Annotating the Sentiment of Homeric Text

John Pavlopoulos, Alexandros Xenos, Davide Picca
Athens University of Economics and Business, Greece
University of Lausanne, Switzerland
annis, a.xenos20@aueb.gr, davide.picca@unil.ch

Abstract

Sentiment analysis studies are focused more on online customer reviews or social media, and less on literary studies. The problem is greater for ancient languages, where the linguistic expression of sentiments may diverge from modern linguistic forms. This work bridges the gap by introducing a dataset with sentiment annotations, collected by Modern Greek readers for the first Book of Iliad. The ground truth time-series of the sentiment of the readers is explored and an empirical investigation is undertaken.

Keywords: Sentiment analysis, language resources, Homeric text

1. Introduction

Sentiment analysis is a series of methods, techniques, and tools about detecting and extracting subjective information, such as opinion and attitudes, from language (Liu, 2009). The field has experienced a rapid growth during the last years, with most of the thousands of papers being published after 2004 (Mäntylä et al., 2018). The main focus of sentiment analysis is the analysis of online customer reviews and, more recently, social media texts, but applications also range to financial market prediction, business review analysis, politics, demonetisation, crime prediction, and disaster assessment (Yadav and Vishwakarma, 2020).

The challenge of sentiment analysis is more demanding for ancient sources for which the linguistic expression of sentiments may be significantly divergent from modern linguistic forms. Nonetheless, attempting to overcome such an issue, researchers employed the corresponding modern Chinese translation (Zhao et al., 2014) in an approach to classify the sentiment of ancient Chinese literature. The same approach was followed in (Yeruva et al., 2020), where the authors used an English translation of randomly chosen sentences from Aeschylus's ancient Greek tragedies to classify the sentiment. The authors found that in-context annotation (i.e., the preceding verses) leads to different perceived sentiment compared to out-of-context annotation, which is probably due to the temporal order of sentiments in long texts.

Motivated by the lack of sentiment annotations for ancient languages, specifically for ancient Greek, in this work we annotate an ancient Greek poem, translated in Modern Greek, verse by verse, in order to maximise the context provided to the annotators. We model the poem's perceived sentiment as a multi-variate time series, where sentiment is defined as the fraction of annotators that found a verse as belonging to a specific class. Furthermore, we experimented with a state of the art deep learning masked language model, pre-trained on Modern Greek and fine-tuned on our data.

The contributions of this work are the following:

- We present the sentiment annotation of Homeric text, vis. the first Book of Iliad. Our dataset is publicly released.[1]
- To annotate the sentiment of text in ancient Greek, we employ a Modern Greek translation, which is expected to be closer to the original compared to other languages. Hence, we expect our dataset to serve as an accurate ground truth, that can be used to benchmark sentiment classifiers in any of the languages that the same text exists today (e.g., English and French).
- We annotate the sentiment by providing the annotators with one verse at a time, hence maximising the use of context (i.e., previous verses and narration) in the annotators' eyes.
- Experimenting with annotating the perceived sentiment of readers compared to annotating the sentiment that the author aimed to provoke, we showed that the former leads to higher inter-annotator agreement.
- We provide a list of emotions that were extracted by the annotators during their sentiment annotation task, employed as a means to increase the annotators' focus.
- Experimenting with GreekBERT (Koutsikakis et al., 2020) we registered a macro-averaged mean square and absolute error of 0.063 and 0.187 respectively. This promising result indicates that our dataset can be used to build effective sentiment estimators, which could be used to mechanically annotate the rest twenty three Books of Iliad and facilitate distant reading in Digital Humanities.

2. Related Work

The analysis of sentiments is certainly one of the most explored fields. It has been applied to several tasks ranging from product reviews (Markopoulos et al.,

[1] https://github.com/ipavlopoulos/sentiment_in_homeric_text

2015; Redhu, 2018), to twitter posts (Singh et al., 2014; Bhuta and Doshi, 2014). In recent years there has also been a growth of interest in all those aspects related to humanistic content. In fact, with the rise of Digital Humanities, the interest in the sentiment analysis for classical content, has become more prominent in the scientific community (Kim and Klinger, 2018; De Greve, Lore and Martens, Gunther and Van Hee, Cynthia and Singh, Pranaydeep and Lefever, Els, 2021) encouraging scholars to create tools and methods for the analysis of literary texts (Picca and Egloff, 2017; Schmidt and Wolff, 2021; Picca and Gay-Crosier, 2021).

