Comparing Transformer and Dictionary-based Sentiment Models for Literary Texts: Hemingway as a Case-study

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

Pascale Feldkamp
Center for Humanities Computing
Aarhus University, Denmark
pascale.moreira@cc.au.dk

Abstract

The literary domain continues to pose a challenge for Sentiment Analysis methods, due to its particularly nuanced and layered nature. This paper explores the adequacy of different Sentiment Analysis tools – from dictionary-based approaches to state-of-the-art Transformers – for capturing valence and modelling 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, compare Transformer and rule-based systems' scores with human annotations, and shed light on the complexities of analyzing sentiment in narrative texts. Finally, we emphasize the potential of model ensembles.

1 Introduction

Recent years have seen a significant increase in the methods available for Sentiment Analysis (SA). While dictionary-based approaches like VADER (Hutto and Gilbert, 2014) seem to consistently 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). Transformer-based models provide a much richer semantic representation texts, but also display shortcomings (Tabinda Kokab et al., 2022). 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 unexplored. Literary language is particularly intriguing to test SA tools (Chun, 2021), because it often aims to evoke rather than explicitly communicate, operating at multiple narrative levels (Jakobson, 1981; Rosenblatt, 1982; Booth, 1983). In this study, 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-based and Transformer-based SA systems.[1] Hemingway's writing style is known for its emotional subtlety, often described as an "iceberg" or "omissive" writing, that evokes more than it describes: "the emotion is plentiful, though hidden but not exposed" (Daoshan and Shuo, 2014). With its directness and limited use of figurative language (Heaton, 1970), Hemingway avoids "overt emotional display" (Strychacz, 2002) in a way that may pose a particular challenge to SA. Building on the literary analysis tradition that seeks to model sentiment arcs in literary texts (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.

2 Related works

In literary studies, what is often called the "affective turn" (Armstrong, 2014) has led to a stronger focus on sentiment expressed in narrative texts (Ngai, 2007), and SA has often been employed in computational literary studies to profile texts and model the "shape of stories" (Reagan et al., 2016a). To capture meaningful aspects of the reading experience, previous works tested the potential of SA (Alm, 2008; Jain et al., 2017) at the word (Mohammad, 2011, 2018), 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). Certain 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

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

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 219–228

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, modelling sentiment arcs holds 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; Reborа, 2023). For example, dictionary-based methods seem to perform well even on so-called "non-linear" narratives (Richardson, 2000; Elkins and Chun, 2019) although they appear to do poorly on a word-basis (Reagan et al., 2016b). On the other hand, more recent Transformer-based approaches have shown both potential and pitfalls in the analysis of sentiment (Elkins, 2022).

3 Methods

3.1 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. 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. 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.[2] 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. 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 which was discrete in ten categories, we consider correlation measures to be more adequate than categorical inter-annotator agreement measures. A representation of the detrended sentiment arc of each annotator is visualized in the Appendix, along with their detrended mean.

After detrending the arcs, the correlation between the annotators' arcs 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. 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 to elicit higher correlations between models.

3.2 Automatic Annotation

All annotations were performed on a sentence-basis (not considering context).[3]

3.2.1 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).[4]

[2] 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) finished 1, respectively 12 years ago (the BA).

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

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

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.

3.2.2 Dictionary-based models

To compare against Transformers, we chose two dictionary-based approaches: (i) the nltk 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. The dictionary is extracted from 165,000 human coded sentences from contemporary literary novels, developed in the Nebraska Literary Lab (Jockers, 2015b). Both models dictionary- and rule-based, and return continuous scores ranging from -1 (negative) to +1 (positive).

3.3 Detrending 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" 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 dynamics of sentiment arcs (Hu et al., 2020; Bizzoni et al., 2022b). Arcs are based on the second polynomial fit (m=2).

4 Results

To evaluate the models we use the average of the annotators' scores. In Table 1 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.

Our results show that large pretrained Transformers correlate with human judgments on the valence 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 literary sentences (see the Appendix for a detailed plot of raw values). Still, it is notable that the dictionary-based systems outperform half 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 1) - 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 rule-based model, Syuzhet, is most removed from the human evaluation appear to be long sentences with complex semantic interplays, for which it displays a positivity bias. Finally, the sentences with most disagreement between the two models are often sentences that were also difficult for human annotators. In the Appendix we show a small selection of such sentences (Table 4).

