Breaking Bard: Unsupervised Adventures in the Homeric Epics

Venika Vachani
Dept. of Computer Science
Fordham University
New York, United States
vvachani@fordham.edu

***Abstract*—Scholars have been preoccupied with the Homeric Question– that is, the question of the identity of the two Ancient Greek epic poems, the Iliad and the Odyssey, traditionally attributed to Homer– since classical antiquity. This question is complicated by the fact that it is possible for multiple authors to have contributed to a single poem, reflecting not just the oral tradition of poetry that the Homeric epics were composed within, but also the possibility of later interpolations. This project attempts to address this question using unsupervised authorship analysis, by analyzing individual sections of the Homeric poems in order to identify potential author(s). By ignoring the traditional book divisions of the Homeric texts and dividing them into smaller segments of text, this work aims to identify stylistic outliers on a more granular level than that of previous computational analyses of the Homeric corpus.**

I. Introduction

Scholars have been preoccupied with the Homeric Question– that is, the question of the identity of the two Ancient Greek epic poems, the Iliad and the Odyssey– since classical antiquity. As the name suggests, both poems have traditionally been attributed to a single author, Homer. However, stylistic and thematic differences between the texts have inspired theories that "Homer" was really two or more poets. The Homeric Question is further complicated by the oral tradition of poetry that the Homeric epics were likely composed within: this may suggest the influence of multiple poets or bards who transmitted their own versions of the Iliad and the Odyssey for an unknown duration of time before "Homer" wrote them down. Finally, there is the very real possibility of interpolations and textual corruption [1].

Certain books and sections in both the Iliad and Odyssey have long faced skepticism of their authorship or authenticity based on traditional philological analyses. For example, Book 10 of the Iliad is considered by many scholars to be a later addition. But interpolations and manuscript corruption do not always occur on so large a scale. In reality, the authorship of the Iliad and Odyssey may vary even by line.

There are no other works, or sections within the Iliad or Odyssey, that can definitively be attributed to Homer. Indeed, we do not know how many authors might have contributed to the poems. Therefore, this is an unsupervised learning task. Specifically, this paper will attempt to apply clustering algorithms in order to identify segments of the poems that may share a common author, using linguistic features common in authorship analysis/text classification.

Previous attempts to apply computational techniques to the Homeric texts have used the traditional divisions of the poems into 24 books each. However, these book divisions are generally considered to be later (non-Homeric) additions: scholars have suggested dates as early as the 6th century B.C., to as late as the 2nd century B.C. This paper will ignore the traditional book divisions and divide each poem into smaller chunks for the purposes of clustering, with the aim of investigating whether the book divisions obscure potential differences in authorship (or the lack thereof).

II. Experimental Methodology

A. Dataset

The primary dataset consists of the two Homeric poems, the Odyssey and the Iliad, in Ancient Greek, provided in full through the Tesserae project and accessed through the CLTK [2]. The data was preprocessed to remove all non-Greek characters, punctuation, line numbers, etc. In addition, personal names and place names were removed, in order that the feature sets might capture stylistic traits peculiar to authors, rather than topical differences. The dataset was slightly biased towards the Iliad, which has a total of 15,693 lines to the Odyssey's 12,109. Each poem was then segmented into chunks of 100 lines each, for a total of 277 chunks.

B. Feature set

Character n-grams have effectively been used for authorship identification in prior work. In this study, the feature set consists of character trigram frequencies that were reweighted by the Inverse Document Frequency vector. Rather than determine the optimal values of maximum and minimum document frequency for this particular application, we obtained the entire feature set (39900 features) before performing dimensionality reduction, since being able to identify individual trigrams is not important for this task. We chose truncated single value decomposition (SVD) for this purpose, since it can handle sparse matrices efficiently [3]. We were able to reduce the dimensions to 250, while still capturing 93% of the variance.

