rely on terms like emotional "evocativeness" and "understatement" to describe writing styles (Strychacz [2002], Daoshan and Shuo [2014]).

One writing style is known for its emotional subtlety is that of Ernest Hemingway. It is characterized (also by Hemingway himself) by its "iceberg" (Hemingway [1996]), or "omissive"

[6]Jakobson, for example, claims the "poetic function" to be distinct from it's "emotive or expressive function", which "aims a direct expression of the speaker's attitude toward what he is speaking about" ([Jakobson, 1981, p. 66])

technique, where: "the emotion is plentiful, though hidden and not exposed" (Daoshan and Shuo [2014]). As Hemingway noted, "[t]he dignity of movement of an iceberg is due to only one-eighth of it being above water" (Hemingway [1996]), implying that the "dignity" or expressiveness of his prose relies on some implicit strategy where more is evoked than is said.

It is unclear whether the implicit, evocative, and expressive strategies of literature – or of Hemingway in particular – can be reliably tracked in texts and whether more implicit types of narration display linguistically recognizable marks. If anything, literary language may be an optimal case study for testing whether this is the case. In our particular case of Hemingway, some aspects of his style hint at possible features of implicitness: the understated quality or "omissiveness" of his style and concreteness of his language.

Firstly, Hemingway's style is described as direct and limited in its use of figurative language (Heaton [1970]). It, moreover, avoids "overt emotional display", presenting actions and situations that *imply* emotions, and leaving their inference up to the reader (Strychacz [2002]): the reader is left to decode not only the life and background of the characters but also their feelings and the intensity of their experience. As such, it may be that Hemingway's "omissive" writing can be tracked by looking at the amount and intensity of sentiment expressions detectable in sentences itself, compared to how "expressive" readers perceive these sentences to be. The sentiment polarity, precision, and intensity of words have been formalized in the NRC VAD lexicon as valence, arousal and dominance (Mohammad [2018a]), a dictionary that has been used also in the literary context (Bizzoni [2022]).

Secondly, Hemingway's aversion to "emotional display and rhetorical overflow" has been linked to the Modernists' and New Critics' emphasis on *concreteness* over abstraction (Strychacz [2002]). The idea here is generally that expressive literature leverages more concrete language. The connection between concreteness and emotional expression is continually formalized in modern literary theory, as in popular notions of "show don't tell", where the most prominent concept is probably that of the *objective correlative* of T.S. Eliot. Eliot defined this as "a set of objects, a situation, a chain of events which shall be the formula of [a] particular emotion" (Eliot [1948]), suggesting a focus on concrete objects and actions over explicit emotion expression as *the* effective method for communicating emotion in literature. This idea has found some support in Auracher and Bosch [2016], which indicates that the concreteness of literary language impacts the emotional engagement of readers and their experience of literary suspense. As such, the evocativeness of literary language may coincide with instances where concrete objects and situations are more heavily described.

Concreteness of words has been measured on a scale from abstract (i.e., what cannot be experienced directly but the meaning of which is defined by other words) to concrete (i.e., what can be experience directly through one of the five senses)(Brysbaert et al. [2014]), ratings which have been widely used (Charbonnier and Wartena [2019]) also in the literary domain (Auracher and Bosch [2016], Flor and Somasundaran [2019]).

III METHODS

We first compare the raw arcs extracted with various models, as well as their detrended versions,[7] to human annotations. We then seek to probe instances of significant disagreement for textual features of implicitness

[7]That is, the same arcs after applying an adaptive filtering technique to reduce noise and "smoothen" the time-series.

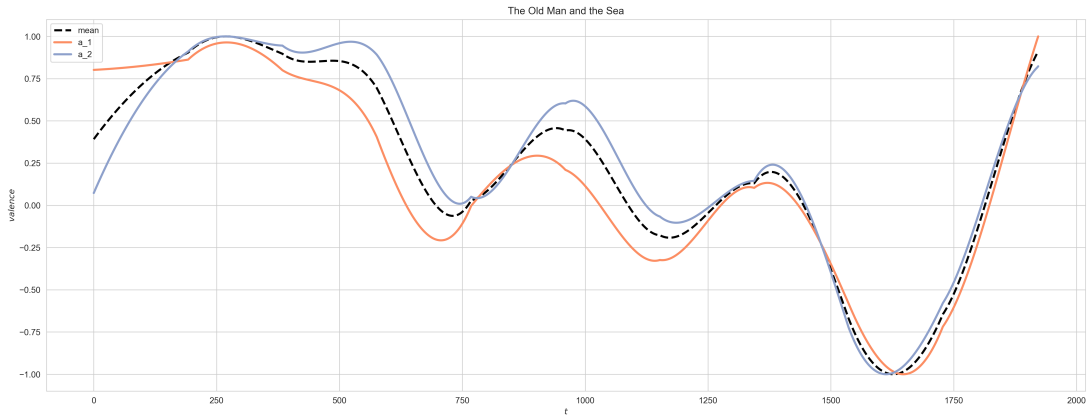

Figure 1: Arc of *The Old Man and the Sea* based on annotator (n=2) values. The dashed line represents the mean value of annotators.

3.1 *The Old Man and the Sea* as a case-study

The present study uses perhaps the most famous novel by Hemingway, *The Old Man and the Sea* to: (i) compare the performance of dictionary and transformer-based SA tools, and (ii) test the two hypotheses about human-model discrepancies – the positive role of understatement of sentiment-associated terms and the positive role of concreteness for experiences of sentiment in human readers.

As with Hemingway in general, the style of the novel is simple and direct. While the feelings of the characters are sometimes stated, their experiences and states of mind are often left to the reader to interpret from similes and object descriptions. This makes sentiment annotation a challenging task also for human annotators.

For example, the protagonist is introduced as a fisherman who hasn't caught a fish in a long time. Instead of mentioning his feelings, the narrator describes his scars: "They were as old as erosions in a fishless desert". This simile can be seen as a case of implicit sentiment as it arguably evokes a sense of despair for the lack of success but without using any explicit sentiment expression.

The reference to the pain and the fear of the characters is also often powerfully implied without any direct mention: "'Ay', he said aloud. There is no translation for this word and perhaps it is just a noise such as a man might make, involuntarily, feeling the nail go through his hands and into the wood". These descriptions, full of concrete objects such as the nail going through the hand, may be seen as a prime example of Eliot's *objective correlative*, where a "set of objects" is set in place to evoke emotion in the reader. Furthermore, when the protagonist is challenged in his final reckoning with the sharks, his fear and tension are rarely stated, but implied in the description of the sharks themselves.

While such passages may appear powerful for the human reader, it is likely that standard SA models would miss their sentimental charge. Words such as "nail" and "hand" gain emotional charge only in Hemingway's particular context and composition, but will not appear emotionally charged when observed as isolated words, as in sentiment lexicons. Our second experiment deals with this possible discrepancy by closely examining sentences of the novel in terms of the selected features valence, dominance, arousal, and concreteness.

