Interpretation of Sentiment Analysis in Aeschylus's Greek Tragedy

Vijaya Kumari Yeruva
Dept. of CSEE
Univ. of Missouri-KC
vyq4b@mail.umkc.edu

Mayanka Chandrashekar
Dept. of CSEE
Univ. of Missouri-KC
mckw9@mail.umkc.edu

Yugyung Lee
Dept. of CSEE
Univ. of Missouri-KC
leeyu@umkc.edu

Jeff Rydberg-Cox
Dept. of English
Univ. of Missouri-KC
rydbergcoxj@umkc.edu

Virginia Blanton
Dept. of English
Univ. of Missouri-KC
blantonv@umkc.edu

Nathan A Oyler
Dept. of Chemistry
Univ. of Missouri-KC
oylern@umkc.edu

Abstract

Recent advancements in NLP and machine learning have created unique challenges and opportunities for digital humanities research. In particular, there are ample opportunities for NLP and machine learning researchers to analyze data from literary texts and to broaden our understanding of human sentiment in classical Greek tragedy. In this paper, we will explore the challenges and benefits from the human and machine collaboration for sentiment analysis in Greek tragedy and address some open questions related to the collaborative annotation for the sentiments in literary texts. We focus primarily on (i) an analysis of the challenges in sentiment analysis tasks for humans and machines, and (ii) whether consistent annotation results are generated from the multiple human annotators and multiple machine annotators. For human annotators, we have used a survey-based approach with about 60 college students. We have selected three popular sentiment analysis tools for machine annotators, including VADER, CoreNLP's sentiment annotator, and TextBlob. We have conducted a qualitative and quantitative evaluation and confirmed our observations on sentiments in Greek tragedy.

1 Introduction

Recent advancements in NLP and machine learning have created unique opportunities for digital humanities research. In particular, sentiment analysis toolkits provide a way to explore the representation of emotions in literary texts such as ancient Greek Tragedy. Aristotle defined *tragedy* as a medium for bringing out emotions, especially pity and fear. Greek tragedies express a plethora of emotions via the characters and their narratives. Recent advancements in the NLP and machine learning make it possible to conduct a systematic analysis of these sentiments and emotions using computational tools.

Recent work has explored the differences between sentiments and emotions in Greek Tragedy and contemporary society. There have been two contrasting views on a universal emotion across time and space. On the one hand, Kalimtzis used David Cairn's school of thought that "cultures exhibit points of overlap that make them mutual intelligible," in other words, having "naive assumption of shared humanity" (Kalimtzis, 2014). In contrast, Konstan endorses the opposing view that emotions observed in Greeks of the classical period are different from the modern ones (Konstan, 2015; Muellner and others, 1996). In recent years, there has been scholarly work focused on comparing the contemporary population's emotional impact based on Greek Tragedy and horrific contemporary events (Munteanu, 2017), or even focusing on the role of emotions in ancient Greek diplomatic practice (Gazzano, 2019).

Sentiment analysis could be used as a training step for machines to perform more complex tasks like emotion detection. However, sentiment analysis is not an easy task for a machine because of the multiple and often unpredictable variables applied to interpret a given sentence. Typically, sentiment in context is

Proceedings of LaTeCH-CLfL 2020, pages 138–146
Barcelona, Spain (Online), December 12, 2020.

an incredibly complex task for machines. A study (Min and Park, 2019) presented a highly complex and dynamic system for reflecting the rich structure of human interaction and communication and identifying associated sentiments and topics by characterizing relationships explicitly. It demonstrated how these methods could be used to explore Victor Hugo's Les Misérables.

In our study, sentiment analysis on Greek Tragedy was conducted using both social media trained sentiment analysis tools and human annotators. This work is an initial step in exploring promising research on the modern understanding of ancient emotion. Furthermore, to conduct advanced sentiment analysis in machine learning, we need well-annotated data that can be used to teach machines about emotion in Greek Tragedy. The human-in-the-loop (Wu et al., 2019; Tsakalidis et al., 2018) has received attention for the potential of human and machine collaboration. More reliable data can be collectively annotated for in-depth study in machine learning from this kind of collaboration.

In this paper, we explore the challenges and benefits from the human and machine collaboration for sentiment annotation about sentiments in Greek Tragedy and address some open questions related to the collaborative annotation. We are particularly interested in analyzing why similar or different behaviors or opinions may be observed from the machine and human annotators–and what a comparison of human annotators and machine annotators can teach us about how humans and machines read emotion. We mainly focus on (i) an analysis of the challenges in sentiment analysis tasks for humans as well as machines and (ii) whether consistent annotation results are generated from multiple human annotators and multiple machine annotators.