When detrending the series of valences, we find that the picture changes: Syuzhet outperforms all Tranformers (Table 1). It is possible that in the case of Syuzhet, errors at the level of raw scores,

	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
Pearson r**, 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
Pearson r**, per annot.**	.88/.71	.70/.69	.92/.85	.87/.81	.62/.70	.90/.97	.96/.93	.95/.93

Table 1: **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. For all correlations, *p-values* < 0.01.

where humans set a negative and Syuzhet a positive score (see Appendix, Fig. 4),[5] are big enough to impact the overall correlation with human annotations, but are still few enough to be "cancelled" out in detrending, so that dictionary-based arcs are the closest to the human arc. The detrending essentially flattens out raw scores, so that scores that are proximate are more alike. In this sense, detrending gives us a pictures of the annotation tendencies at each point of the arc, and smoothens out scores that diverge suddenly from the tendencies.

5 Analysis

Literary language is a challenge to SA due to its subtlety and 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 specially 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 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 two 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.

6 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 Transformer's outputs.

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

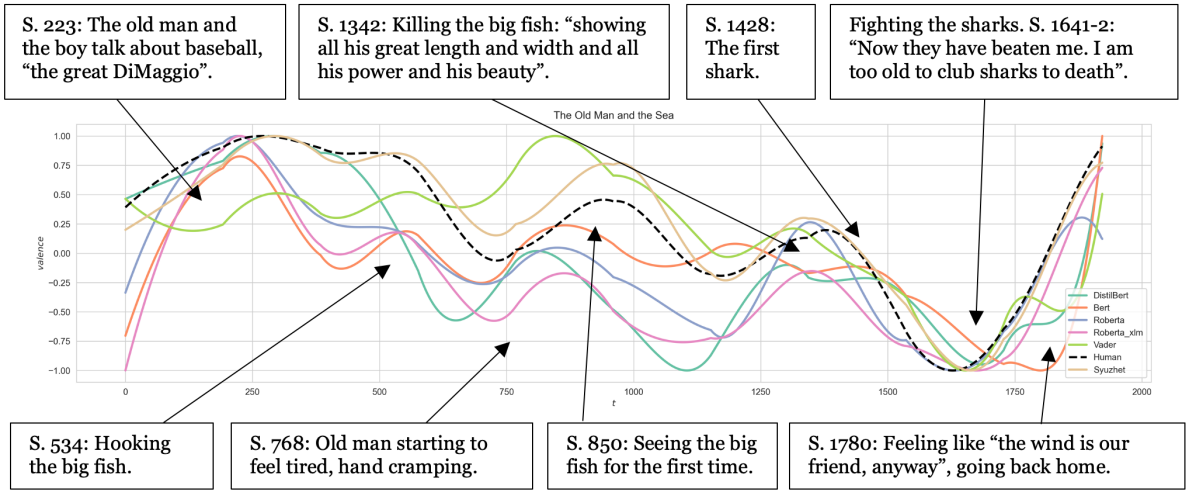

Figure 1: Arcs of *The Old Man and the Sea* based on various methods, with manual annotations of narrative events. The added dashed line represents the mean value of human annotators.

	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 2: Mean difference and standard deviation between human and model valence.

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 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 VADER and Syuzhet. When looking at the detrended version of the arcs, the picture is reversed: despite serious shortcomings of the tool (Kim, 2022), detrended arcs made with the Syuzhet package appear to be the most closely related to the detrended version of human arcs (Fig. 1). 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 1). 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 the human intensities: on average, VADER and Syuzhet have a smaller mean distance from human intensity (as does the roBERTa), and a lower standard deviation (Table 2). [6] 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 opposite scores. 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."*

Limitations

As sentiment annotation is a difficult task, this study has attempted to make the process as robust as possible, and we have sought to make our

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

procedure by various SA methods as transparent as possible. Regardless, identifying sentiment in text is always subjective and difficult to measure, and may be subject to cultural understandings of sentiment expression – which inevitably situates our analysis in the Anglophone cultural context. Moreover, it should be noted that our annotators were academics, and though their annotation may reflect their knowledge of literary devices and language, it also reflects the cultural understandings of a particular class. As a case-study moving towards a comparison and better understanding of sentiment analysis methods, it should also be noted that the analysis limits itself to one, and a particularly canonic, Anglophone literary novel. We trust that any interpretation of our findings will have these limitations in mind.

References

Huwail J. Alantari, Imran S. Currim, Yiting Deng, and Sameer Singh. 2022. 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.

