Deciphering Emotional Landscapes in the Iliad: A Novel French-Annotated Dataset for Emotion Recognition

Davide Picca, John Pavlopoulos
University of Lausanne, Athens University of Economics and Business
Switzerland, Greece
davide.picca@unil.ch, annis@aueb.gr

Abstract

One of the most significant pieces of ancient Greek literature, the Iliad, is part of humanity's collective cultural heritage. This work aims to provide the scientific community with an emotion-labeled dataset for classical literature and Western mythology in particular. To model the emotions of the poem, we use a multi-variate time series. We also evaluated the dataset by means of two methods. We compare the manual classification against a dictionary-based benchmark as well as employ a state-of-the-art deep learning masked language model that has been tuned using our data. Both evaluations return encouraging results (MSE and MAE Macro Avg 0.101 and 0.188 respectively) and highlight some interesting phenomena.

Keywords: Emotion Recognition, Computing in classical studies, Digital Humanities

1. Introduction

The Iliad has long captivated scholars across various disciplines, including personality psychology and emotional research (Bolen, 2004; Caldwell, 1993; Egloff et al., 2019). Its intricate emotional fabric renders it an exemplary subject for exploring the evolution of human psychology. However, the lack of extensive, emotion-annotated datasets in classical literature has significantly hindered such inquiries.

This study aims to propel the domain of emotion analysis in classical literature forward, with a particular emphasis on the ancient Greek epic, "The Iliad." To achieve this, we present the first publicly available, emotion-annotated dataset of the Iliad[1]

This dataset constitutes a groundbreaking resource for the scientific community, facilitating nuanced investigations into the emotional intricacies of classical texts.

Our dataset not only addresses this void but also offers a multivariate sentiment time series. We utilize this dataset to assess a dictionary-based benchmark and to train a cutting-edge deep learning masked language model for sentiment analysis. The empirical findings, elaborated upon in Sections 3 and 4, unveil compelling patterns and phenomena within the Iliad's emotional landscape.

2. Related work

2.1. Theory of emotions

Researchers have advanced multiple emotional models to elucidate the nature and manifestations of emotions. Plutchik's theory, which stems from Russell's circumplex model of emotions (Russell, 1980), posits over 90 definitions of emotion. In Russell's model, valence and arousal serve as the horizontal and vertical axes, respectively. Plutchik extends this by likening his emotional framework to a color wheel, where proximate emotions are closely situated, and antithetical emotions are positioned 180 degrees apart (Plutchik, 2001). He further introduces a third dimension to represent emotional intensity, thereby transforming the circumplex into a cone-shaped structure. Building on Plutchik's work, Cambria (Cambria et al., 2012) formulated the Hourglass of Emotions model. This model employs eight emotions—pleasantness, fear, eagerness, sadness, calmness, anger, disgust, and joy—as descriptors for affective states and serves as the cornerstone of our research.

2.2. Emotion Recognition in classical literature

The analysis of emotions, along with sentiment analysis, is certainly one of the most explored fields (Alswaidan and Menai, 2020).

Emotion recognition and sentiment analysis are distinct but interconnected fields within affective computing (Cambria and Hussain, 2012; Liu, 2012a). While sentiment analysis has been widely applied in various domains, including the humanities (Kim et al., 2010; Sebe et al., 2005; De Greve, Lore and Martens, Gunther and Van Hee, Cynthia and Singh, Pranaydeep and Lefever, Els, 2021; Sprugnoli et al., 2021; Yeruva et al., 2020), emotion recognition in classical texts like Greek or Latin texts remains less explored (Pavlopoulos et al., 2022; Picca and Richard, 2023). Our work contributes by annotating ten books of the Iliad, thereby providing

[1] (Omitted for review)

LREC-COLING 2024, pages 4462–4467
20-25 May, 2024.

a unique resource for the scientific community.