3.2 Human Annotation

The first contribution of this paper is to provide a valence-annotated version of *The Old Man and the Sea*. Human annotators (n=2) read it from beginning to end and scored its 1923 sentences on a 1 to 10 valence scale: 1 signifying the lowest, and 10 the highest valence (see Fig. 1). Here, valence was intended as the sentiment expressed by the sentence. The annotators were instructed to avoid rating how a sentence made them feel and to try to report only on the sentiments actually embedded in the sentence, i.e., to think about the valence of each sentence individually, without overthinking the story's narrative to reduce contextual interpretation.[8] This naturally is far from an obvious or objective task, which created several interesting cases of uncertainty or ambiguity.

Both annotators have extensive experience of literary analysis, and hold degrees in literature.[9] They worked independently, not discussing nor subsequently changing scores. The task was not explicitly categorical: the annotators could use, in principle, decimals or even more fine-grained representations of their perceived valence. Nonetheless, both annotators resorted to using discrete values only. A representation of the detrended sentiment arc of each annotator is visualized in Fig. 1, along with their detrended mean.

As mentioned, *The Old Man and the Sea* is an advantageous case study for SA. While the story arc is linear and the style is simple, it is often ambivalent, shifting perspectives and narrative sympathies between the natural and human world, so that it can be difficult to annotate even for a human reader. For example, the sentence "Then the fish came alive, with his death in him, and rose high out of the water showing all his great length and width and all his power and his beauty" is stylistically simple, but offers a tension between contrasting emotions that challenges linear valence scales.

Accordingly, the correlation between the human annotators is not perfect, albeit very robust (Pearson: 0.652; Spearman: 0.624). The Cohen-Kappa score is 0.342. While this is relatively low, seeing as the annotators were working on a continuous valence space that was divided into ten discrete categories, we consider correlation measures to be more adequate than categorical inter-annotator agreement measures.

After detrending the arcs, the correlation between the annotators' arcs (also Fig. 1) is much more robust, with a Pearson correlation of 0.92. In short, this means that humans differ more on their sentence-by-sentence judgment of valence than they differ on the overall sentiment arc of the novel. The detrended arcs are, in fact, an attempt to draw the shape of the overall sentiment progress of a text, independently from the "noise" of individual sentences' ups and downs. As such, they tend to be more linear, more robust, and they tend to elicit higher correlations between models.

[8] While context-less scoring could be achieved by shuffling the sentences, we sought to make the reading experience itself as natural as possible (i.e., linear). The carry-over of sentiment from one sentence to the next which might be experienced by annotators also makes our task of distinguishing individual sentences solely by textual features harder, whereby we see our results as all the more significant (i.e., detectable even through the "noise" of the carry-over sentiment).

[9] Both were academics, male and female, at ages 31 and 34, who were non-native but very proficient English speakers, and who finished their literature degree (MA and BA) 2 years (MA) and 12 years ago (the BA).

3.3 Automatic scoring

All annotations were performed on a sentence basis (not considering the context).[10]

3.3.1 Transformers

For the automatic annotation of the novel's sentences, we used four SOTA transformers: (i) DistilBERT base uncased fine-tuned on SST2 (Sanh et al. [2020]), (ii) BERT base uncased fine-tuned on product reviews for SA (Peirsman [2020]), (iii) roBERTa base fine-tuned for SA on tweets (Barbieri et al. [2020]), (iv) roBERTa base fine-tuned for multilingual SA on tweets (Barbieri et al. [2022]).[11]

The first model returns two possible categories, *positive* or *negative*; models 3 and 4 also have the *neutral* category. Instead, model 2 returns five different categories, from 1, most negative, to 5, most positive. It's important to remember that, unlike dictionary-based models, transformers' output is categorical in nature. To use their output for representing continuous sentiment arcs, we have used the confidence score of their labels as a proxy for sentiment intensity. So if the model classifies a sentences as *positive* with a confidence of, for example, 0.89, we interpret it as a valence score on the sentence of +0.89. If the model classifies a sentences as *negative* with a confidence of 0.89, we interpret it as a valence score on of the sentence of -0.89. However, we couldn't do the same for the *neutral* category (or category 3 in system (iii)), so we simply converted these cases to a score of 0. Naturally, this may make the comparison less fair for these models than for the models already designed for a continuous scoring approach. On the other hand, our quest is precisely to find out, which model(s) approximate a human continuous valence rating on literary texts.

3.3.2 Dictionary-based models

To compare against transformers, we chose two dictionary-based approaches: (i) nltk's implementation of VADER (Hutto and Gilbert [2014]), arguably the most widespread dictionary-based method for SA. (ii) Syuzhet (Jockers [2014]), a widespread implementation designed to model literary arcs.[12] Both of these models are dictionary- and rule-based, and return continuous scores ranging from -1 (negative) to +1 (positive).

3.4 Sentiment arcs

A sentiment arc refers to a simple 1D representation of sections of a literary work (e.g., the valence of words, sentences, or paragraphs). Because narratives and derived arcs based on the valences are inherently noisy and nonlinear, studies typically apply some technique for detrending or "smoothing" the arcs to reduce noise and extract the global narrative trends - from a simple moving average window to more complex noise reduction techniques (Chun [2021], Jockers [2015a], Bizzoni et al. [2021], Gao et al. [2016]). As wavelet approaches typically used for noise reduction are not ideal for nonlinear series, Jianbo Gao et al. [2010] proposed an adaptive filtering technique for nonlinear series. Studies have demonstrated the usefulness of adaptive filtering applied to sentiment arcs, especially in the context of estimating the dynamics of sentiment arcs (Hu et al. [2020], Bizzoni et al. [2022b]).

[10]Sentences were tokenized using the nltk tokenize package: https://www.nltk.org/api/nltk.tokenize.html

[11]We included the multilingual roBERTA to test this model for future work on multilingual literary corpora.

[12]Though tools developed less broadly tend to rely on less data, the Syuzhet dictionary is relatively large: extracted from 165,000 human-coded sentences from contemporary literary novels, developed in the Nebraska Literary Lab (Jockers [2015b]).

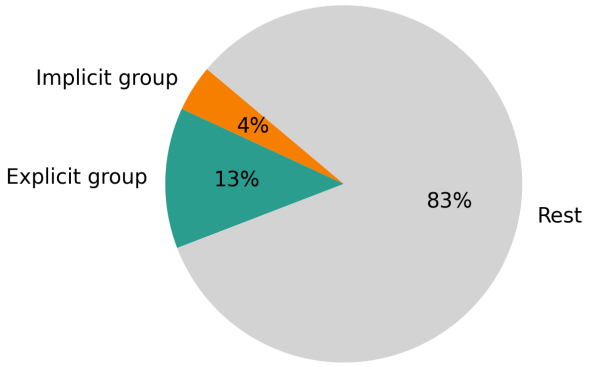

Figure 2: Division of sentences of *The Old Man and the Sea* into groups of: 81 sentences where human and model sentiment scoring diverged significantly, and 245 sentences where it converged.