2 Study Domain: Greek Tragedy

For this study, Aeschylus's essays were obtained in TEI conformant XML texts from the Perseus Digital Library (Smith et al., 2000; Rydberg-Cox, 2011). We extracted sentences based on stratified sampling, as shown in Table 1. We used the sentiment annotations ranging from 'extremely positive' to 'extremely negative' (also known as diversity sampling (Munro, 2019)).

Table 1: Greek Tragedy Survey Dataset

Essay Name	Sentences	Total# of Sentence
Eumenides	8	115
Prometheus Bound	11	138
Seven Against	7	82
Agamemnon	12	188
Suppliant Women	4	107
Persians	8	122
Total	50	752

2.1 Research Questions

There are two main goals of this paper:
RQ1: *What is the level of agreement between multiple human and machine annotators when evaluating sentiments? If the agreement is low, what are the reasons behind it? From the human annotators' evaluation, what is the impact of context towards their sentiment rankings?* To appropriately characterize or measure the mutual (dis)agreement between human annotators, we performed the statistical analysis and studied: (i) the correlation between sentiment annotation and the change in sentiment annotation when read in context, (ii) the correlation between sentiment annotation and survey sentence length, and (iii) the correlation between sentiment annotation and the number of words expressed emotions or sentiments.

RQ2: *What are the primary properties of annotators (humans or machines) that can "coexist" with sentiment annotations and (dis)agreement of sentiments in different segments and episodes?* We will assess the annotation by comparing human and machine agreement and disagreement and conduct both qualitative and quantitative evaluation regarding agreement and disagreement between human annotators. For the qualitative assessment, we will identify the most controversial question that shows the highest standard deviation among the human sentiment annotations. The survey questions will be explored further to explain the challenges of sentence-level annotation, contextual-level annotation, machine-level annotation, sentimental terms, and the survey sentence length. We will calculate Cohen's Kappa scores to describe the quality of human annotators' annotations for the quantitative evaluation.

Table 2: Survey Questions

ID	Questions	Answers
Q1	Sentiment for a given sentence (dialogue or partial of dialogue)	*Extremely Positive, Moderately Positive, Neutral, Moderately Negative, Extremely Negative*
Q2	Words that lead to sentiment discussion	*Word list from a given sentence*
Q3	Sentiment for a given sentence with the context defined	*Extremely Positive, Moderately Positive, Neutral, Moderately Negative, Extremely Negative*

2.2 Survey Design for Human Annotation

For the human annotation study, we have used a survey-based approach to access the human ability to analyze sentiment in Greek tragedy. For our study, 61 college students in the humanities were asked to rate the sentiments expressed in sentences extracted from the texts.

2.2.1 Design of Survey Questions

We asked students to focus on: (1) "sentiment" in a sentence (referred to as the target sentence), (2) the words in the target sentence that contributed to the annotated "sentiment." and (3) the "sentiment" of the target sentence within a given context (referred to as the sentence-in-context).

- Question 1 (Q1) aims to capture the sentiment expressed by a sentence, with no knowledge about the speaker or the play. This question is designed to capture the sentiment perceived by the annotator by reading a sentence in isolation.
- Question 2 (Q2) aims to understand why that annotator thinks a sentence exhibits a particular sentiment. The second question is designed to realize the interpretability component of the sentiment selected in Q1.
- Question 3 (Q3) aims to capture the sentence's sentiment within a broader context and if the human understanding of sentiment in a given sentence changes when that sentence is read in context. We represent the context by providing sentences before and after the given sentence from Question 1.

Table 2 shows the three questions that were posed for each sentence that the students were asked to annotate. The sentiment was categorized into the following categories: [*Extremely Positive, Moderately Positive, Neutral, Moderately Negative, Extremely Negative*].

2.2.2 Hypothesis and Observations

We had several assumptions about human annotators completing this task. First, the human annotators would not be broadly familiar with the Greek tragedies from which the sentences were drawn and would not be able to infer the broader context from a single sentence. Determining sentiment accurately in a single sentence without context is not an easy task. The sentiment annotation might change when it is understood within a broader context. Second, the sentiment expressed in many sentences is ambiguous and there may be subtle differences between the ways that different annotators perceive sentiment in a given sentence (e.g., extremely negative and negative). This ambiguity is also correlated with the number of words that express sentiment in any given sentence. Thus, it is even tricky for multiple human annotators to offer a consistent assessment of each sentence.

In our study, 50 sentences were randomly selected from the corpus of Greek tragedy for human annotation. On average, each sentence was annotated by fifteen human annotators (13-17 annotators per question). Of the human annotators, three types of sentiment questions were posed. (Q1) "sentiment" in a sentence (referred to as the target sentence), (Q2) the words in the target sentence that contributed to annotated "sentiment," and (Q3) the "sentiment" of the target sentence within a given context (referred to as the sentence-in-context).