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

Nancy Armstrong. 2014. The Affective Turn in Contemporary Fiction. *Contemporary Literature*, 55(3):441–465. Publisher: [Board of Regents of the University of Wisconsin System, University of Wisconsin Press].

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification.

Yuri Bizzoni, Pascale Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023. 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, 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*, NLP4DH.

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).

Wayne C. Booth. 1983. *The Rhetoric of Fiction*, 2nd edition edition. University of Chicago Press, Chicago.

Katherine Bowers and Quinn Dombrowski. 2021. Katia and the Sentiment Snobs. Blog: Datasitter's Club.

Jon Chun. 2021. SentimentArcs: A Novel Method for Self-Supervised Sentiment Analysis of Time Series Shows SOTA Transformers Can Struggle Finding Narrative Arcs. ArXiv:2110.09454 [cs].

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

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

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

Hady Elsahar and Matthias Gallé. 2019. 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. Association for Computational Linguistics.

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

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

C. P. Heaton. 1970. Style in *The Old Man and the Sea*. *Style*, 4(1):11–27. Publisher: Penn State University Press.

Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2020. Dynamic evolution of sentiments in Never Let Me Go: Insights from multifractal theory and its implications for literary analysis. *Digital Scholarship in the Humanities*, 36(2):322–332.

Clayton Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

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.

Roman Jakobson. 1981. Linguistics and poetics. In *Linguistics and Poetics*, pages 18–51. De Gruyter Mouton.

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. 2014. A Novel Method for Detecting Plot. Matthew L. Jockers Blog.

Matthew Jockers. 2015a. Revealing Sentiment and Plot Arcs with the Syuzhet Package. Matthew L. Jockers Blog.

Matthew L. Jockers. 2015b. *Syuzhet: Extract Sentiment and Plot Arcs from Text*.

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

Hoyeol Kim. 2022. Sentiment analysis: Limits and progress of the Syuzhet package and its lexicons. *Digital Humanities Quarterly*, 16(2).

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*, 79:1–13.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. pages 34–41.

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.

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

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

Sianne Ngai. 2007. *Ugly Feelings*. Harvard University Press, Cambridge, MA.

Bruno Ohana, Sarah Jane Delany, and Brendan Tierney. 2012. A Case-Based Approach to Cross Domain Sentiment Classification. In *Case-Based Reasoning Research and Development*, Lecture Notes in Computer Science, pages 284–296, Berlin, Heidelberg. Springer.

Emily Öhman and Riikka H. Rossi. 2022. 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. Association for Computational Linguistics.

Yves Peirsman. 2020. nlptown/bert-base-multilingual-uncased-sentiment · Hugging Face.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016a. The Emotional Arcs of Stories Are Dominated by Six Basic Shapes. *EPJ Data Science*, 5(1):1–12.

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

Simone Rebora. 2023. Sentiment Analysis in Literary Studies. A Critical Survey. *Digital Humanities Quarterly*, 17(2).

Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29.

Brian Richardson. 2000. Linearity and Its Discontents: Rethinking Narrative Form and Ideological Valence. *College English*, 62(6):685–695.

Louise M. Rosenblatt. 1982. The literary transaction: Evocation and response. *Theory Into Practice*, 21(4):268–277.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv:1910.01108 [cs].

Thomas Strychacz. 2002. "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.

Annie Swafford. 2015. Problems with the Syuzhet Package. *Anglophile in Academia: Annie Swafford's Blog*.

Sayyida Tabinda Kokab, Sohail Asghar, and Shehneela Naz. 2022. Transformer-based deep learning models for the sentiment analysis of social media data. *Array*, 14:100157.

A Appendix

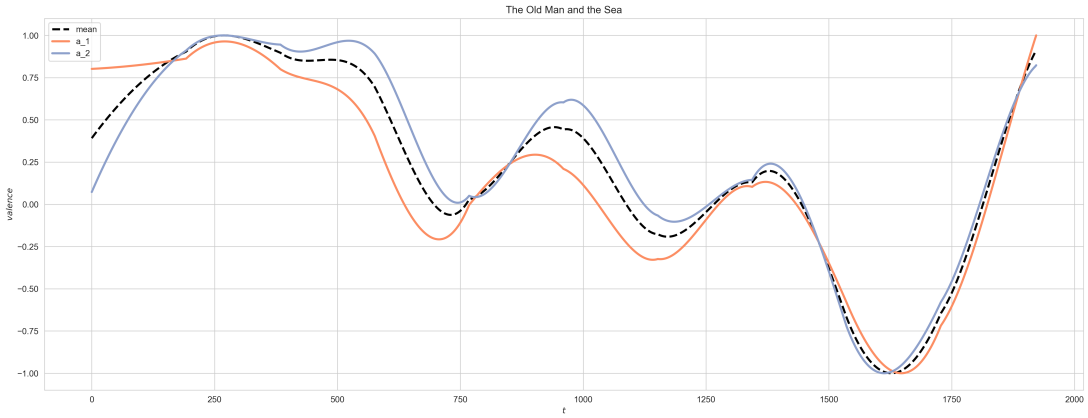

Figure 2: Arc of *The Old Man and the Sea* based on annotator (n=2) values. The dashed line represents the mean value of annotators.

