Below the Sea (with the Sharks): Probing Textual Features of Implicit Sentiment in a Literary 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

Literary language presents an ongoing challenge for Sentiment Analysis (SA) due to its complex, nuanced, and layered form of expression. It is often suggested that effective literary writing is evocative, operates beneath the surface and understates emotional expression. To explore features of implicitness in literary expression, this study takes Ernest Hemingway's *The Old Man and the Sea* as a case-study, focusing specifically at implicit sentiment expression in this text. We examine sentences where automatic sentiment scoring shows substantial divergences from human sentiment annotation, and probe these sentences for distinctive traits. We find that sentences where humans perceived a strong sentiment while models did not are significantly lower in arousal and higher in concreteness than sentences where humans and models were more aligned, suggesting the importance of simplicity and concreteness for implicit sentiment expression in literary prose.

1 Introduction

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 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, though critics often rely on terms like emotional "evocativeness" and "understatement" to describe writing styles (Strychacz, 2002; Daoshan and Shuo, 2014). It is far from clear whether implicit, evocative and expressive strategies can be reliably tracked in text and whether more implicit types of narration display linguistically recognizable marks.

In this study, we use *The Old Man and the Sea*, often considered the exemplary masterpiece of Ernest Hemingway, as a case study for exploring such implicitness.[1] Hemingway's writing style is known for its emotional subtlety and is characterized (also by Hemingway himself) by its "iceberg" (Hemingway, 1996), or "omissive" technique, where: "the emotion is plentiful, though hidden and not exposed" (Daoshan and Shuo, 2014). Moreover, Hemingway's style is direct and limited in use of figurative language (Heaton, 1970). It thus avoids "overt emotional display", presenting actions and situations that *imply* emotions, and leave their inference up to the reader (Strychacz, 2002). As such, it may be that Hemingway's "omissive" writing can be tracked by looking at the amount and intensity of emotion expressions detectable in the text itself, comparing this to how "expressive" the text is perceived by readers.

2 Related works

Literary language may convey emotions in a variety of ways beyond simply using words directly associated with emotional states (e.g., "happy").In the case of Hemingway, the apparent aversion to "emotional display and rhetorical overflow" in his prose has been linked to the Modernists' and New Critics' emphasis on *concreteness* over abstraction (Strychacz, 2002). A key example of this perspective is Brooks and Warren (1976)'s seminal description of poetry as "incorrigibly particular and concrete – not general and abstract". The connection of concreteness to emotional expression is continually formalized in modern literary theory, also with regard to prose, where the most prominent concept is probably that of the *objective correlative* of T.S. Eliot. Eliot defined it as "a set of objects, a situation, a chain of events which shall be the formula of [a] particular emotion" (Eliot, 1948),

[†]The authors contributed equally to this work.

[1]The annotated text is available at: https://github.com/PascaleFMoreira/Annotated_Hemingway

Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language, pages 54–61
March 21, 2024 ©2024 Association for Computational Linguistics

suggesting a focus on concrete objects and actions over explicit emotion expression as the effective method for communicating emotion in literature. In support of this idea, Auracher and Bosch (2016) indicate that the concreteness of literary language impacts the emotional engagement of readers and their experiences of literary suspense.

We concentrate our study on implicitness in the expression and readers' experience of sentiments in *The Old Man and the Sea*. Sentiment Analysis (SA) has become an increasingly central method for computational literary studies research (Rebora, 2023), often used as a tool to gauge the sentiment arcs of novels (i.e., the consecutive highs and lows of sentiment throughout a narrative)(Jockers, 2014; Reagan et al., 2016) also in connection with assessing reader appreciation (Bizzoni et al., 2023). While divergences between human and model SA scores generally indicate shortcomings in SA methods, we suggest that such divergences may also be informative – both for model improvement and for gaining a deeper understanding of sentiment expression in literary texts – if we test whether certain textual features characterize such instances. First, we seek to find sentences where human sentiment annotation diverges from model scores, the latter of which may not capture implicit or omissive sentiment as well (Zhou et al., 2021; Li et al., 2021). Then, we test whether these sentences of implicit sentiment expression can be told apart from other by certain features, the choice of which are informed by the mentioned literary theory and descriptions of implicitness in Hemingway's style: the mean valence,[2] arousal,[3] and dominance,[4] as well as their mean concreteness.[5]