2.3 Model Design for Machine Annotation

For machine annotation, we selected three popular sentiment analysis tools, VADER (Hutto and Gilbert, 2014), CoreNLP's sentiment annotator (Socher et al., 2013), and TextBlob (Loria, 2017). VADER (Hutto and Gilbert, 2014) is a rule-based model for sentiment analysis that was empirically constructed by a gold standard list of linguistic features and sentiment in microblog-like contexts. Developers demonstrated its effectiveness compared to the state-of-practice benchmarks and shallow machine learning algorithms.

VADER's sentiment intensity valence is ranged from -1 (most extreme negative) to +1 (most extreme positive). Stanford CoreNLP's sentiment annotator (Socher et al., 2013), which was designed with Recursive Neural Tensor Networks and the Stanford Sentiment Treebank, achieved 80.7% accuracy on fine-grained sentiment prediction. It has five sentiment classes, very negative to very positive (from 0 to 4) at a sentence level. TextBlob (Loria, 2017) determines sentiment in two measures, namely polarity and subjectivity. The polarity score describes the sentiment intensity in a range from -1.0 to 1.0. The subjectivity score ranges from 0.0 to 1.0, where 0.0 is very objective, and 1.0 is very subjective.

3 Experimental Results and Evaluation

These models for sentiment analysis are based on the lexical, grammatical, and syntactical conventions of model sentiment. Expressing sentiment intensity is determined by rules that might differ from sentiments as they are expressed in the literary texts. Our initial hypothesis is that human annotations of sentiments in Greek tragedy will be similar to the machine annotations and validate the applicability of these toolkits for literary texts. This is determined as follows: First, machine annotators face similar challenges to those of human annotators. Since machine annotators rely on a dictionary of word meanings, they can be consistently applied for sentiment analysis. However, some terms from Greek tragedy may not be adequately determined by machine annotators. Contemporary meanings in social media, for which the sentiment analysis tools were developed, may not align. Second, machine annotators have their own discrete sentiment annotations. For example, the annotation ratings for VADER, CoreNLP, and TextBlob vary. For simplicity, we have normalized them into five annotations, consistent with the human annotations ranging from extremely positive (1) to extremely negative (5). Finally, since machine annotators cannot determine a sentiment annotation when the input is given with the surrounding context, it is difficult for multiple machine annotators to get a consistent annotation result.

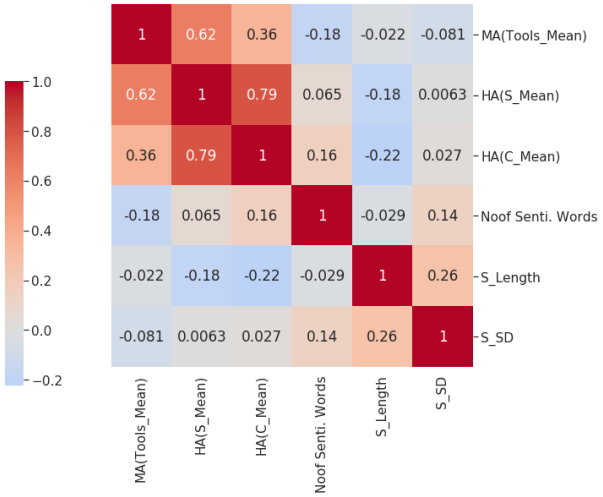

Figure 1: Correlation Testing

We used correlation tests to determine whether the values of quantitative variables change in conjunction with each approach. First, we computed the pairwise correlation coefficients for quantitative variables using the Pearson coefficient (r) and represented them in a heatmap. Second, we checked whether the variation between the sets of variables is monotonic (increasing or decreasing) or whether the data's underlying distribution is normal. Correlation coefficients ranged from -1.00 to +1.00. A positive value indicates a positive correlation – one variable is increasing, and so does the other – while a negative value indicates a negative correlation – one variable is increasing and the other decreasing.

Table 3 shows six hypothetical variables and Figure 1 illustrates the correlations of these variables: positive correlations are displayed in red and negative correlations in light blue. Color intensity is proportional to the correlation coefficients. We have also extended the Pearson coefficient correlation (r) with additional coefficients such as Spearman (ρ) and Kendall's tau (τ), and the overall graphs shown in Figure 2 are shown with consistent correlations for five variables and three coefficient values.