More recently, the interest in dataset development in the literary field has grown strongly, spanning from the textual content to multimodal channels such as video, speech and music. Hence, datasets have been created and corpora have been collected for diverse tasks. For example, in (Kulkarni and Alicea, 2021) the authors created a database of more than 1,076 English titles in order to index and search texts taking into account interest and emotional makeup of readers. So authors came out with search and indexing that was based on sentiment progression for locating and recommending books. Special attention was also given to the use of other multimodal channels such as sentiment analysis of video and audio in the performance of theatrical plays. An example is provided in (Schmidt and Burghardt, 2018), where the authors acquired a video recording of a 2002 theatre performance in order to evaluate textual lexicon-based sentiment analysis and two state-of-the-art audio and video sentiment analysis tools. Sentiment annotation, however, can be present in a variety of cultural production spaces, such as the creation of song lyrics (Akiki and Burghardt, 2021). The authors of this work empirically collected valence, dominance, and arousal scores, based on user-generated tags that are available for 90,001 songs available on LAST.FM.

Despite this production of published studies in a variety of application fields, as well as on diverse channels (e.g., sound and video), a smaller degree of attention has been paid to classical literally work, especially regarding classical Greek. An exception is the work of Sprugnoli et al., who created a small gold standard consisting of eight Horatian poems (Sprugnoli et al., 2021). Each sentence was manually labeled as positive, negative, neutral, or mixed, and two automatic approaches were then assessed for sentiment classification. Another fortunate exception is (Yeruva et al., 2020), where the authors studied sentiment analysis of Aeschylus (Greek) text, by relying on a survey-based approach with approx. 60 college students and three popular English sentiment analysis tools as machine annotators. To the best of the authors knowledge, however, no empirical sentiment annotation has been provided for any of the twenty four Books of Iliad (15,693 verses), a war poem (Liddell et al., 2011), nor for the post-war poem of Odyssey. That is despite the fact that theoretical sentiment studies are not rare in literature (Scott, 1979; Koziak, 1999; Braund and Most, 2004). Our work attempts a first step in this direction, by establishing a sentiment annotation of the 1st Iliad Book.

3. The Dataset

3.1. Defining the annotation schema

In a preliminary experiment, 92 Iliad verses were isolated along with their translation to modern Greek.[2] The verses were taken from the 6th and the 24th Book, comprising the meeting of Hector with Andromache and the ransom of Hector, respectively. These two Books were selected due to the passionate nature of their dialogues, which are expected to comprise emotion-rich verses. Fourteen graduate students were given the 92 verses and they were divided randomly into two groups.[3]

Annotators of one group (GROUP A) were asked to annotate the sentiment (positive, negative, mixed or neutral) and the exact emotion (free text; advised to use a single word when they could) they felt when reading the verse. The annotators of the other group (GROUP B), instead of focusing on their own sentiment and emotions when reading each verse, they annotated those that the poet tried to provoke to the reader. The list of extracted emotions was continuously updated by the annotators during their task and it was accessible to all annotators at all times. The exhaustive list of emotions that was compiled is shown in Table 1.

For each annotator, we averaged and rounded the polarity annotations of the rest, in order to compute their agreement. A much higher percentage agreement was achieved for GROUP A (74.28%) compared to GROUP B (51.52%) for the task of subjectivity detection (neutral vs. positive, negative, or mixed). The same analogy held using Cohen's kappa (38.10% for GROUP A compared to 13.61% for GROUP B). We interpret this result in two ways. First, even at the higher level of detecting whether a sentiment exists or not in the verse and even for GROUP A, agreement is relatively low. Second, the emotion the author wanted to provoke to the reader is harder to label than the perceived emotion of the reader. This result is consistent with findings in literature, where readers have been found to underestimate the writer's emotions (Kajiwara, Tomoyuki and Chu, Chenhui and Takemura, Noriko and Nakashima, Yuta and Nagahara, Hajime, 2021).