3. Empirical Analysis

3.1. Building the dataset

The dataset was carefully assembled to feature excerpts from specific books of the Iliad, chosen for their richness in dialogic and emotional content as identified by scholars in the field of Greek literature.[2] In particular, Book 1 and Books 16 to 24 are widely acknowledged for containing more dialogues and greater emotional expressiveness.

The annotation framework was structured to allow for the presence of multiple emotions within individual excerpts, thus enabling a multi-label classification approach. In instances where multiple emotions were annotated for a single excerpt, we explored the possibility of distilling these annotations to a predominant emotion to facilitate certain types of analysis. This reduction was approached with consideration of the inherent complexity of emotional expressions and the potential nuances lost in such a simplification. The choice to distill complex emotional annotations to single emotions was made judiciously, with an understanding of how this might impact both the analytical depth of our dataset and its utility in training more nuanced emotion recognition models. The taxonomy of emotions for annotation was adapted from Cambria's work (Cambria et al., 2012), as elaborated in Section 2.1, and was selected for its comprehensive scope and empirical validation in psychological studies.

Twenty-three annotators participated in this study and were randomly assigned to one of the twenty-four books. Each annotator read the entire book and annotated the emotions expressed in any excerpt of direct speech, excluding the narrator's text. Excerpts were annotated by a minimum of one and a maximum of four annotators, each working independently.

The resulting dataset comprises 468 extracted excerpts from ten books, each annotated with zero or more of the eight specified emotions.[3] The average length of the excerpts is 503 characters, ranging from 35 to 3,571 characters. As shown in Figure 1, the distribution is notably skewed towards the categories of "eagerness," "anger," "sadness," and "calmness." This skewness likely arises from the blurred boundaries that define these emotions compared to the more clearly delineated categories of "fear," "joy," "disgust," and "pleasure."

Emotion detection poses a greater challenge than sentiment classification due to the complexity and variety of emotions. While sentiment classification generally involves assigning a positive or negative value to text, emotion detection necessitates the recognition of a broader spectrum of emotions, including nuanced ones. This complexity is evident in Figure 3, which shows the diversity of emotions present in a single book and underscores the subjectivity involved in their detection. The low inter-annotator agreement for specific emotions (Fig. 2) further attests to this complexity. These phenomena are discussed in greater detail in Sections 3.2 and 3.3.

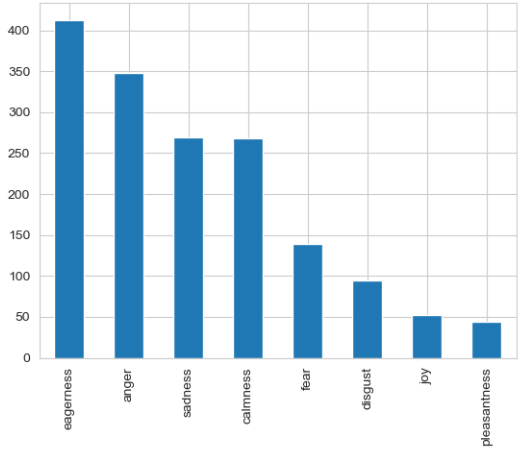

Figure 1: The overall frequency of each emotion across the entire dataset.

3.2. Inter-annotator Agreement

Upon grouping the codes by excerpt, we observe that 35.82% of the excerpts were assigned an emotion by a single annotator, 30.70% by two annotators, 24.73% by three, and 8.74% by four annotators.

The primary aim of our annotation task was to capture emotions as they were intended by Homer, mindful of the nuances introduced through translation into French and the subjective interpretations of individual annotators. Given the variable number of annotators for each text, we chose Krippendorff's alpha (Krippendorff, 2013) as the metric for measuring inter-annotator agreement.[4] Krippendorff's alpha ranges from 0 to 1, with higher values indicating better inter-rater reliability. Examining the results for each individual book, as depicted in Figure 2, reveals that certain books (e.g., 18, 19, and 22) yield higher inter-annotator agreement compared to others (e.g., 1, 16, and 20).