Table 3: Hypothetical Variables

Variable	Description
MA	*Machine Annotation (Mean)*
HS	*Human Annotation in Sentence (Mean)*
HC	*Human Annotation in Context (Mean)*
SW	*#Sentiment Words*
SL	*Sentence Length*
SD	*HA Standard Deviation*

As seen from Figure 1 and Figure 2, there are four positive correlations (HS-HC, HS-MA, HC-MA, SL-SD) and two negative correlations (HS-SL, HC-SL). The human-to-human annotation (HS-HC),

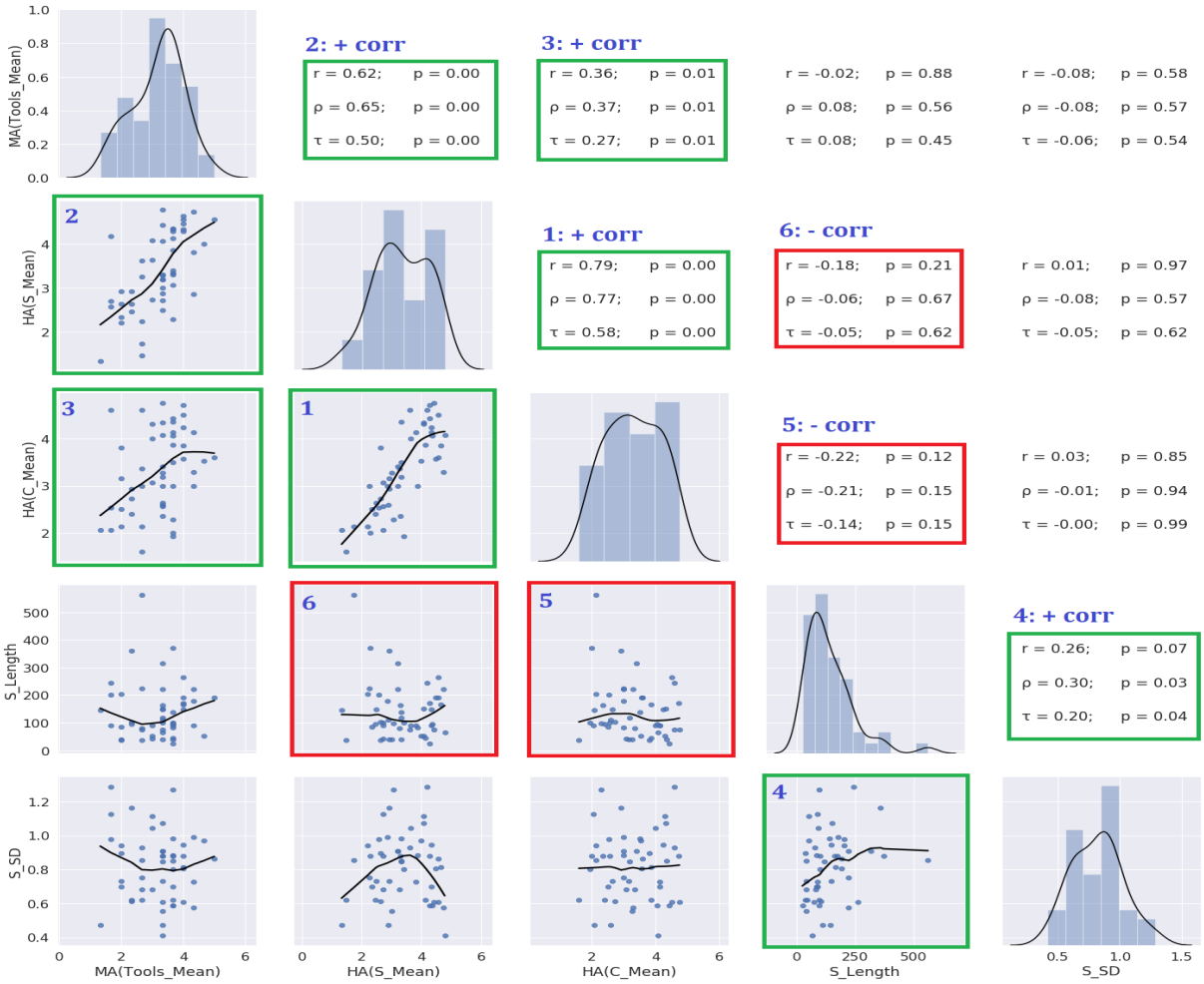

Figure 2: Correlation Testing between Machine Annotations (MA), Human Annotations (HA), Survey Sentence Length (S_Length), and Standard Deviation of Human Annotations (S_SD). Where r: Pearson, ρ: Spearman, and τ: Kendall's tau correlation's value. The green boxes (Cases 1, 2, 3, 4) indicate a positive correlation while the red boxes (Cases 5,6) indicate a negative correlation.

which is even in two different settings such as sentence and context, and the human-to-machine annotation (HS-MA) show the two highest correlations with coefficient values of 0.79 and 0.62, respectively. The human annotations for a sentence and the sentence length (HS-SL) and human annotations for a sentence in context and the sentence length (HC-SL) show the two lowest correlations with coefficient values of -0.22 and -0.18. Figure 1 shows no significant correlation between the human annotations for a sentence and the number of sentiment words (HS-SW).