Based on the findings of this preliminary experiment, we opted to continue our experiments by instructing the annotators to label their own sentiment that was perceived while reading. Also, we decided to provide them with the compiled list of emotions and to ask them to

[2] We used the translation by N. Kazantzakis and I. T. Kakridis, Athens 1955.

[3] All students were native Greek speakers, enrolled students of an MSc in Digital Humanities, with background in Linguistics or in Greek Literature. Two were male and twelve were female.

hope, fear, joy, distress pride, shame, admiration, reproach, love, hate, anger, remorse, relief, satisfaction, mercy, empathy, anxiety, worry, awe, willingness, pain, complaint, sorrow, surprise, guilt, shocking, question, grief, suspense, insecurity, loneliness, sadness, humiliation, compassion, fury, dispassion, affinity, disdain, Self denial, pain, irony/sarcasm, injustice, Being moved, rejection, longing, respect, jealousy, certainty, homesickness, Self pity, grudge, confidence, compassion, bravery, acknowledgement, despair, awareness

Table 1: Emotions suggested by the annotators during their task.

select the proper emotion. Emotion detection is a much harder task then sentiment classification, but we added this subtask to keep the annotator's focus on the task. For the same goal, and to simplify the task, we asked the annotators to write the name of the hero talking in the verse and to not annotate any sentiment/emotion when it was the narrator speaking.

3.2. Building the dataset

By using the same edition and translation, we gave all the verses of the first Book of Iliad to eight annotators.[4]

Inter-annotator Agreement

Percentage agreement when using the three sentiment (positive, negative, neutral) and the single narrator class was found to be 0.50. Krippendorff's alpha was found to be 0.39 and the free marginal kappa was 0.33. When we measured the agreement only for the narrator class (one vs. rest; is it the narrator talking in the verse or not), we observe an alpha of 0.83 and a Kappa of 0.95, which indicate that the annotation task was performed as planned, but there are verses for which the annotators will disagree with regards to the perceived sentiment.[5] Hence, any two readers may experience different emotions while reading the same verse.

Exploratory analysis

Based on our finding that the same verse may be perceived with different emotions by different readers, we modelled four temporal variables. One time-series has been created for each category, reflecting the fraction of annotators who labelled the verse with the respective category. Figure 1 presents these four time-series. The results show that there is a complementary nature between the positive and the negative class (Pearson's Rho correlation: -0.58), which appear to occur one after the other (i.e., a high negative fraction appears after a high positive fraction, and vise versa). The narrator time series appears to have a consistently low score, but we note that the rolling window hides the existing high peaks (all annotators picked the narrator class), because there are not many consecutive verses where it is only the narrator speaking.

[4]All students were native Greek speakers and graduate students of an MSc in Data Science. Seven male and one female students were enrolled.

[5]We also experimented with other binarisation approaches, which brought no improvement with regards to the agreement. The agreement remained low even when we grouped positive with negative, and neutral with narrator.

The 611 verses of this Book are 55 characters long, on average, with a standard deviation of 3. The shortest verse has 44 characters and the longest 69. We model the verse polarity ($P(v)$) as the product between the fraction of positive (f^+) and the fraction of negative (f^-) annotations per verse: $P(v) = f^+(v) * f^-(v)$. By sorting the verses based on polarity, we found that ten verses were annotated by half of the annotators as positive and by the other half as negative. These verses are presented in Table 2. Additionally, we found fifteen verses which all the annotators found as positive (verse number: 85, 89, 127, 208, 209, 210, 213, 214, 262, 274, 277, 278, 279, 283, 298, 443, 447, 456, 472, 474) and nineteen which all the annotators found negative (26, 28, 29, 32, 103, 104, 105, 106, 107, 108, 176, 187, 324, 325, 413, 414, 415, 416, 417, 418).
By looking at the verses based on the sentiment class assigned by the majority of the annotators (Fig. 2), we found that in 69 verses three annotators classified the respective verse as positive and three annotators classified it as negative while the other two annotators classified it as neutral or annotator. In the other 542 verses, 38 were classified as neutral by the majority of the annotators, 233 as positive, 232 as negative and 39 said that it was the narrator speaking.