We find that annotating author-intended emotions is generally more challenging than identifying reader-perceived emotions, corroborating ex-

[2] https://www.gutenberg.org/ebooks/14285

[3] Our twenty-three annotators could not cover all twenty-four books, so we opted for assigning at least two annotators per book.

[4] The percentage agreement across all eight emotions was found to be *0.263*.

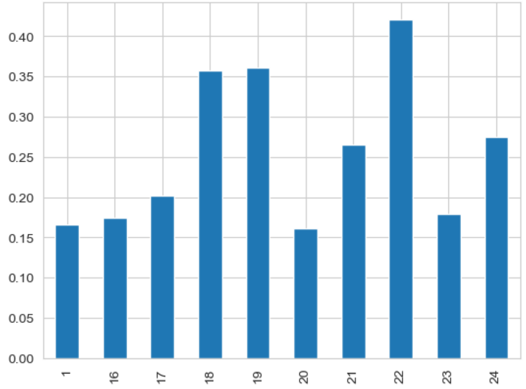

Figure 2: Krippendorff's alpha per book, indicating significant variability, primarily due to the difficulty in reaching consensus on emotions.

isting literature that suggests inconsistent emotional perception among readers (Kajiwara, Tomoyuki and Chu, Chenhui and Takemura, Noriko and Nakashima, Yuta and Nagahara, Hajime, 2021; Pavlopoulos et al., 2022). However, this is not universally applicable; books 15, 19, and 22 exhibit higher agreement, indicating that the effectiveness of annotation guidelines may differ depending on the specific book.

3.3. The Ground Truth

To generate a machine-readable target value for our classification task, we calculated a score for each emotion in each text within the dataset. This score represents the fraction of annotators who labeled the given emotion in that specific text. Figure 3 displays the scores for the eight emotions across all texts in the dataset, revealing that the emotions of "pleasantness," "fear," "disgust," and "joy" are often present at lower levels compared to other emotions. Utilizing majority voting allows us to assign a single emotion label to each instance in the dataset, thereby creating a target value suitable for training and evaluating machine learning models.

4. Experiments and Evaluation

Our dataset serves dual purposes: evaluating existing emotion classifiers and training new ones. To investigate these capabilities, we conducted two distinct experiments. The first experiment involved a comparison between manual and automatic labeling, as suggested by (Abdaoui et al., 2017), a French adaptation of the English NRC-Lex dataset (Mohammad and Turney, 2013). In the second experiment, we fine-tuned Multilingual-BERT (Devlin et al., 2019) specifically for emotion classification.

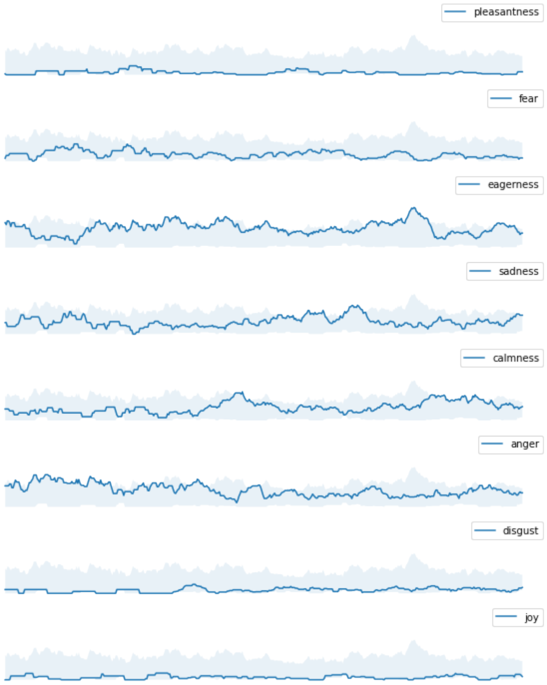

Figure 3: The ground-truth emotion signals in the poem. Each signal illustrates the fraction of annotators (vertically, from 0 to 1) who labeled the respective emotion per text, with texts arrayed horizontally. The shaded background represents the minimum (lower) and maximum (higher) fraction found for any emotion in that particular text.