3.1 Human and Machine Collaboration in Sentiment Annotation

Human-machine co-annotation is the first step towards interactive machine learning. Most of the current work in interactive machine learning ultimately uses human annotations as interpretability for the existing system (Wu et al., 2019; Smith-Renner et al., 2020; Lertvittayakumjorn and Toni, 2019). In this collaborative annotation system, humans and machines are given the same task to facilitate a compare and contrast analysis. We assess the questionnaire's ability to detect agreement or disagreement of human and machine annotators and then determine if there were significant correlations between variables for human and machine collaboration for sentiment annotation in Greek tragedy.

3.1.1 Agreement Among Human Annotators

We have evaluated the degree of the agreement among multiple annotators (13 to 17 annotators) using two approaches. First, we analyzed the standard deviation for the human annotators (13 to 17 annotators per question) for 50 questions. We have categorized the 50 questions into six categories: (1) SD $\leq$ 0.55,

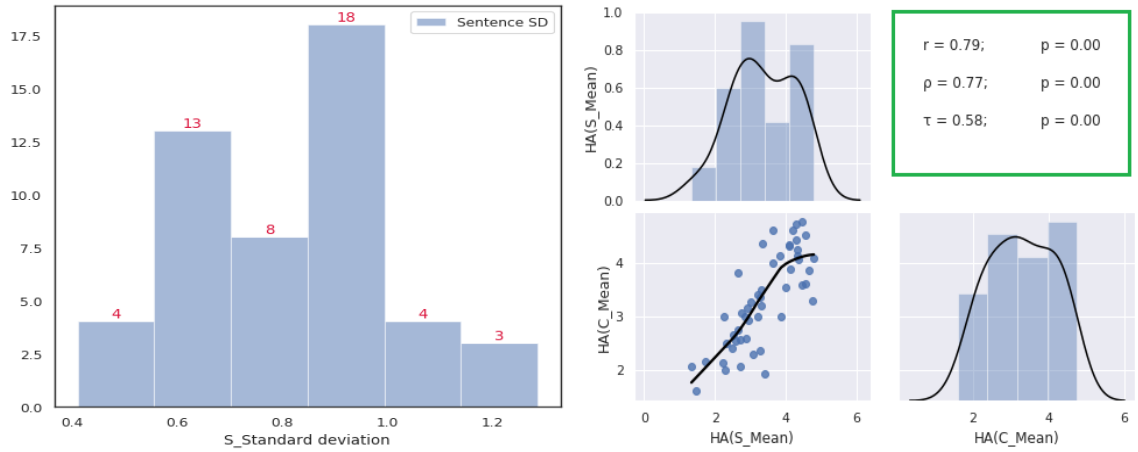

Figure 3: (a) Standard Deviation Distribution of HS (b) Correlation between HS and HC

(2) $0.56 \leq SD < 0.70$, (3) $0.70 \leq SD < 0.84$, (4) $0.85 \leq SD < 1.00$, (5) $1.00 \leq SD \leq 1.15$, and (6) $1.16 \leq SD < 1.30$ as shown in Figure 3(a). The mean of human annotations for both sentence and context are positively correlated with a Pearson correlation value of 0.79, as shown in Figure 3(b). One of the three most controversial survey questions ($1.16 \leq SD < 1.30$) is shown in Table 4. This survey question from the play *Agamemnon* was rated 3.07 (Neutral) in the sentence in isolation to 2.28 (Moderately Positive) in context by 13 annotators.

Second, we considered the correlation analysis using three coefficient measures, such as the Pearson coefficient correlation (r), Spearman (ρ), and Kendall's tau (τ). Figure 4 shows that the human annotations for both sentence (HA) and context (HC) are positively correlated to the machine annotations (MA) with the Pearson correlation values of 0.62 and 0.36, respectively. Similarly, Spearman (ρ) and Kendall's tau (τ) values show the positive correlations for HA and MA, HC, and MA.