3.5 Identifying and probing instances of disagreement

For our second experiment, we distinguished a subset of sentences that may be seen to represent discrepancies between human and model sentiment perception – that appear powerful to human readers but not for automatic annotation systems.

Sentence	DistilBert	Bert	Roberta	Roberta_xlm	Vader	Syuzhet	Human
Then he felt the gentle touch on the line and he was happy.	**0.9998**	4.42	0.94	0.68	0.76	0.42	6.5
Blessed art thou among women and blessed is the fruit of thy womb, Jesus.	0.9982	**5.91**	0.84	0.86	0.83	0.45	6.5
"Tomorrow is going to be a good day with this current," he said.	0.9991	4.37	**0.98**	0.89	0.44	0.19	6.5
Bed will be a great thing.	0.9996	5.59	0.95	**0.91**	0.62	0.14	7.5
But he was such a calm, strong fish and he seemed so fearless and so confident.	0.9997	5.38	0.85	0.75	**0.95**	0.72	**8.0**
The boy had given him two fresh small tunas, or albacores, which hung on the two deepest lines like plummets and, on the others, he had a big blue runner and a yellow jack that had been used before; but they were in good condition still and had the excellent sardines to give them scent and attractiveness.	0.9972	4.5	0.8	0.45	0.94	**1.0**	7.0

Table 1: A comparative performance overview of the models, presenting sentences that elicited the highest scores. Values are not normalized; all models return a score between -1 and 1 (except for BERT, which ranges from 1 to 6). Human ratings range from 0 to 10.

To create a subset of such sentences, we used the distance between SA models' and humans' annotations of sentences. We selected those cases in which human readers perceived sentimental charge (whether positive or negative), but where models did not. We proceeded in the following way: (1) We selected all sentences that were *not* scored neutral or near-neutral by human annotators (all sentences scoring lower than 5 or higher than 6), i.e., where human readers did detect some sentiment. This subset accounted for less than half of the sentences of the novel: a total of 687 out of 1923 sentences.
(2) Of this subset, we selected only those sentences that did *not* elicit a strong sentiment score from 3 SA models, using the best-performing ones: the VADER dictionary, Syuzhet dictionary, and roBERTa base.

We thus only kept sentences that had normalized absolute scores smaller than 0.1 in *all three*

	DistilBert	Bert	Roberta	Roberta_xlm	Vader	Syuzhet	Average	Select
Kendall τ	0.39	0.28	**0.50**	**0.50**	0.36	0.36	0.48	0.50
Spearman r	0.51	0.36	0.57	**0.59**	0.43	0.45	0.59	0.61
Pearson r	0.42	0.36	**0.63**	**0.63**	0.46	0.48	0.65	0.70
Pears. per annot.	.41/.48	.35/.30	.59/.56	.59/.55	.45/.39	.46/.41	.62/.55	.66/.61
Kendall τ	0.62	0.49	0.75	0.73	0.41	**0.84**	0.83	0.84
Spearman r	0.80	0.68	0.90	0.89	0.57	**0.96**	0.96	0.96
Pearson r	0.80	0.71	0.90	0.85	0.68	**0.96**	0.96	0.96
Pears. per annot.	.88/.71	.70/.69	.92/.85	.87/.81	.62/.70	.90/.97	.96/.93	.95/.93

Table 2: **Top**: correlations between *raw* annotations and the human mean values. The last row indicates the Pearson correlation per method to each annotator individually. **Bottom**: Correlations between *detrended* annotations and the human mean values. Again, the last row indicates the Pearson correlation to each annotator. For all correlations, *p-values* < 0.01.

models.[13] In short, we selected all sentences that appeared sentiment-charged to humans while being scored as neutral or almost neutral by all three SA systems. This left us with 81 sentences in what we call the "implicit" group (Fig 2).
(3) For comparison, we selected sentences where humans and models were more aligned in their sentiment scoring, what we call the "explicit" group (Fig. 2).

These are sentences where both humans and models found either a positive or negative sentiment (above an absolute model score of 0.1) and agreed on the sentiment direction (positive/negative).
(4) We then proceeded to compare the "implicit" group of sentences to the where SA models were neutral but humans were not, to the set of sentences where model and human score were more aligned. We compared the groups in terms of the selected features: valence, arousal, dominance,[14] and concreteness.[15] Finally, we used a Mann-Whitney U test to examine differences between the groups.

IV RESULTS

4.1 Comparing models

To evaluate the models we use the average of the annotators' scores (see Table 1 for some selected samples of the models' performance).

In Table 2 we present the correlations between each model and the human baseline. We also add the correlations with two "ensemble" approaches: the average of all SA models' outputs and a select average of the outputs of only Roberta, Roberta-xlm, and Syuzhet: the three best-performing models, and add the mean R2 score for comparison (Table 3). [16]

Our results show that large pre-trained transformers correlate with human judgments on the va-

[13]Normalized scores are between -1 and 1, the absolute score of 0.1 refers to the interval between -0.1 and 0.1.
[14]We used the VAD lexicon (Mohammad [2018a]) to retrieve the valence, arousal, and dominance scores for each word, averaging scores over each sentence: https://saifmohammad.com/WebPages/nrc-vad.html
[15]To retrieve concreteness scores of words and lemmatized sentences individually, we used the concreteness lexicon by Brysbaert et al. [2014]: http://crr.ugent.be/archives/1330
[16]The R2 score is a statistical measure for regression models that represents the proportion of the variance of a dependent variable that is explained by an independent variable: in other words, it represents how well the model fits real data. It ranges between 0 (no explanation), to 1 (complete explanation of the variance).

	DistilBert	Bert	Roberta	Roberta_xlm	Vader	Syuzhet
Raw	0.13	0.11	0.39	0.33	0.15	-1.03
Detrended	0.34	-0.38	0.43	-0.11	0.23	0.91

Table 3: R2 scores for time series compared to the human mean values.

lence of sentences better than the rule-based VADER and Syuzhet. Thus, despite transformer's output on each sentence being categorical, it appears that their confidence scores can be successfully used as proxies for valence intensity even on a set of literary sentences (see Fig. 3 for a visual comparison of the values' distribution).