4.1. FEEL Experiment

In the initial experiment, we utilized our dataset to evaluate an emotion classifier based on the French Expanded Emotion Lexicon (FEEL) (Abdaoui et al., 2017). FEEL is an enriched French lexicon that accounts for both polarity and emotion, developed through semi-automatic translation and synonym expansion of the English NRC Word Emotion Association Lexicon (NRCEmoLex) To conduct this experiment, we matched words from our dataset's excerpts against the FEEL lexicon's categorized words, assigning always an emotion label based on the highest frequency of emotion-associated words within each excerpt. We then compared these automatic classifications with the manual annotations provided in our dataset to assess the concordance and discrepancies.

The results are visualized through a confusion matrix 4, highlighting the alignment and mismatches between the FEEL-based classifications and our manual annotations. This allowed us to observe the effectiveness of lexicon-based emotion detection on classical literature and to identify potential areas where the FEEL lexicon could be further refined for better alignment with human-annotated emotions. The FEEL algorithm's emo-

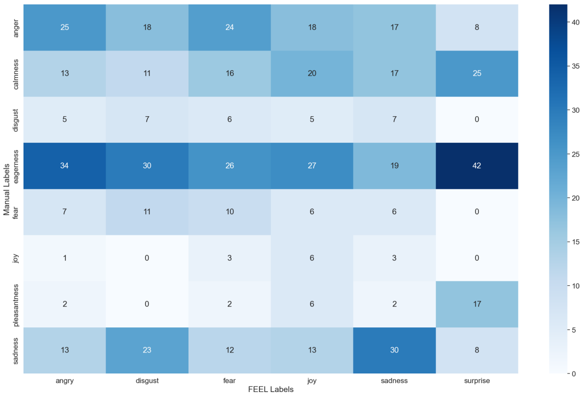

Figure 4: Confusion matrix for FEEL classification, with results displayed in percentages.

tional categories do not align perfectly with those based on Plutchik's theory, as evident in Figure 4. We refrained from aggregating these categories to highlight specific phenomena. For instance, FEEL often identifies "eagerness" as the dominant category, even though this category encapsulates a range of emotions with ambiguous boundaries. Another noteworthy observation is the frequent association of FEEL's "surprise" category with "pleasantness," "calmness," and "eagerness" (see Figure 4). This aligns with psychological literature, confirming that the emotion of surprise can be a composite of pleasure, calmness, and eagerness (Ortony et al., 1990).

4.2. Multilingual-BERT Experiment

For the second experiment, we employed our dataset to fine-tune Multilingual-BERT (Devlin et al., 2019) for emotion recognition. We utilized the "distilbert-base-multilingual-cased" model, a pre-trained masked language model capable of handling multiple languages. The maximum input sequence length was set to 500, and a batch size of 32 was used for training. We employed the "autofit" method with a learning rate of 1e-4 for automatic learning rate adjustment. Early stopping was implemented to halt training after 16 epochs when the validation loss ceased to improve.

Multilingual-BERT, pre-trained on monolingual corpora from 104 languages, achieves state-of-the-art performance across various multilingual tasks according to (Pires et al., 2019). These configurations allowed us to train an efficient, high-performing model on our dataset.

Given the skewed nature of our dataset, as noted in Section 3.1, we applied an over-sampling algorithm to minority classes to ensure representative data. This algorithm, proposed by (Menardi and Torelli, 2014), involves random selection of samples from minority classes, with replacement, to enhance their representation in the training dataset. This strategy balances the class representation, ensuring that the dataset accurately reflects the broader population.

	MSE	MAE
EAGERNESS	0.148	0.234
CALMNESS	0.117	0.206
ANGER	0.058	0.152
DISGUST	0.150	0.227
JOY	0.043	0.130
FEAR	0.127	0.228
PLEASANT	0.138	0.222
SADNESS	0.025	0.101
AVG	0.101	0.188

Table 1: MAE and MSE per emotion and macro-averaged of Multilingual-BERT.