3 Method

In this preliminary analysis of implicit or omissive writing, we focus on the sentiments in *The Old Man and the Sea*. As noted, 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. 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 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 the certain composition that Hemingway creates, but will not appear emotionally charged when observed as isolated words. To create a subset of such sentences that appear powerful to human readers but may not be so for automatic annotation systems, we used the distance between SA models' and humans' annotations of sentences. We thus operationalized "implicit sentiment" as those cases in which human readers perceived sentimental charge (whether positive or negative), but where models did not, selecting all such sentences. We proceeded by the following steps:

3.1 Annotation, scoring and selection

1) Two independent human annotators scored each sentence of the novel on a 1-10 sentiment scale. The annotators were instructed to avoid rating how a sentence made them feel but assess the valence of each sentence, without overthinking the story's narrative, reducing – as far as possible – contextual interpretation. We thereafter assigned each sentence of the novel the mean annotator score for each sentence.[6] Both annotators had extensive experience of literary analysis, and hold degrees in

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

[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).

[6]The Spearman correlation between annotators is 0.65.

literature.[7] Annotators worked independently, not discussing nor changing their scores.

2) There are a variety of SA methods from machine learning to dictionary-based approaches, each displaying advantages and shortcomings (Öhman, 2021). (Reagan et al., 2017). More recent Transformer-based approaches have shown both potential and pitfalls in SA for literary texts (Elkins, 2022), so that an ensemble of models has been suggested (Elkins, 2022). We used several SA models, transformer- and dictionary-based, to score the same book for valence. Our chosen models for annotation on a sentence-base were:
(i) The **VADER dictionary** (Hutto and Gilbert, 2014), arguably the most widespread dictionary-based method for SA.
(ii) The **Syuzhet dictionary** (Jockers, 2014), extracted from 165,000 human coded sentences of contemporary literary novels.[8]
(iii) **roBERTa base**, fine-tuned for SA on tweets (Barbieri et al., 2020).[9]

3) Excluding mid-valued sentences, we selected all the sentences that the human annotators scored as having some sentimental charge (all sentences scoring lower than 5 or higher than 6). Since the human readers did detect some sentiment in these sentences, they are candidates for implicit sentiment expression. This subset accounted for less than half of the sentences of the novel: a total of 835 out of 1923 sentences.

4) Of this subset, we selected only those sentences that did *not* elicit a strong sentiment score from the SA models: we only kept sentences which normalized absolute score was smaller than 0.1 in *all three models*. 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 101 sentences in what we call the "implicit" group (Fig. 1).

5) For comparison, we selected sentences where

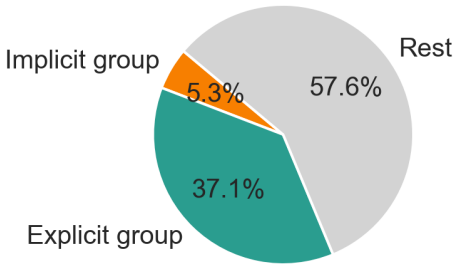

Figure 1: Division of sentences of *The Old Man and the Sea* into groups of: 101 sentences where human and model sentiment scoring diverged significantly, and 714 sentences where it converged.

human and models were more aligned in their sentiment scoring, what we call the "explicit" group (Fig. 1). These are sentences where both humans and models found either a positive or a negative sentiment (above an absolute 0.1), and agree on the sentiment direction (positive/negative).

We then compared 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,[10] and concreteness.[11] Finally, we used a Mann-Whitney U test to examine differences between the groups (to further validate our results, we performed additional tests; see the Appendix for an overview of these results).

4 Results

Our selected group of 101 sentences represent 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

[7] Both were academics, male and female, at ages 31 and 34, who were non-native but very proficient English speakers, and who finished a literature degree more than 2 years ago.

[8] Developed by Matthew L. Jockers in the Nebraska Literary Lab (Jockers, 2015).