	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.

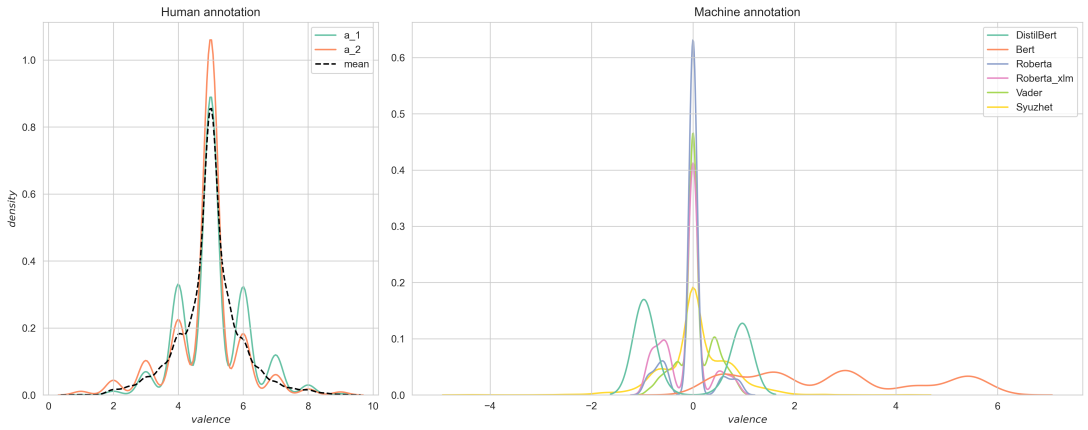

Figure 3: Kernel density plots visualize the distributions of values (0 or neutral being the most common). Note that value ranges differ: the BERT model, for example, assigns valence on a 5-point scale, while human annotators could assign any (round) value between 0 and 10.

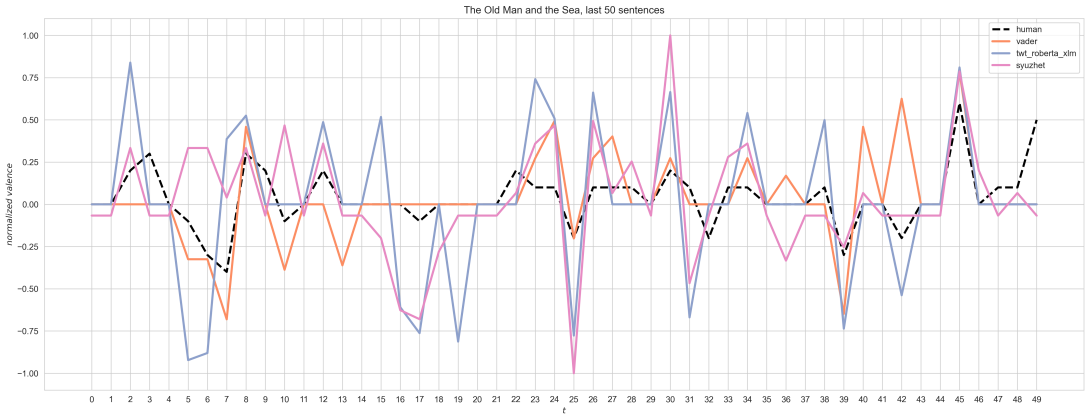

Figure 4: Arc of the last 50 sentences of *The Old Man and the Sea* with on transformer and dictionary-based annotation. The added dashed line represents the mean value of human annotators. Note that sentences like [5]: "I am not lucky" and [10] "I do not care" are systematically misjudged as positive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

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 Syuzhet (central rows the table). Roberta XLM is most off track for short, relatively ambiguous sentences; 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 the table. These sentences are often also complex for human annotators to judge.

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 5: To give a short overview of the models' comparative performance, we present the sentences that elicited the highest score for each model.