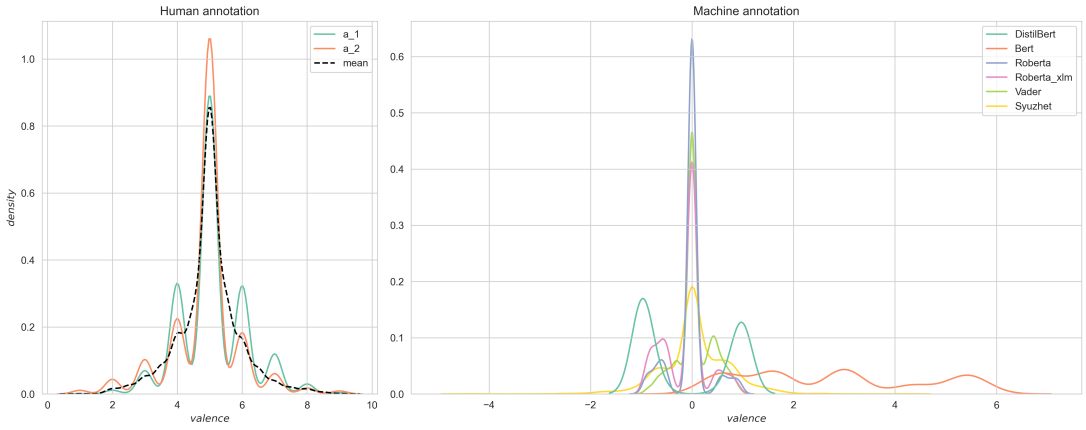

Figure 3: Kernel density plots visualize the distributions of values (0 or neutral being the most common). Note that value ranges differ: the BERT model, for example, assigns valence on a 5-point scale, while human annotators could assign any (round) value between 0 and 10.

Still, it is notable that the dictionary-based systems outperform half of our transformer population. Interestingly, the correlation of each model with each individual human is *lower* than the correlation of each model with the average human annotation (Table 2) - in other words, sentiment seems to act almost as an objective measure, with individual stochastic "errors" reduced through repeated annotation. If we observe the sentences with the highest disagreement between (average) human judgment and the best performing transformer, Roberta-xlm, we find that these sentences tend to be short, where the model displays a negativity bias; while the sentences where the best performing dictionary-based model, Syuzhet, is most removed from the human evaluation appear to be long sentences with a complex semantic interplay, for which it displays a sort of positivity bias (Table 4). In Fig. 4 we show a visualization of the raw sentiment annotations of the last 50 sentences of the novel.

Finally, the sentences with the most disagreement between the two models are often sentences that were also difficult for human annotators. As illustration, we show a small selection of such sentences in Table 4.

When detrending the series of valence we find that the picture changes: Syuzhet now outperforms all of the Tranformers (Table 2). It is possible that in the case of Syuzhet, the errors at the level of raw scores, where humans set a negative score and Syuzhet a positive score (see Fig. 4),[17] are big enough to impact the overall correlation with human annotations, but are still few enough to be "canceled" out in detrending so that dictionary-based arcs are the closest ones

[17]This may be due to systematic errors, such as the issue with negations in Syuzhet.

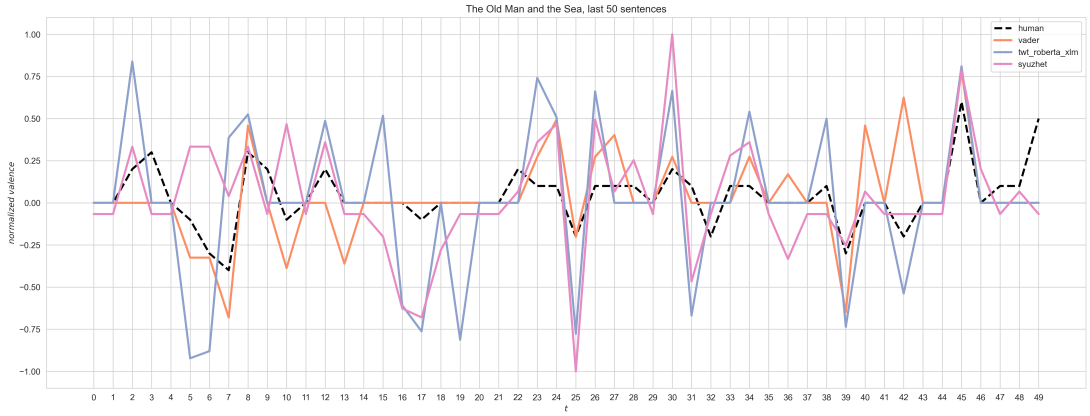

Figure 4: Arc of the last 50 sentences of *The Old Man and the Sea* with on transformer and dictionary-based annotation. The added dashed line represents the mean value of human annotators. Note that sentences like [5]: "I am not lucky", and [10] "I do not care", are systematically misjudged as positive in the Syuzhet, annotation despite the negations.

to the human arcs. The detrending process essentially flattens out raw scores, so that scores that are proximate are more alike. In this sense, detrending the series gives us a picture of the annotation tendencies at each point of the arc, and smoothens out scores that diverge suddenly from the overall tendencies.

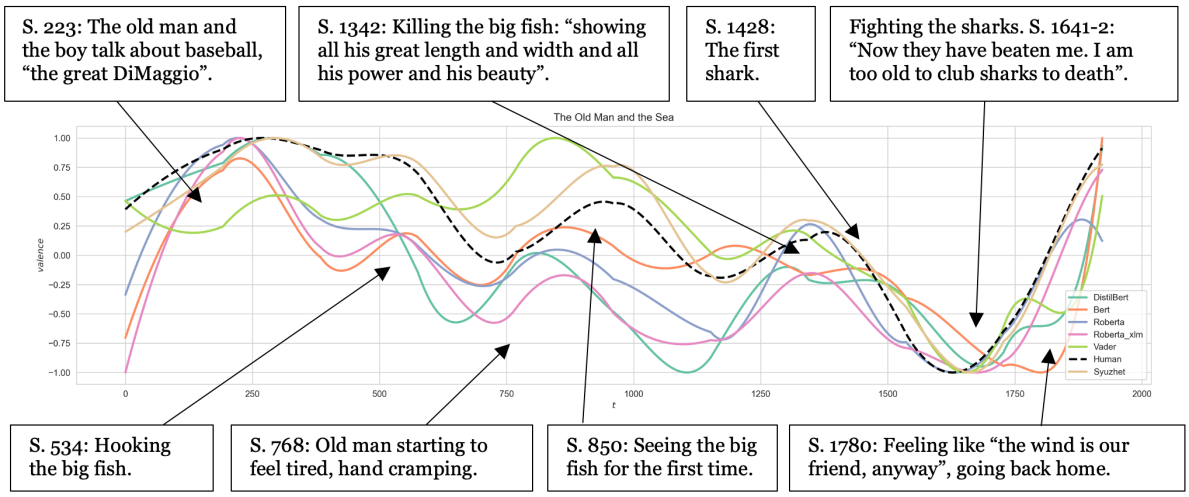

Figure 5: Arcs of *The Old Man and the Sea* based on various methods, with manual annotations of corresponding narrative events. The added dashed line represents the mean value of human annotators.