[9] Note that we converted the categorical Transformer output is to continuous SA scores by using the confidence score of roBERTA's labels as a proxy for sentiment intensity. If the model classifies a sentences as *positive* with a confidence of, for example, 0.89, we interpret it as a valence score of +0.89 for this sentence, and so on. Note that we converted scores of the *neutral* category to 0.0. This procedure of translating SA Transformer output to a continuous scale is detailed in Bizzoni and Feldkamp (2023).

[10] We used the VAD lexicon (Mohammad, 2018) to retrieve the valence, arousal and dominance scores for each word, averaging scores over each sentence: `https://saifmohammad.com/WebPages/nrc-vad.html`

[11] 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`

		Valence	Dominance	Arousal	Concreteness
Word-based	Implicit	0.581 ±0.163	0.476 ±0.152	0.379 ±0.155	2.759 ±1.174
	Explicit	0.559 ±0.230	0.482 ±0.170	0.433 ±0.189	2.677 ±1.146
	MWU test	724.263	696.247	582.587*	6196.182*
Sentence-based	Implicit	0.596 ±0.109	0.495 ±0.118	0.401 ±0.106	2.732 ±0.37
	Explicit	0.572 ±0.164	0.494 ±0.110	0.446 ±0.110	2.649 ±0.328
	MWU test	39.308	36.558	25.146*	45.660*

Table 1: Mean and st.d. feature values of the implicit and explicit groups, where features are computed, respectively, on a **word** basis (rows above) and on a **sentence** basis (rows below), as well as the results of the MWU test between the groups in each setup. In the implicit group: sentences perceived non-neutral by humans but neutral by models (below an absolute score of .1); in the explicit group: sentences where human and models were more aligned. * p-value < 0.05.

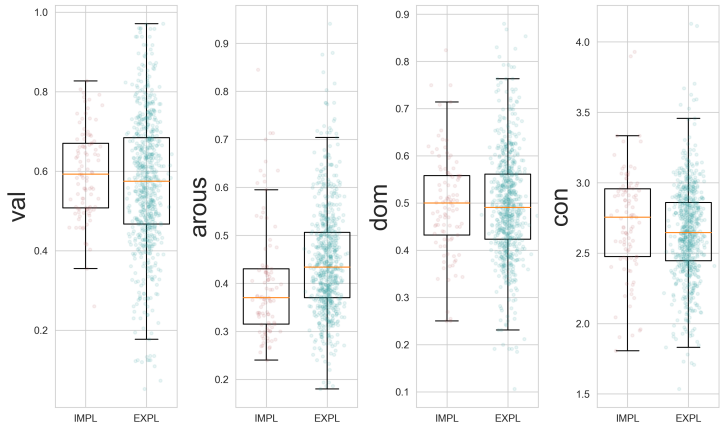

Figure 2: Boxplots comparing implicit (n=101) and explicit (n=714) groups of sentences by scores of each of the four features.

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 tokenized all the sentences using WordNet's lemmatizer. For each sentence lemmatized, we computed the average Valence, Arousal and Dominance using the NRC-VAD-Lexicon. These measures attempt to position a word in a three-dimensional sentiment 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 concreteness 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 (101 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" sentence – as can be seen in Table 1. Two of the four feature dimensions appear to be significant in the sentences that implicitly ex-

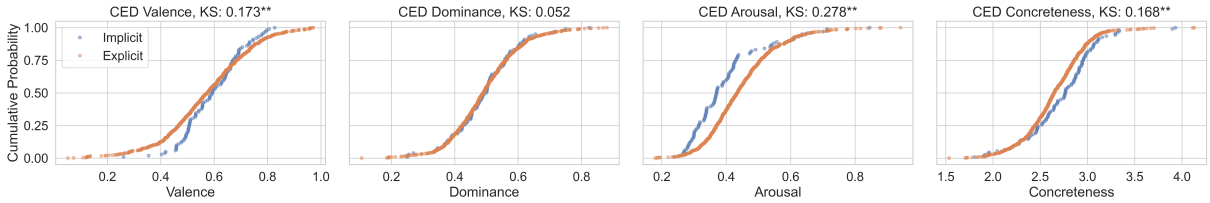

Figure 3: Cumulative Empirical Distribution (CED) of features per group and statistics of the two-sample Kolmogorov-Smirnov test (KS) for goodness of fit (on top). **p-value < 0.01.

press a sentiment: their level of concreteness and their level of arousal.[12] 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. The statistical significance of the two relevant categories is even stronger when they are measured on a sentence- rather than word base (Table 1).