In this study, MAE serves as a straightforward metric for gauging average model performance, while RMSE offers sensitivity to outliers, essential for capturing extreme emotional states. As illustrated in Table 1, our findings align with those of other researchers in the field (Pavlopoulos et al., 2022). The current experiment's outcomes echo those of a previous one (see Section 4.1), indicating that the emotion "eagerness" is particularly challenging to classify. This is attributed to the ambiguous definition of "eagerness," making it difficult to distinguish from other emotions. This consistency across different experiments and datasets suggests that the challenge in classifying "eagerness" is not an isolated issue. Conversely, emotions with clear definitions, such as "joy," "anger," and "sadness," are more readily identifiable, even by machine algorithms.

5. Discussion and Conclusion

The realm of emotions is inherently complex, necessitating a nuanced approach due to the unique characteristics of emotional responses and the challenges they present for both quantification and cross-domain comparison (Kim and Klinger, 2018). In this paper, we introduce a dataset annotated for emotions, created by native French speakers reading the Iliad. We cautiously compare this dataset with existing work, particularly the Modern Greek annotations by Pavlopoulos et al. (Pavlopoulos et al., 2022). Although the lack of a standardized methodology and verse-level analysis hampers comprehensive comparison, our study offers valuable insights. We demonstrate that both lexicon-based and deep learning-based approaches are effective for emotion classification in this context. Looking ahead, our research aims to extend the dataset to encompass additional books from both the Odyssey and the Iliad. Such an expansion will facilitate a more thorough exploration of the emotional landscape in these ancient texts and enable cross-cultural and cross-lingual studies. Existing research supports the notion that emotional responses remain consistent across different translations of the same work (Gygax et al., 2010; Hamby et al., 2022). Consequently, our ongoing and future work holds significant potential for scholars interested in the cultural, historical, and emotional facets of classical literature.

References

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. FEEL: A French Expanded Emotion Lexicon. *Language Resources and Evaluation*, 51(3):833–855.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

Jean Shinoda Bolen. 2004. *Goddesses in Everywoman: Powerful Archetypes in Women's Lives*, 1st quill ed edition. Quill, New York.

Richard Caldwell. 1993. *The Origin of the Gods: A Psychoanalytic Study of Greek Theogonic Myth*. Oxford Univ. Press, New York, N.Y.

Erik Cambria and Amir Hussain. 2012. *Sentic computing: A common-sense-based framework for concept-level sentiment analysis*. Springer.

Erik Cambria, Andrew Livingstone, and Amir Hussain. 2012. The Hourglass of Emotions. In *Cognitive Behavioural Systems*, Lecture Notes in Computer Science, pages 144–157, Berlin, Heidelberg. Springer. 147 citations (Crossref) [2022-12-01].

De Greve, Lore and Martens, Gunther and Van Hee, Cynthia and Singh, Pranaydeep and Lefever, Els. 2021. Aspect-based sentiment analysis for German : Analyzing 'talk of literature' surrounding literary prizes on social media. In *Computational Linguistics in the Netherlands (CLIN 31), Abstracts*, Ghent, Belgium.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. 2020. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592.

Mattia Egloff, Davide Picca, and Alessandro Adamou. 2019. Extraction of character profiles from the gutenberg archive. In *Metadata and Semantic Research*, pages 367–372, Cham. Springer International Publishing.

Paul Ekman. 2015. Darwin and facial expression: A century of research in review. *null*.

Paul Ekman and Wallace V. Friesen. 1969. Nonverbal leakage and clues to deception. *Psychiatry MMC*.

Paul Ekman, Wallace V. Friesen, and Phoebe C. Ellsworth. 1972. Emotion in the human face: Guidelines for research and an integration of findings. *null*.

N.H. Frijda. 1993. Moods, emotion episodes and emotions. In M. Lewis and J.M. Haviland, editors, *Handbook of Emotions.*, pages 381–403. Guilford Press, New York.