4.2 Comparing annotators

Literary language is a challenge to SA models due to its subtlety and its complexity. Narrative sentences can be as complex as those of any other domain, yet because literary texts aim for their readers to experience rather than just be informed, they seem especially difficult to annotate. Looking at the human scores of *The Old Man and the Sea*, we found that annotators used almost the whole range (1 to 10), going from 2 to 9. Though annotators were instructed not to overthink the narrative to reduce contextual scoring, this was not always easy. Hemingway's direct style partly facilitated annotation e.g. ("Fish," he said, "I love you and respect you very much"), but underlying complexity sometimes sparked uncertainty and disagreement for human annotators. Despite being negative agents in the story, the sharks, for example, are still described

Sentence	Roberta_xlm	Syuzhet	Human
They were immune to its poison	-.87	-.05	.3
Perhaps he is too wise to jump	-.68	-.14	.3
”I wish the boy was here”, he said aloud and settled himself against the rounded planks of the bow and felt the strength of the great fish through the line he held across his shoulders moving steadily toward whatever he had chosen.	.42	.63	-.1
There is no one worthy of eating him from the manner of his behaviour and his great dignity.	-.92	.2	-.1
The old man’s head was clear and good now and he was full of resolution but he had little hope	-.85	.15	-.2

Table 4: Examples of sentences with the largest disagreement *between machine and (normalized) human score* for Roberta-xlm (upper rows of the table) and for Syuzhet (central rows of the table). Roberta-xlm is most off track for short and relatively ambiguous sentences, while Syuzhet appears to disagree more with long and complex sentences. Examples of sentences that instead elicit a large disagreement *between the two models* are in the lower rows of the table. These sentences are often also complex to judge for human annotators.

as “beautiful”, and the protagonist is portrayed as both “beat” and “undefeated”. Several of the larger inter-annotator disagreements were often due to the presence of co-existing valences in the same sentence. Several of such sentences elicited differing judgments from the models as well: for example the sentence “*The old man hit him on the head for kindness and kicked him, his body still shuddering, under the shade of the stern*” elicited scores of 6 and 2 from the annotators, -.97 from DistilBert and +.46 from VADER (normalized values).

We have already observed that almost all models correlate less with individual annotators than with the mean of the annotators, an effect that is magnified when we also compute the mean of all the models’ scores. The average annotation of all the models (after normalization) correlates with the human judgments better or as well as the individual models, both for the raw scores and for the detrended arcs.

4.3 Comparing annotators and models

Our selected group of sentences represents a divergence between human and text-based SA systems: humans found them to express some form of sentiment not detected by the three SA models. Notably, the average absolute human score of the “implicit” group was slightly higher (0.23) than the average score of the “explicit” group (0.22). For example, the sentence *“The other watched the old man with his slitted yellow eyes and then came in fast with his half circle of jaws wide to hit the fish where he had already been bitten*” is perceived as negative by human annotators, but does not contain any of the explicit expressions of negative emotion that text-based SA models usually pick up on.

We computed the average valence, arousal and dominance (VAD) using the NRC-VAD-Lexicon. These measures attempt to position a word in a three-dimensional space, detailing different aspects of a word’s affective semantics. For example, *lion* is higher than *shark* in valence and dominance, but lower in arousal.

For concreteness, we used Brysbaert et al. [2014]’s lexicon of English lemmas. This resource complements the elements modelled by the NRC Lexicon, as it attempts to quantify the con-

	Valence	Arousal	Dominance	Concreteness
Implicit	0.583 ±0.111	0.409 ±0.119	0.499 ±0.108	2.735 ±0.347
Explicit	0.563 ±0.198	0.467 ±0.113	0.500 ±0.119	2.609 ±0.327
MWU test	8536.5	12047.0*	9048.5	7634.0*

Table 5: Mean and st.d. feature values of the implicit and explicit groups, as well as the results of the MWU test between the groups in each setup. In the implicit group: sentences perceived as non-neutral by humans but as neutral by models (below an absolute score of .1); in the explicit group, sentences where humans and models were more aligned in their recorded intensity (models' score above an absolute of .1, humans' score above 6 or below 5). * p-value < 0.01.

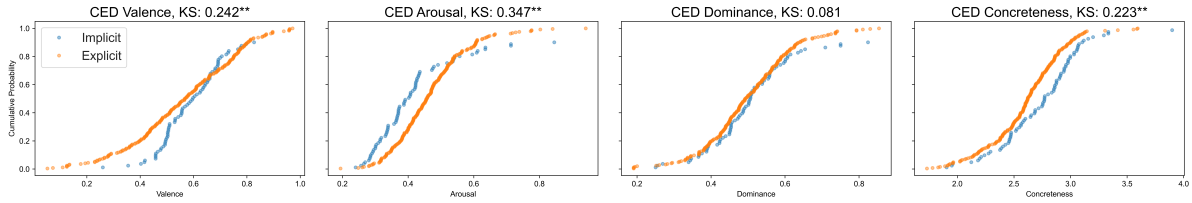

Figure 6: Cumulative Empirical Distribution (CED) of features per group and statistics of the two-sample Kolmogorov-Smirnov test (KS) for goodness of fit (on top). **p-value < 0.01.

creteness of each word independently from its affective aspect, even if it has been suggested that abstract words are connected to a stronger valence than concrete words (Kousta et al. [2011]). These dimensions of lexical semantics can appear quite uncorrelated, but their interplay appears evident when looking at many of the "implicit sentiment" sentences from the novel, like the one cited above.

We then compared the average valence, arousal, dominance, and concreteness of the words used in the sentences perceived by at least one SA model as having an absolute sentimental intensity stronger than .1 (714 sentences) with those of the words used in the sentences that only humans perceived as sentimentally charged (81 sentences). Using the Mann Whitney U test, we computed which of the differences in textual features between the two groups are significant. Here, we find that while valence and dominance do not show significant differences between the two groups, "implicit sentiment" sentences have a much lower arousal and a slightly higher concreteness, on average, than the set of "explicit" sentences – as can be seen in Table 5. Two of the four feature dimensions appear to be significant in the sentences that implicitly express a sentiment: their level of concreteness and their level of arousal.[18] Valence in sentences with lower arousal and higher concreteness appear more detectable to the human eye than to models, pointing to a discrepancy between them. Concrete words seem to contribute to create the implicit effect that SA models have a hard time detecting – SA models have a harder time picking the sentiment of words like *blood* and *teeth*. The statistical significance of the two relevant categories is even stronger when they are measured on a sentence- rather than word base (Table 5).

This interplay could be precisely one of the components of the "*omissive* prose" effect.

For example, one sentence which was perceived very positive by human readers and neutral by models also holds high concreteness (2.78): *"The boy took the old army blanket off the bed and spread it over the back of the chair and over the old man's shoulders"*. It seems to exemplify

[18]The lack of difference in valence is likely an effect of groups confounding positive and negative sentences.