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 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*.[13]

To further validate these results, we examined the distribution of our data, performing the The Kolmogorov-Smirnov (KS) test[14] on the empirical cumulative distribution of the groups (Fig. 3). 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. 3).[15]

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

[13]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.

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

[15]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.

5 Conclusions and Future Works

In examining human and model sentiment annotations in *The Old Man and the Sea*, 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 humans in these cases pick up on contextual information not available to the models, 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* technique evoke a perception of sentiments in human readers without any explicit emotional reference and without using words directly associated to emotional states. 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. 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. Finally, 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.

Limitations

We want to underline that the present work is an examination of one work of fiction only, also due to the fact that large-scale annotation of texts is a complex and costly undertaking. Moreover, as this study examined and drew conclusions from what can be considered a particularly "canonical" text of Western literary production, we note that it situates the study in (prestigious) Western literary culture, where certain norms of writing style may prevail. As such, further study is needed to draw more far-reaching conclusions, and the present study should be considered only a step toward a more comprehensive examination of implicit sentiment expression in literary fiction.

References

Jan Auracher and Hildegard Bosch. 2016. Showing with words: The influence of language concreteness on suspense. *Scientific Study of Literature*, 6(2):208–242.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

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

Yuri Bizzoni, Pascale Moreira, 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.

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

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

Cleanth Brooks and Robert Penn Warren. 1976. *Understanding Poetry*. Holt, Rinehart and Winston, New York.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

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.

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

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

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

Ernest Hemingway. 1996. *Death in the Afternoon*. Simon & Schuster, New York.

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.

Matthew Jockers. 2014. A novel method for detecting plot.

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

Stavroula-Thaleia Kousta, Gabriella Vigliocco, David P. Vinson, Mark Andrews, and Elena Del Campo. 2011. The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, 140(1):14–34.

Xiaotao Li, Shujuan You, Yawen Niu, and Wai Chen. 2021. Learning embeddings for rare words leveraging Internet search engine and spatial location relationships. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 278–287, Online. 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.

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

Andrew J. Reagan, Christopher M. Danforth, Brian Tivnan, Jake Ryland Williams, and Peter Sheridan Dodds. 2017. Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. *EPJ Data Science*, 6(1):1–21.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):1–12.

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

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.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45:1191–1207.

Deyu Zhou, Jianan Wang, Linhai Zhang, and Yulan He. 2021. Implicit sentiment analysis with event-centered text representation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6884–6893, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

Test	**Valence**	**Dominance**	**Arousal**	**Concreteness**
MWU	724.263.0	696.247	582.588**	619.618*
T-test	1.8548	-0.7048	-5.6028**	2.3346*
T(W)-test	2.4353	-0.7703	-6.4615**	2.2885*
T(W)-test, 100 permutations	2.4353	-0.7703	-6.4615**	2.2885**
MWU	39.308	36.558	25.146**	45.660**
T-test	1.4119	0.1118	-3.8562**	2.4327*
T(W)-test	1.8972	0.1148	-3.6547**	2.2209*
T(W)-test, 100 permutations	1.8972	0.1148	-3.6547**	2.2209*

Table 2: Additional test between groups where features were calculated per word (above) and sentence (below). Regarding the t-test, we also ran it without assuming equal population variance, we thus performed a Welch's (W) t-test with and without permutations (n=200). * p-value < 0.05, ** p-value < 0.05. Note that the p-value for concreteness tends to be higher than for arousal (even if in all cases < 0.05, which might indicate that the difference between groups are more strongly distinguished by arousal.

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

Table 3: 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=714+101), using *the difference between human and roBERTa sentiment score* as the dependant 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.