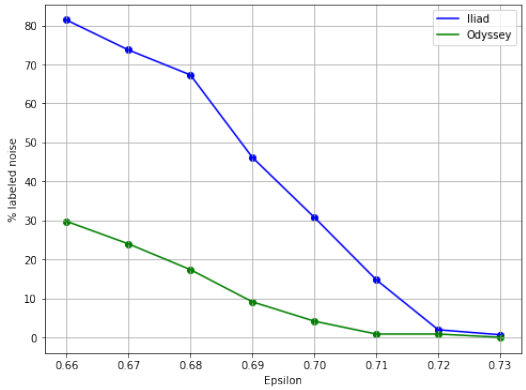

Fig. 1. Epsilon vs percent of poem labeled noise

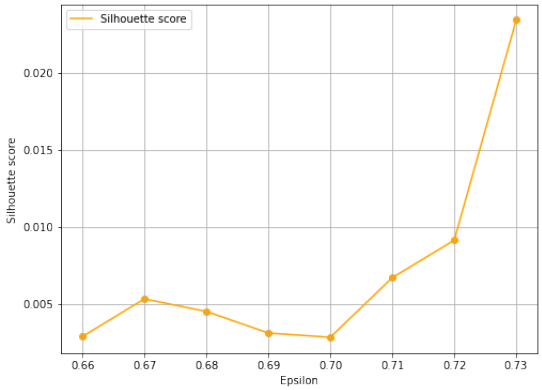

Fig. 2. Epsilon vs silhouette score

C. DBScan clustering

Our first line of investigation was to apply DBScan [4] to the entire dataset, varying the value of epsilon to identify clusters and outliers. The value of min_samples was fixed at 3: we judged 300 lines roughly long enough to identify a cluster, given that the shortest book in either the Iliad or the Odyssey is 331 lines, and that previous studies have applied clustering based on book divisions. Cosine similarity was selected for the distance metric. For each value of epsilon, we calculated the percentage of the Iliad and the Odyssey classified as noise, evaluated the clusters formed, and computed the silhouette score.

III. Results

A. Overview

Our preliminary results were as follows:

1) DBScan was significantly more likely to label chunks of the Iliad as noisy than chunks of the Odyssey (see Fig. 1).
2) At each value of epsilon tested, DBScan always modeled the data as consisting of 1 large cluster, 0-6 small clusters, and noise.
3) The small clusters never contained chunks from the Odyssey.
4) As a result of 3), any given chunk of the Odyssey was either part of the large cluster or noise.
5) Chunks from the Odyssey made up the majority of the large cluster, but it still contained some Iliad chunks at each value of epsilon tested.

What can we conclude from this? We expected to see two main clusters corresponding to the Iliad and Odyssey, with some chunks classified as noise and perhaps a few small clusters corresponding to later additions (eg: Book 10 of the Iliad, which made up its own cluster in Sandell & Bozzone's work [5]). Instead, our results are quite different. DBScan appears to have modeled the Homeric corpus as mostly falling into one large cluster (which we shall call the major cluster). This seems to roughly correspond to an "Odyssean cluster", in that most of its data points tend to be from the Odyssey.

The model also tends to identify 0-6 small clusters of the Iliad, consisting of roughly 300-500 lines each, which we shall call minor clusters. In addition, the model seems to identify the Iliad as far more noisy than the Odyssey.

Next, we must evaluate our clusterings and select an optimal value of epsilon. This is an unsupervised learning problem with no ground truth to evaluate our results against, but we can still use the silhouette scores for this purpose (Fig. 2.)

The higher silhouette scores from $\epsilon = 0.71 - 0.73$ are not particularly meaningful. At $\epsilon = 0.72$ DBScan identifies 4 noise points, and at $\epsilon = 0.73$ just 1. That leaves us with 2 potential values for ϵ: 0.67, 0.68, which we can evaluate in greater detail.