Table 4: Sentiment Annotation with High Standard Deviation (SD=1.27)

Essay Name		*Agamemnon*
Sentence	**Survey Question**	**Survey Question in Context**
High SD	(Mean: 3.07) *Exile though I was, I laid my hand upon my enemy, compassing every device of cunning to his ruin.*	(Mean: 2.28) *But grown to manhood, justice has brought me back again. Exile though I was, I laid my hand upon my enemy, compassing every device of cunning to his ruin. So even death would be sweet to me now that I behold him in justice's net.*

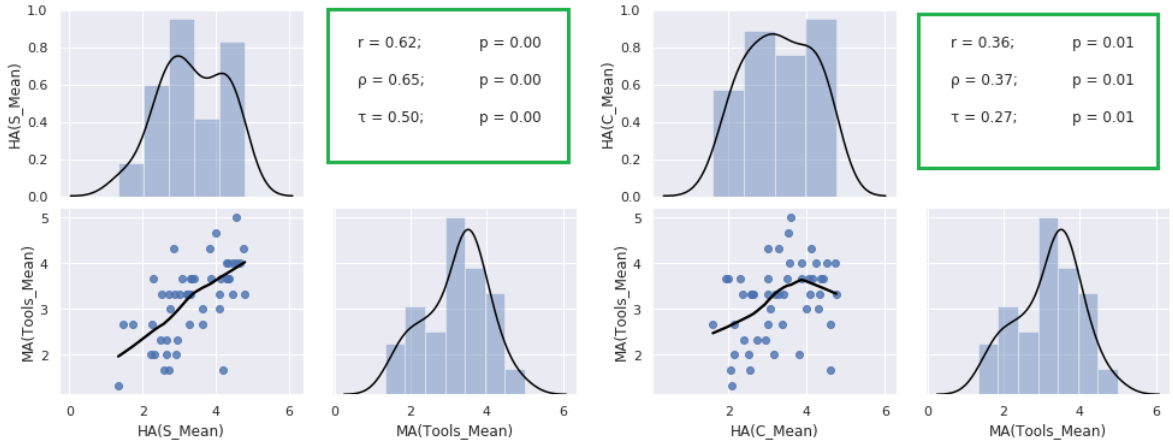

Figure 4: (a) Correlation between HS and MA (b) Correlation between HC and MA

3.1.2 Agreement between Human Annotators and Machine Annotators

For human and machine annotation, we computed Cohen's Kappa Correlation (Cohen, 1960), one of the most commonly used statistics to test inter-rater reliability (Tsakalidis et al., 2018). Kappa value of $<$ 0 indicates Poor agreement, 0.01 – 0.20: Slight agreement, 0.21 – 0.40: Fair agreement, 0.41 – 0.60: Moderate agreement, 0.61 – 0.80: Substantial agreement, and 0.81 – 1.00: Almost perfect agreement.

The inter-rater reliability for HS and MA was evaluated, and the kappa value of 0.11 was computed for the HS-MA. Then, we conducted it for three individual machine annotators (VADER, CoreNLP, and TextBlob). The kappa values of 0.23, 0.13, -0.05 were reported for HS-VADER, HS-CoreNLP, and HS-TextBlob. Among machine annotators, the agreement between VADER and human annotators (HS) shows the best kappa value, 0.23, compared to others, which is a *fair agreement*, according to Kappa Correlation.

3.2 Sentiment Change With Context

We have evaluated the impact of context on the survey question regarding the annotators' sentiment annotations. The introduction of context is an attempt to determine, using the theory of Michel Foucault, if sentiment in the context of Greek tragedy is a 'discursive object', a product of 'discourse' (Kalimtzis, 2014; Foucault, 1970). Figure 5 depicts the distribution of sentiment annotations of sentences from Greek tragedy without context and with context. The figure shows the sentiment annotation distributions of the survey questions without context in blue and with the context in yellow. Table 5 shows the most change in sentiment annotations. First, the mean of annotation by 14 annotators changes negatively from 3 (Neutral) to 3.8 (leaning towards Moderately Negative) by adding the context to the survey question. Second, the mean of annotation by 15 annotators changes negatively from 3.4 (Neutral) to 1.9 (leaning towards Moderately Positive) by adding the context to the survey question. This shows the impacts of the sentiment change due to the existence of context.