Pascal Gygax, Jane Oakhill, and Alan Garnham. 2010. The representation of characters' emotional responses: Do readers infer specific emotions? *Cognition & Emotion*, 17:413–428. 43 citations (Crossref) [2023-04-27].

Anne Hamby, Daphna Motro, Zared Shawver, and Richard Gerrig. 2022. Examining readers' emotional responses to stories: An appraisal theory perspective. *Journal of Media Psychology: Theories, Methods, and Applications*, pages No Pagination Specified–No Pagination Specified. 1 citations (Crossref) [2023-04-27] Place: Germany Publisher: Hogrefe Publishing.

Kajiwara, Tomoyuki and Chu, Chenhui and Takemura, Noriko and Nakashima, Yuta and Nagahara, Hajime. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104.

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

Youngmoo E. Kim, Erik M. Schmidt, Raymond Migneco, Brandon G. Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A. Speck, and Douglas Turnbull. 2010. Music emotion recognition: A state of the art review. In *Proc. Ismir*, volume 86, pages 937–952.

David Konstan. 2015. Affect and emotion in greek literature. *null*.

Klaus Krippendorff. 2013. *Content analysis: An introduction to its methodology*. Sage Publications.

Jung-Hoon Lee, Hyun-Ju Kim, and Yun-Gyung Cheong. 2020. A Multi-modal Approach for Emotion Recognition of TV Drama Characters Using Image and Text. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 420–424, Busan, Korea (South). IEEE.

Bing Liu. 2012a. Sentence subjectivity and sentiment classification. In *Sentiment analysis and opinion mining*, pages 37–48. Springer International Publishing, Cham.

Bing Liu. 2012b. *Sentiment Analysis and Opinion Mining*, 1 edition. Synthesis Lectures on Human Language Technologies. Springer Cham. EBook ISBN: 978-3-031-02145-9, Published: 31 May 2022.

Jean Shinoda M. D. Bolen. 1993. *Gods in Everyman: Archetypes That Shape Mens Lives*, edition unstated edition. Harper Paperbacks.

Giovanna Menardi and Nicola Torelli. 2014. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

A. Ortony, G.L. Clore, and A. Collins. 1990. The cognitive structure of emotions.

John Pavlopoulos, Alexandros Xenos, and Davide Picca. 2022. Sentiment Analysis of Homeric Text: The 1st Book of Iliad. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7071–7077, Marseille, France. European Language Resources Association.

Davide Picca and Caroline Richard. 2023. Unveiling emotional landscapes in plautus and terentius comedies: A computational approach for qualitative analysis. In *Ancient Language Processing Workshop*, pages 88–95.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is Multilingual BERT?

R. Plutchik. 2001. The nature of emotions. *American scientist*, 89(4):344–350.

J. Posner, J.A. Russell, and B.S. Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734.

J.A. Russell. 1980. A circumplex model of affect. *J Personal Soc Psychol*, 39(6):1161–1178.

Anvita Saxena, Ashish Khanna, and Deepak Gupta. 2020. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79.

Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S. Huang. 2005. Multimodal approaches for emotion recognition: A survey. In *Internet Imaging VI*, volume 5670, pages 56–67. SPIE.

Rachele Sprugnoli, Francesco Mambrini, Marco Passarotti, and Giovanni Moretti. 2021. Sentiment analysis of latin poetry: First experiments on the odes of horace. In *Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022*, volume 3033 of *CEUR Workshop Proceedings*. CEUR-WS.org.

C. Vinola and K. Vimaladevi. 2015. A survey on human emotion recognition approaches, databases and applications. *ELCVIA: electronic letters on computer vision and image analysis*, pages 00024–44.

Vijaya Kumari Yeruva, Mayanka ChandraShekar, Yugyung Lee, Jeff Rydberg-Cox, Virginia Blanton, and Nathan A Oyler. 2020. Interpretation of sentiment analysis in aeschylus's Greek tragedy. In *Proceedings of the the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 138–146, Online. International Committee on Computational Linguistics.