B. Analysis of minor clusters

TABLE I
MINOR CLUSTERS, $\epsilon = 0.67$

Cluster #	Passages
1	Il. 2.490-3.12
2	Il. 4.452-5.7, Il. 5.608-5.707, Il. 13.421-13.520, Il. 16.517-16.616, Il. 17.250-17.349
3	Il. 5.208-5.307, Il. 8.88-8.187, Il. 23.319-23.618
4	Il. 5.708-5.807, Il. 8.288-8.487

TABLE II
MINOR CLUSTERS, $\epsilon = 0.68$

Cluster #	Passages
1	Il. 2.490-3.12
2	Il. 4.452-5.7, Il. 13.521-13.620, Il. 14.385-14.484, Il. 16.517-16.616, Il. 17.250-17.349
3	Il. 5.108-5.307, Il. 8.88-8.187, Il. 13.421-13.520, Il. 20.248-20.347, Il. 23.319-23.618
4	Il. 5.708-5.807, Il. 8.288-8.487, Il. 10.518-11.38, Il. 19.372-20.47
5	Il. 11.539-11.639, Il. 17.150-17.249, Il. 17.650-17.749

Examining the minor clusters generated in tables 1 and 2, we see some similarities for $\epsilon = 0.67$ and $\epsilon = 0.68$. (Note that although each data point corresponds to 100 lines of text, contiguous chunks have been condensed for clarity in

the tables.) Cluster 1 for $\epsilon = 0.67$ and Cluster 1 for $\epsilon = 0.68$ both correspond to the same segment of Book 2 and the very beginning of Book 3. Cluster 2 for $\epsilon = 0.67$ and Cluster 2 for $\epsilon = 0.68$ both differ by only 1 segment each. The lines in Cluster 3 and Cluster 4 for $\epsilon = 0.67$ are subsets of the lines in Cluster 3 and Cluster 4, respectively, for $\epsilon = 0.68$. The only entirely unique cluster is Cluster 5 for $\epsilon = 0.68$.

C. Analysis of major cluster

TABLE III
ILIAD LINES IN MAJOR CLUSTER FOR ϵ = 0.67, 0.68

Major cluster membership	Passages
$\epsilon = 0.67$ and $\epsilon = 0.68$	Il. 1.1-1.100, Il. 1.301-1.500, Il. 2.190-2.289, Il. 2.390-2.489, Il. 3.13-3.312, Il. 7.270-7.469, Il. 9.27-9.326, Il. 9.631-10.217, Il. 10.318-10.417, Il. 19.172-19.271, Il. 24.322-24.521
$\epsilon = 0.67$ only	Il. 6.499-7.69, Il. 11.840-12.91, Il. 13.721-13.820
$\epsilon = 0.68$ only	Il. 1.201-1.300 Il. 3.413-4.51, Il. 4.252-4.351, Il. 6.199-6.298, Il. 13.821-14.83

The model classifies so many Odyssey chunks from every book as part of the major cluster that looking at these in detail is not particularly interesting. However, we can evaluate the chunks of the Iliad that the model sorts into the major cluster (perhaps the more "Odyssean" chunks of the Iliad?). Looking at table 3, some common patterns emerge: significant portions (300+ lines) of books 1, 3, 9, & 10, and smaller portions (200 lines) of books 2, 7, & 24, end up in the major classifier for both values of epsilon.

D. Analysis of noise

TABLE IV
ODYSSEY LINES IN NOISE FOR ϵ = 0.67, 0.68

Noise membership	Passages
$\epsilon = 0.67$ and $\epsilon = 0.68$	Od. 3.123-3.322, Od. 5.479-6.185, Od. 6.286-7.154, Od. 8.108-8.507, Od. 11.183-11.382, Od. 11.483-11.582, Od. 14.450-15.16, Od. 19.146-19.245, Od. 20.42-20.141, Od. 22.114-22.313, Od. 23.114-23.213, Od. 23.314-24.41, Od. 24.242-24.341
$\epsilon = 0.67$ only	Od. 3.23-3.122, Od. 5.279-5.478, Od. 12.43-12.142, Od. 13.190-13.289, Od. 14.50-14.149, Od. 19.346-19.545
$\epsilon = 0.68$ only	–