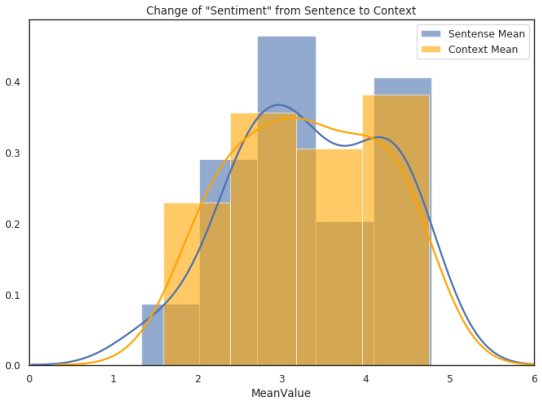

Figure 5: Sentiment Change With Context

Table 5: Sentiment Change in Context

	Essay Name	*Eumenides*
Change	**Survey Question**	**Survey Question in Context**
Negative 3 ⇒ 3.8	(Mean: 3.0) *I will give you strong proof of this.*	(Mean: 3.8) *I am not a suppliant in need of purification, nor did I sit at your image with pollution on my hands. I will give you strong proof of this. It is the law for one who is defiled by shedding blood to be barred from speech until he is sprinkled with the blood of a new-born victim by a man who can purify from murder.*
Positive 3.4 ⇒ 1.9	(Mean: 3.4) *Lord Apollo, you know how to do no wrong; and, since you know this, learn not to be neglectful also.*	(Mean: 1.9) *Lord Apollo, you know how to do no wrong; and, since you know this, learn not to be neglectful also. For your power to do good is assured.*

3.3 Correlation between Sentence Length and Sentiment Annotation

Human sentiment annotation tends to be negative for short sentences while positive for long sentences. Our results show that the higher the standard of deviation, the more the disagreement among annotators. Figure 6 shows that the sentence length (SL) is negatively correlated to human sentiment annotations (HA) and VADER's, while SL is positively related to standard deviations of human annotation (SD). Sentence length and standard deviation of human annotations for the sentence are positively correlated with a Pearson correlation value of 0.26, as shown in Figure 6(d). This indicates that longer sentences in the questionnaires tend to show more disagreement among annotators.

Regarding the relationship between the sentence length and sentiment annotation, Table 6 shows a short sentence is rated as *Extremely Negative* (mean: 4.28) by 14 human annotators and *Extremely Negative* by VADER. At the same time, a long sentence is rated as *Moderately Positive* (mean: 1.73) by 15 human annotators and *Extremely Positive* by VADER. Similar patterns are shown in the sentiment annotations for the question in context.

Table 6: Sentence Length and Sentiment Annotation

	Essay Name	*Eumenides*
Sentence	**Survey Question**	**Survey Question in Context**
Short	(Mean: 4.28) *Oh, oh, the shame of it!*	(Mean: 4.42) *I am breathing fury and utter rage. Oh, oh, the shame of it! What anguish steals into my breast!*
Long	(Mean: 1.73) *For us, the remnant of the Argive host, the gain has the advantage, and the loss does not bear down the scale; so that, as we speed over land and sea, it is fitting that we on this bright day make this boast: The Argive army, having taken Troy, at last, has nailed up these spoils to be a glory for the gods throughout Hellas in their shrines from days of old. Whoever hears the story of these deeds must extol the city and the leaders of her host, and the grace of Zeus that brought them to accomplishment shall receive its due measure of gratitude.*	(Mean: 2.14) *Our misfortunes should, in my opinion, bid us a long farewell. For us, the remnant of the Argive host, the gain has the advantage, and the loss does not bear down the scale; so that, as we speed over land and sea, it is fitting that we on this bright day make this boast: The Argive army, having taken Troy, at last, has nailed up these spoils to be a glory for the gods throughout Hellas in their shrines from days of old. Whoever hears the story of these deeds must extol the city and the leaders of her host, and the grace of Zeus that brought them to accomplishment shall receive its due measure of gratitude.*