4. Empirical Analysis

We used our dataset to fine-tune Greek-BERT (Koutsikakis et al., 2020) in sentiment estimation. Greek-BERT is a large Transformer pre-trained on Greek corpora. It achieves state of the art performance in three Greek Natural Language Processing (NLP) tasks, i.e., part-of-speech tagging, named entity recognition, and natural language inference. We fine-tuned Greek-BERT to estimate the fraction of the annotators that classified a verse to a sentiment category (i.e., positive, negative, neutral) or labelled it as a verse where the narrator was speaking. Regarding the network architecture, we added a feed-forward neural network (FFNN) on top of the top-level embedding of the [CLS] token. The FFNN consists of a dense layer (128 neurons) and a TANH activation function, followed by another dense layer. The last dense layer has four output neurons, with a softmax activation function on top to produce the probability distribution over the four categories (i.e., positive, negative, neutral, narrator). Following the work of (Fornaciari et al., 2021), we trained our model using probabilistic gold labels, to handle instance ambiguity, instead of using the standard one-hot labels that ignore the disagreement between the annota-

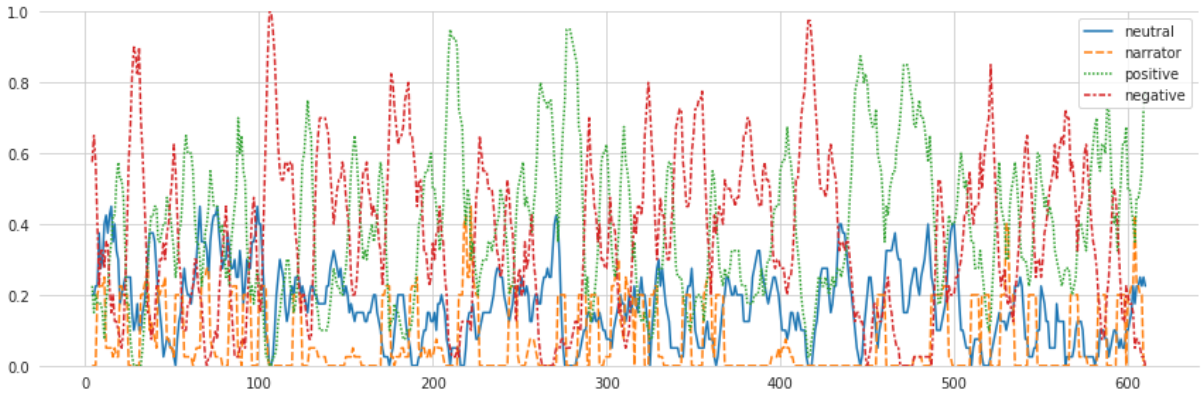

Figure 1: Fraction of annotators per class (positive, negative, neutral, narrator) per verse of the 1st Iliad Book. A rolling average with a window of 5 verses is shown.

Figure 2: Fraction of the majority of the annotators per sentiment per verse. With blue, green, red color are verses classified as neutral, positive and negative respectively by the majority of the annotators. Orange are verses in which the majority of the annotators said the the narrator was speaking. Black are verses where three annotators classified them as positive and three as negative while the other two annotators classified it as neutral or annotator.

tors over instances. We used mean square error (MSE) as our loss function and early stopping with patience of 5 epochs. For our experiment we used a 80/10/10 percent train/validation/test split respectively, but keeping the temporal order of the verses (the verses of training set preceded the ones for validation, which preceded the ones for testing purposes).

Experimental results

Table 3 presents the mean squared error and mean absolute error per sentiment dimension (i.e., the error of the predicted compared to the gold fraction), as well as macro-averaged. The highest error is for the positive while the lowest error is achieved for the neutral dimension. The mean error across dimensions was 0.063 (MSE) and 0.187 (MAE). When exploring the predicted and the gold sentiment of the final 61 verses that were used for testing (Figure 3), we observe that the model's predictions (dashed) fall close to the gold ones for the three sentiment dimensions. Errors are obvious in the narrator dimension, which means that there are verses that the model cannot distinguish with regards to the speaker (narrator or hero).