	DistilBert	Bert	Roberta	Roberta_xlm	Vader	Syuzhet
Avg. difference	0.86	0.48	0.19	0.26	0.23	0.16
Std	0.22	0.32	0.22	0.26	0.24	0.15

Table 6: Mean difference and standard deviation between human valence and models' valence.

the notion of objective correlative – that is, the literary technique of transmitting sentiment to readers without using emotion associated words, through an exposition of concrete *objects* or *actions*.[19]

To further validate these results, we examined the distribution of our data performing the Kolmogorov-Smirnov (KS) test[20] on the empirical cumulative distribution of the groups (Fig. 6).

Considering the test values, we may reject the null hypothesis that the two groups are drawn from the same continuous distribution in the case of valence, arousal, and concreteness (see Fig. 6 and Table 7).[21]

	Constant	Valence	Arousal	Dominance	Concreteness
Coefficient	-2.1609	-3.4922**	-7.2940**	8.9520**	1.1254**

Table 7: The table presents the coefficients and associated p-values resulting from the Ordinary Least Squares (OLS) regression analysis. We performed the regression on the combined "implicit"/"explicit" groups of sentences (n=81 / 245) using *the difference between human and roBERTa sentiment score* as the dependent variable. The coefficients represent the estimated effect of each independent variable (our four features) on the dependent variable, score divergence. * p-values < 0.01 indicate that all variables have a statistically significant impact on score divergence.

V DISCUSSION AND CONCLUSIONS

For this case-study in comparing sentiment annotation methods for literary analysis, we have compared the correlations between human annotations and several SA systems' annotations of the sentences of the novel *The Old Man and the Sea*. While sentiment analysis is often tackled as a classification problem (with two or three categories at most), we found this approach to be exceedingly coarse-grained to verify the efficacy of SA models on literary texts, and we preferred to model it as a continuous scoring task. Most of the time human annotators would have been unable to fit a sentence into a binary classification, and the most interesting behaviours of the models happen when looking at their ability to position a sentence on a nuanced continuum. Naturally, it is now possible to operate the opposite operation and convert the continuous annotations into two or three categories, to compare them directly with the transformers' outputs.

We have observed interesting differences between transformer- and dictionary-based methods. Still, it should be noted that our analysis was performed on one story only, even though the

[19]We only suggest this effect as the method we use – the VAD and concreteness scores – may be considered a relatively crude way of operationalizing this concept.

[20]We used the implementation of this test in the SciPy library: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html.

[21]The significance of valence is predictable, as we have selected the sentences based on their valence. However, it is not picked by all models as it "crosses over" the distribution of explicit sentences. That is, implicit sentences are more positive than the most negative explicit sentences, and more negative than the most positive explicit sentences.

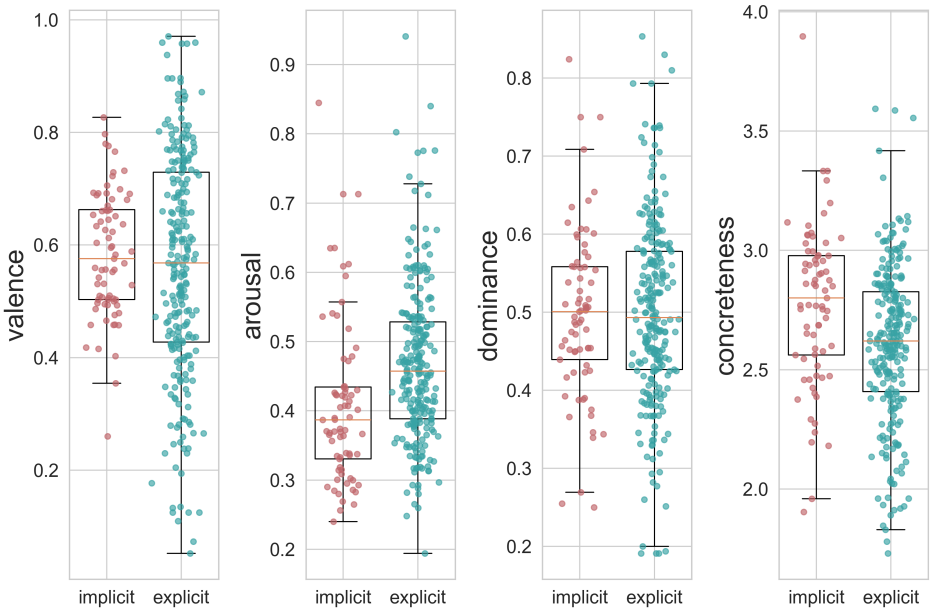

Figure 7: Boxplots comparing implicit (n=81) and explicit (n=245) groups of sentences by scores for each of the four features.

particular example of *The Old Man and the Sea* appears particularly apt as case-study for Sentiment Analysis, considering its emotionally understating literary style. Despite being categorical in nature, the largest transformers of our collection proved to hold strong correlations with human judgments in the sentence-level annotation – higher than the dictionary-based models VADER and Syuzhet. When looking at the detrended versions of the arcs, the picture is reversed: despite serious shortcomings of the tool (Kim [2022]), the detrended arc made from the Syuzhet package's scores appear to be the most closely related to the detrended version of the human arcs (Fig. 5). In both cases, the best results are achieved when using both transformer and dictionary-based systems, as they appear to be at least partly complementary, and our best model correlates with the mean human score almost as much as humans correlated with each other (Table 2). We have observed that average human judgments seem to be more aligned to models than individual judgments, and average automatic scores from different sources seem to work better than the scores of any individual model. Moreover, at the sentence level, while roBERTa correlated with human judgments best, VADER and Syuzet are closer to human intensities: on average VADER and Syuzhet have a smaller mean distance from human intensities, and a lower standard deviation (Table 6).[22]

Beyond providing the best correlation with human judgments, it's possible that a compound approach, integrating the scores of two or more models, would be greatly beneficial for something else: the detection of confounding or polarizing sentences, likely to elicit differing or opposite scores.

In fact, while comparing human and model sentiment annotations in the novel, we observed a distinct group of sentences that garnered high human scores, but received neutral ratings from our three SA models. Looking into textual features of this group, we found that they can be distinguished by their levels of arousal and concreteness. Because we might assume that

[22]We also observe that, when inspecting raw scores, transformers seem to be more "extreme" in their judgement than human and dictionary-based models. See Fig. 4 for a visualization.

humans in these cases pick up on contextual information not available to the models - even if human annotators were instructed to focus on individual sentences - we find the difference in terms of textual features between the groups particularly interesting. More than just context appear to be giving these sentences an evocative strength that is not captured by the models.