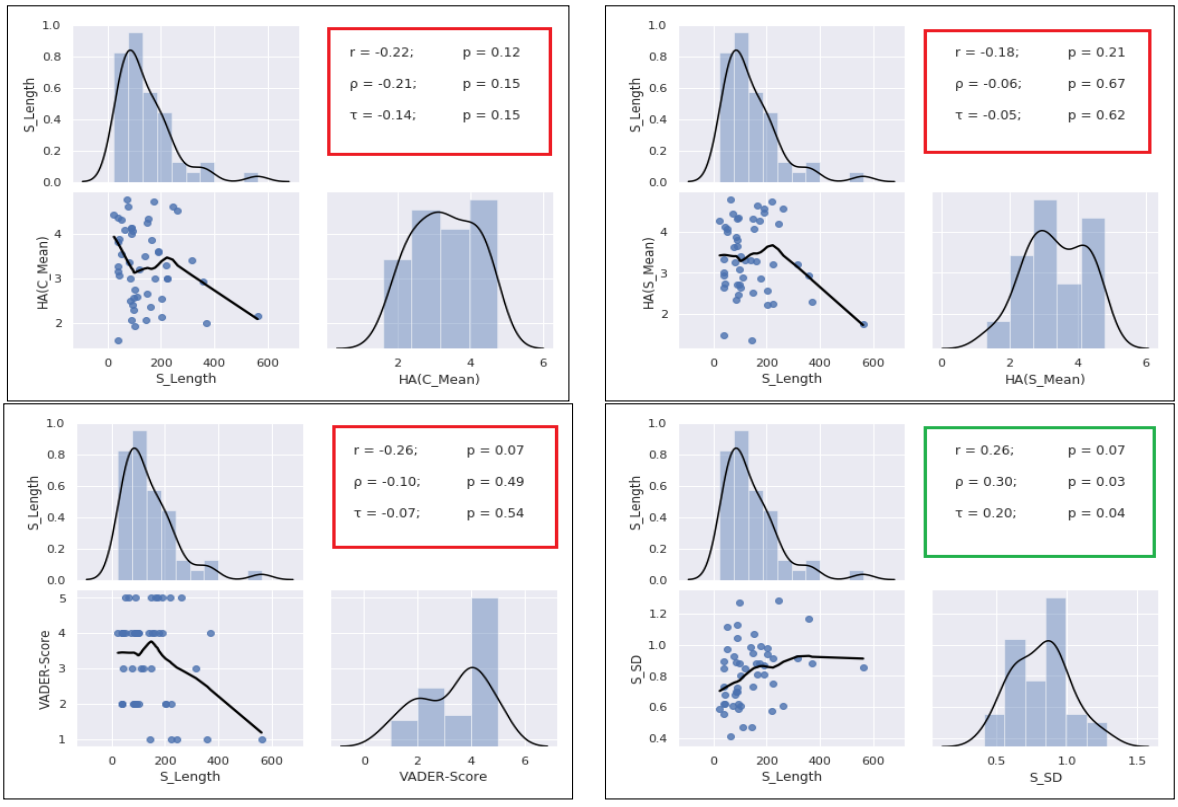

Figure 6: Correlations: (a) SL and HC (b) SL and HS (c) SL and VADER (d) SL and SD

4 Conclusions

This study explored the following questions: *What is the level of agreement between multiple human and machine annotators when evaluating sentiments in Greek tragedy? If the agreement is low, what are the reasons behind it?* First, we have conducted a coefficient correlation analysis with six variables

using Pearson, Spearman, and Kendall and found that there are positive correlations for human-to-human annotation as well as human-to-machine annotation and negative correlations for human annotation and sentence length. Second, we have conducted the inter-rater reliability between human and machine annotators, and the results are either fair or slight agreement. The inter-rater reliability between human and machine annotators confirms the high performance of computational sentiment analysis (especially VADER) and their applicability to literary texts such as Greek tragedies.

References

J Cohen. 1960. A coefficient of agreement for nominal scales educ psychol meas. *SAGE Publications Inc*, 20:37–46.

Michel Foucault. 1970. The archaeology of knowledge. *Information (International Social Science Council)*, 9(1):175–185.

Francesca Gazzano. 2019. Greek ambassadors and the rhetoric of supplication. some notes. *KTÈMA Civilisations de l'Orient, de la Grèce et de Rome antiques*, 44:53–69.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Kostas Kalimtzis. 2014. *Taming Anger: The Hellenic Approach to the Limitations of Reason*. A&C Black.

David Konstan. 2015. Affect and emotion in Greek literature. *Oxford Handbooks Online*.

Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation methods for text classification. 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 5198–5208.

Steven Loria. 2017. Textblob: Simplified text processing [a python (2 and 3) library for processing textual data].

Semi Min and Juyong Park. 2019. Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling. *PloS one*, 14(12):e0226025.

Leonard Charles Muellner et al. 1996. *The anger of Achilles: Mēnis in Greek epic*. Cornell University Press.

Robert Munro. 2019. Human-in-the-loop machine learning.

Dana LaCourse Munteanu. 2017. The paradox of literary emotion: An ancient Greek perspective and some modern implications. *Nuntius Antiquus*, 13(2):263–283.

Jeff Rydberg-Cox. 2011. Social networks and the language of Greek tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(3).

David A Smith, Jeffrey A Rydberg-Cox, and Gregory R Crane. 2000. The perseus project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25.

Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Adam Tsakalidis, Symeon Papadopoulos, Rania Voskaki, Kyriaki Ioannidou, Christina Boididou, Alexandra I Cristea, Maria Liakata, and Yiannis Kompatsiaris. 2018. Building and evaluating resources for sentiment analysis in the Greek language. *Language resources and evaluation*, 52(4):1021–1044.

Tongshuang Wu, Daniel S Weld, and Jeffrey Heer. 2019. Local decision pitfalls in interactive machine learning: An investigation into feature selection in sentiment analysis. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(4):1–27.