5. Conclusion

This study presented a dataset based on the 1st Iliad Book, including annotations of sentiment as this was perceived by Modern Greek native speakers when reading the poem. Verses with polarised sentiment and unanimous annotations were presented, while an empirical analysis showed that an existing deep learning sentiment estimation model can achieve a low error. The latter finding is particularly important for potential mechanical annotation of other Homeric texts. In

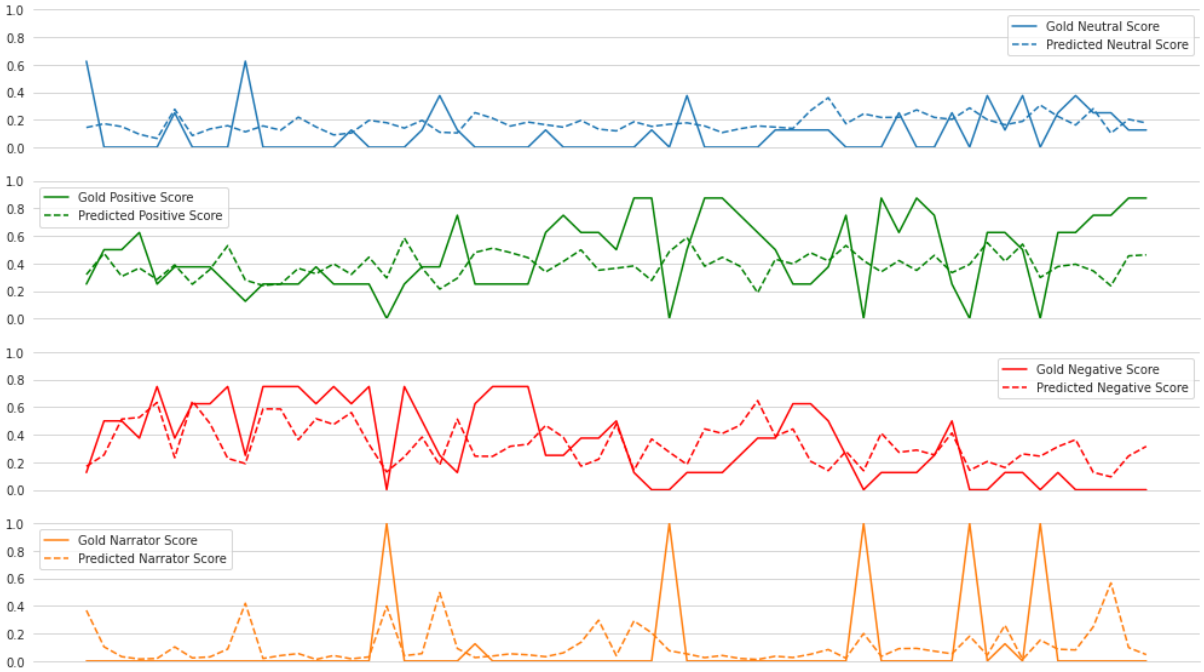

Figure 3: Predicted (dashed line) and ground truth (solid line) fraction of annotations per category of the last 61 verses, when fine-tuning GreekBERT on all the previous verses.

#	ILIAD BOOK 1: TRANSLATED VERSE
42	οι Δαναοί με τις σαγίτες σου τα δάκρυα που 'χω χύσει!
44	κι απ' την κορφή του Ολύμπου εχύθηκε θυμό γεμάτος, κι είχε
46	κι αντιβροντούσαν οι σαγίτες του στις πλάτες, μανιασμένος
194	και το τρανό σπαθί του ανάσερνε, να τη η Αθηνά απ' τα ουράνια'
197	Πίσω του εστάθη και τον άρπαξεν απ' τα ξανθά μαλλιά του
545	Ήρα, το κάθε που στοχάζομαι καθόλου μην τ' ολπίζεις
550	μη θες να το ρωτάς ανώφελα και μην ψιλοσκαλίζεις.
552	Ύγιέ του Κρόνου τρομερότατε, τι λόγια αυτά που κρένεις;
553	Ποτέ να σε ρωτώ δε θέλησα και να ψιλοσκαλίζω,
581	να μας πετάξει... τι στη δύναμη πολύ τρανότερος μας.