The finding of higher levels of concreteness and lower levels of arousal of this group of sentences aligns with literary theories suggesting that writing styles that employ techniques like "omissive writing" or the *objective correlative* evoke a perception of sentiments in human readers without any explicit emotional reference and without using words directly associated to emotional states. In other words, the fact that sentences appearing affective to humans but not models stand out in terms of arousal and concreteness suggests that the sentimental effect is not achieved through any direct reference to feelings or emotions. Rather, the evocative strength of these sentences relies at least in part on words with a low arousal profile, and higher concreteness levels, managing to be particularly subtle in how sentiment charge is transmitted to the reader. Our findings support supplementing sentiment models with feature detection when dealing with the literary domain, since it may be that fiction texts use language differently than non-fiction, e.g., employing objective correlatives to evoke sentiment in the reader, as we have seen in this study. Further exploration into arousal and concreteness may hold promise for a more comprehensive understanding of sentiment in prose in fiction with that in non-fiction. Broader quantitative studies of fiction would help understanding how concreteness and arousal resonate with readers, particularly regarding their appreciation of implicit sentiments' evocation in prose. Finally, further analyses of literary texts where different scores of sentimental intensity diverge significantly promises to shed light on literary techniques that go beyond description and into the evocation of feelings in the reader's experience. After all, some of the sentences with the largest difference between rule-based and transformer-based scores are beautifully complex to judge for human readers alike, such as the sentence that elicited the the highest disagreement between models: *"I killed him in self-defense," the old man said aloud. "And I killed him well."*

References

Sara Ahmed. *The cultural politics of emotion*. Edinburgh University Press, Edinburgh, second edition edition, 2014. ISBN 978-0-7486-9114-2 978-0-7486-9113-5.

Huwail J. Alantari, Imran S. Currim, Yiting Deng, and Sameer Singh. An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. *International Journal of Research in Marketing*, 39(1):1–19, March 2022. ISSN 01678116. doi: 10.1016/j.ijresmar.2021.10.011. URL https://linkinghub.elsevier.com/retrieve/pii/S0167811621000926.

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

Nancy Armstrong. The Affective Turn in Contemporary Fiction. *Contemporary Literature*, 55(3):441–465, 2014. ISSN 0010-7484. URL https://www.jstor.org/stable/43297971. Publisher: [Board of Regents of the University of Wisconsin System, University of Wisconsin Press].

Derek Attridge. *Peculiar Language*. Routledge, 1988.

Jan Auracher and Hildegard Bosch. Showing with words: The influence of language concreteness on suspense. *Scientific Study of Literature*, 6(2):208–242, December 2016. ISSN 2210-4372, 2210-4380. doi: 10.1075/ssol.6.2.03aur. URL http://www.jbe-platform.com/content/journals/10.1075/ssol.6.2.03aur.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification, October 2020. URL http://arxiv.org/abs/2010.12421.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond, May 2022. URL http://arxiv.org/abs/2104.12250.

Yuri Bizzoni. Correlations between GoodReads Appreciation and the Sentiment Arc Fractality of the Grimm

brothers' Fairy Tales. In *Proceedings of the Computational Humanities Research Conference*, page 12. CEUR-WS, 2022.

Yuri Bizzoni, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 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, 2021. NLP Association of India (NLPAI). URL https://aclanthology.org/2021.nlp4dh-1.1.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 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*, NLP4DH, 2022a. doi: 10.46298/jdmdh.9154.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 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, nov 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.nlp4dh-1.5.

Yuri Bizzoni, Pascale Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 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, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wassa-1.2.

Wayne C. Booth. *The Rhetoric of Fiction*. University of Chicago Press, Chicago, 2nd edition edition, February 1983. ISBN 978-0-226-06558-8.

Eleonora Borelli, Davide Crepaldi, Carlo Adolfo Porro, and Cristina Cacciari. The psycholinguistic and affective structure of words conveying pain. *PloS one*, 13(6):e0199658, 2018. doi: 10.1371/journal.pone.0199658.

Katherine Bowers and Quinn Dombrowski. Katia and the Sentiment Snobs, 2021. URL https://datasittersclub.github.io/site/dsc11.html. Blog: Datasitter's Club.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911, September 2014. ISSN 1554-3528. doi: 10.3758/s13428-013-0403-5. URL https://doi.org/10.3758/s13428-013-0403-5.

Jean Charbonnier and Christian Wartena. Predicting word concreteness and imagery. In Simon Dobnik, Stergios Chatzikyriakidis, and Vera Demberg, editors, *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 176–187, Gothenburg, Sweden, May 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-0415. URL https://aclanthology.org/W19-0415.

Jon Chun. SentimentArcs: A Novel Method for Self-Supervised Sentiment Analysis of Time Series Shows SOTA Transformers Can Struggle Finding Narrative Arcs, October 2021. URL http://arxiv.org/abs/2110.09454. arXiv:2110.09454 [cs].

MA Daoshan and Zhang Shuo. A discourse study of the Iceberg Principle in *A Farewell to Arms*. *Studies in Literature and Language*, 8(1):80–84, 2014.

T.S. Eliot. *Selected Essays by T. S. Eliot*. Faber & Faber, 1948.

Katherine Elkins. *The Shapes of Stories: Sentiment Analysis for Narrative*. Cambridge University Press, July 2022. ISBN 978-1-00-927040-3 978-1-00-927039-7. doi: 10.1017/9781009270403. URL https://www.cambridge.org/core/product/identifier/9781009270403/type/element.

Katherine Elkins and Jon Chun. Can Sentiment Analysis Reveal Structure in a Plotless Novel?, August 2019. URL http://arxiv.org/abs/1910.01441. arXiv:1910.01441 [cs].

Hady Elsahar and Matthias Gallé. To Annotate or Not? Predicting Performance Drop under Domain Shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1222. URL https://aclanthology.org/D19-1222.

Michael Flor and Swapna Somasundaran. Lexical concreteness in narrative. In Francis Ferraro, Ting-Hao 'Kenneth' Huang, Stephanie M. Lukin, and Margaret Mitchell, editors, *Proceedings of the Second Workshop on Storytelling*, pages 75–80, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3408. URL https://aclanthology.org/W19-3408.

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

Adam Hammond. The double bind of validation: distant reading and the digital humanities' "trough of disillusionment". *Literature Compass*, 14(8):e12402, 2017. ISSN 1741-4113. doi: 10.1111/lic3.12402. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/lic3.12402. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lic3.12402.

C. P. Heaton. Style in *The Old Man and the Sea*. *Style*, 4(1):11–27, 1970. ISSN 0039-4238. URL https://www.jstor.org/stable/42945039. Publisher: Penn State University Press.

Ernest Hemingway. *Death in the Afternoon*. Simon & Schuster, New York, 1996. ISBN 978-0-684-80145-2.

Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 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, June 2020. ISSN 2055-7671. doi: 10.1093/llc/fqz092.

Clayton Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, pages 216–225, 2014. doi: 10.1609/icwsm.v8i1.14550.

Swapnil Jain, Shrikant Malviya, Rohit Mishra, and Uma Shanker Tiwary. 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, December 2017. NLP Association of India. URL https://aclanthology.org/W17-7515.

Roman Jakobson. Linguistics and poetics. In *Linguistics and Poetics*, pages 18–51. De Gruyter Mouton, 1981. doi: 10.1515/9783110802122.18.