Table 2: Polarised verses (half annotators found the verse positive while the rest found it negative) with their verse number shown on the left. The edition's translation by N. Kazantzakis and I. T. Kakridis, 1955.

future work we plan to expand the dataset with more Books and investigate the accuracy and the potentials of mechanical sentiment annotation.

	MSE	**MAE**
POSITIVE	0.083	0.238
NEGATIVE	0.062	0.204
NEUTRAL	0.033	0.151
NARRATOR	0.075	0.154
MACRO AVG	0.063	0.187

Table 3: Mean absolute error (MAE) and mean squared error (MSE) per class, between the gold and the predicted scores for the last 61 verses, using the rest to fine-tune Greek BERT.

6. Bibliographical References

Akiki, C. and Burghardt, M. (2021). Muse: The musical sentiment dataset. *Journal of Open Humanities Data*, 7, 07.

Bhuta, S. and Doshi, U. (2014). A review of techniques for sentiment analysis of twitter data. *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*. 37 cites:.

Braund, S. and Most, G. W. (2004). *Ancient anger: perspectives from Homer to Galen*, volume 32. Cambridge University Press.

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*.

Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy,

D., and Poesio, M. (2021). Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online, June. Association for Computational Linguistics.

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.

Kim, E. and Klinger, R. (2018). A survey on sentiment and emotion analysis for computational literary studies. *CoRR*, abs/1808.03137.

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., and Androutsopoulos, I. (2020). Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

Koziak, B. (1999). Homeric thumos: The early history of gender, emotion, and politics. *The Journal of Politics*, 61(4):1068–1091.

Kulkarni, H. and Alicea, B. (2021). Sentiment progression based searching and indexing of literary textual artefacts.

Liddell, H. G., Scott, R., Jones, H. S., McKenzie, R., and Project, T. L. G. (2011). Translation of the word μῆνις:the online liddell-scott-jones greek-english lexicon.

Liu, B. (2009). Handbook chapter: sentiment analysis and subjectivity. handbook of natural language processing. *Handbook of Natural Language Processing. Marcel Dekker, Inc. New York, NY, USA*.

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

Markopoulos, G., Mikros, G., Iliadi, A., and Liontos, M. (2015). Sentiment analysis of hotel reviews in greek: A comparison of unigram features. 9:373–383, 01.

Picca, D. and Egloff, M. (2017). Dhtk: The digital humanities toolkit. In *WHiSe@ISWC*.

Picca, D. and Gay-Crosier, C. (2021). An Automatic Partitioning of Gutenberg.org Texts. In Dagmar Gromann, et al., editors, *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 35:1–35:9, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Redhu, S. (2018). Sentiment analysis using text mining: A review. *International Journal on Data Science and Technology*, 4(2):49–49. 12 cites:.

Schmidt, T. and Burghardt, M. (2018). *Toward a tool for sentiment analysis for german historic plays*. epub.uni-regensburg.de.

Schmidt, T. and Wolff, C. (2021). Exploring multimodal sentiment analysis in plays: A case study for a theater recording of emilia galotti. In *CHR*.

Scott, M. (1979). Pity and pathos in homer. In *Acta Classica: Proceedings of the Classical Association of South Africa*, volume 22, pages 1–14. Classical Association of South Africa (CASA).

Singh, S., Paul, S., Kumar, D., and Arfi, H. (2014). Sentiment analysis of twitter data set: survey. *International Journal of Applied . . . .*

Sprugnoli, R., Mambrini, F., Passarotti, M., and Moretti, G. (2021). Sentiment analysis of latin poetry: First experiments on the odes of horace. In Elisabetta Fersini, et al., editors, *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.

Yadav, A. and Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385.

Yeruva, V. K., ChandraShekar, M., Lee, Y., Rydberg-Cox, J., Blanton, V., and Oyler, N. A. (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.

Zhao, H., Wu, B., Wang, H., and Shi, C. (2014). Sentiment analysis based on transfer learning for chinese ancient literature. In *2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2014)*, pages 1–7. IEEE.