Jianbo Gao, H. Sultan, Jing Hu, and Wen-Wen Tung. Denoising Nonlinear Time Series by Adaptive Filtering and Wavelet Shrinkage: A Comparison. *IEEE Signal Processing Letters*, 17(3):237–240, March 2010. ISSN 1070-9908, 1558-2361. doi: 10.1109/LSP.2009.2037773. URL http://ieeexplore.ieee.org/document/5345722/.

Matthew Jockers. A novel method for detecting plot, 2014. URL https://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/.

Matthew Jockers. Revealing sentiment and plot arcs with the syuzhet package, 2015a. URL https://www.matthewjockers.net/2015/02/02/syuzhet/.

Matthew L. Jockers. *Syuzhet: Extract Sentiment and Plot Arcs from Text*, 2015b. URL https://github.com/mjockers/syuzhet.

Evgeny Kim and Roman Klinger. A survey on sentiment and emotion analysis for computational literary studies. *arXiv preprint arXiv:1808.03137*, 2018.

Hoyeol Kim. Sentiment analysis: Limits and progress of the Syuzhet package and its lexicons. *Digital Humanities Quarterly*, 16(2), 2022. URL http://www.digitalhumanities.org/dhq/vol/16/2/000612/000612.html.

Corina Koolen, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. Literary quality in the eye of the Dutch reader: The national reader survey. *Poetics*, 79:1–13, 2020. doi: 10.1016/j.poetic.2020.101439.

Stavroula-Thaleia Kousta, Gabriella Vigliocco, David P. Vinson, Mark Andrews, and Elena Del Campo. The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, 140 (1):14–34, 2011. ISSN 1939-2222, 0096-3445. doi: 10.1037/a0021446. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0021446.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5505. URL https://aclanthology.org/D19-5505.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 246–256, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.22. URL https://aclanthology.org/2021.emnlp-main.22.

Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A. González, and Thamar Solorio. 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, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1114.

Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio. 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, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2042. URL https://aclanthology.org/N18-2042.

Saif Mohammad. 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, June 2011. Association for Computational Linguistics.

URL https://aclanthology.org/W11-1514.

Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia, 2018a. Association for Computational Linguistics. doi: 10.18653/v1/P18-1017. URL http://aclweb.org/anthology/P18-1017.

Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia, July 2018b. Association for Computational Linguistics. URL https://aclanthology.org/P18-1017.

Jan Mukařovský. Standard language and Poetic Language. In Paul L. Garvin, editor, *A Prague School Reader on Esthetics Literary Structure, and Style*, pages 17–30. 1932. Georgetown University Press, 1964. URL https://monoskop.org/images/3/3a/A_Prague_School_Reader_on_Esthetics_Literary_Structure_and_Style_1964.pdf.

Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. 27:16–32, 2018. ISSN 1574-0137. doi: 10.1016/j.cosrev.2017.10.002. URL https://www.sciencedirect.com/science/article/pii/S1574013717300606.

Sianne Ngai. *Ugly Feelings*. Harvard University Press, Cambridge, MA, January 2007. ISBN 978-0-674-02409-0.

Bruno Ohana, Sarah Jane Delany, and Brendan Tierney. A Case-Based Approach to Cross Domain Sentiment Classification. In Belén Díaz Agudo and Ian Watson, editors, *Case-Based Reasoning Research and Development*, Lecture Notes in Computer Science, pages 284–296, Berlin, Heidelberg, 2012. Springer. ISBN 978-3-642-32986-9. doi: 10.1007/978-3-642-32986-9_22.

Emily Öhman. 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, December 2021. NLP Association of India (NLPAI). URL https://aclanthology.org/2021.nlp4dh-1.2.

Emily Öhman and Riikka H. Rossi. Computational exploration of the origin of mood in literary texts. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 8–14, Taipei, Taiwan, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.nlp4dh-1.2.

Yves Peirsman. nlptown/bert-base-multilingual-uncased-sentiment · Hugging Face, 2020. URL https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. The Emotional Arcs of Stories Are Dominated by Six Basic Shapes. *EPJ Data Science*, 5(1):1–12, December 2016a. ISSN 2193-1127. doi: 10.1140/epjds/s13688-016-0093-1. URL https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0093-1.

Andrew J. Reagan, Brian Tivnan, Jake Ryland Williams, Christopher M. Danforth, and Peter Sheridan Dodds. Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. (arXiv:1512.00531), September 2016b. URL http://arxiv.org/abs/1512.00531. arXiv.

Simone Rebora. Sentiment Analysis in Literary Studies. A Critical Survey. *Digital Humanities Quarterly*, 17(2), 2023. URL https://www.digitalhumanities.org/dhq/vol/17/2/000691/000691.html#kim-klinger2018b.

Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, December 2016. ISSN 2193-1127. doi: 10.1140/epjds/s13688-016-0085-1.

Brian Richardson. Linearity and Its Discontents: Rethinking Narrative Form and Ideological Valence. *College English*, 62(6):685–695, 2000. ISSN 0010-0994. doi: 10.2307/379008. URL https://www.jstor.org/stable/379008.

Louise M. Rosenblatt. The literary transaction: Evocation and response. *Theory Into Practice*, 21(4):268–277, 1982. ISSN 0040-5841. URL https://www.jstor.org/stable/1476352.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, February 2020. URL http://arxiv.org/abs/1910.01108. arXiv:1910.01108 [cs].

Thomas Strychacz. "The sort of thing you should not admit": Ernest Hemingway's aesthetic of emotional restraint. In Milette Shamir and Jennifer Travis, editors, *Boys Don't Cry? Rethinking Narratives of Masculinity and Emotion in the U.S.*, pages 141–166. Columbia University Press, 2002. doi: 10.7312/sham12034-009. URL https://www.degruyter.com/document/doi/10.7312/sham12034-009/html.

Annie Swafford. Problems with the Syuzhet Package. *Anglophile in Academia: Annie Swafford's Blog*, March

2015. URL https://annieswafford.wordpress.com/2015/03/02/syuzhet/.

Sayyida Tabinda Kokab, Sohail Asghar, and Shehneela Naz. Transformer-based deep learning models for the sentiment analysis of social media data. *Array*, 14:100157, July 2022. ISSN 2590-0056. doi: 10.1016/j.array.2022.100157. URL https://www.sciencedirect.com/science/article/pii/S2590005622000224.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45:1191–1207, 2013. doi: https://doi.org/10.3758/s13428-012-0314-x.

Albin Zehe, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger, and Fotis Jannidis. Prediction of Happy Endings in German Novels Based on Sentiment Information. In Peggy Cellier, Thierry Charnois, Andreas Hotho, Stan Matwin, Marie-Francine Moens, and Yannick Toussaint, editors, *Interactions between Data Mining and Natural Language Processing*, pages 9–16, Riva del Garda, 2016.

Deyu Zhou, Jianan Wang, Linhai Zhang, and Yulan He. Implicit sentiment analysis with event-centered text representation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6884–6893, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.551. URL https://aclanthology.org/2021.emnlp-main.551.