The model classifies so many Iliad chunks from every book as part of the noise group that looking at these in detail is not particularly interesting. However, we can evaluate the chunks of the Odyssey that the model considers noisy. Looking at table 4, some common patterns emerge: significant portions (300+ lines) of books 8 & 11 and smaller portions (200 lines) of books 3 & 22 are identified as noisy for both values of epsilon.

IV. RELATED WORK

While supervised learning is more commonly applied for authorship analysis than unsupervised learning, certain findings from these supervised techniques are still broadly relevant. Mosteller and Wallace were pioneers of quantitative authorship analysis; in classifying unknown documents within the Federalist Papers, they established broad guidelines for feature selection by demonstrating the effectiveness of function word frequency rather than content word frequency [6].

Recent work has supported the particular effectiveness of character n-grams, the approach used in this paper. Keselj et al. showed that character n-grams could achieve the best results on texts in multiple languages in an authorship attribution competition, standing out particularly in the case of multiple candidate authors that this project shares [7].

Manousakis and Stamatatos applied PCA, SVM, and intrinsic plagiarism detection methods to determine the authorship of Rhesus, a Greek play traditionally attributed to Euripides that scholars now consider to be the work of an unknown playwright. They employed character n-grams not only to determine authorship but to identify known sources of inspiration (other Greek playwrights) in the disputed play [8].

Finally, Sandell and Bozzone also utilized unsupervised learning algorithms in authorship analysis of the Homeric texts, although they used the traditional book divisions and employed hierarchical clustering for their final results. Their results showed distinct clusters of books within the Homeric texts that corresponded to scholarly opinion [5].

V. CONCLUSION

What can we make of these rather confusing findings? The most promising result comes from our analysis of the major cluster. Large (200+ lines) chunks of books 1,2,3,7,9,10, and 24 were sorted into the major cluster alongside a greater proportion of Odyssey chunks. Perhaps surprisingly, these "Odyssean" books correspond perfectly to the books of the Iliad considered linguistically or thematically unusual by scholars [5]. Correlating the minor clusters with scholarly analysis on a line-by-line level is possible, but probably outside the scope of this paper (although the fact that Il. 2.490-3.12 was clustered by itself is promising, as it is book 2 that contains the notoriously-weird Catalogue of Ships.)

Our results also tentatively suggest some utility to evaluating the Homeric texts outside of the traditional book divisions. As one example, roughly the first half of Book 2 was clustered with the major cluster, but the second half was in a cluster by itself, which indicates stylistic variation within the book. Furthermore, even though the Iliad books identified as unusual through membership in the major cluster roughly corresponded

to scholarly views, noticeable chunks of these unusual books were nevertheless missing from the major cluster. For example, even though Book 10 is by far the most readily-identified later addition to the Iliad, lines 10.218-10.317 and 10.418-10.517 were identified as noise rather than as part of the major cluster!

Some of this variation within books, of course, could be explained by artifacts of the admittedly limited methodology used in this experiment. For example, varying the stagger size used to make 100-line divisions, experimenting on slightly different feature sets, or increasing/decreasing the number of lines per division, could all account for some of these oddities. However, previous studies suggest that authorship analysis can be done accurately using even smaller document lengths than the 100 lines used in this study. Furthermore, in the domain of Greek and Latin texts, it is not uncommon for the veracity of an individual line can be doubted. Future computational analyses of the Homeric corpus might benefit by exploring similar smaller divisions and attempting to identify breaks in authorship on a more granular level.

Another limitation of this work is that it was less successful at identifying potential authorship anomalies within the Odyssey. The largest groups of noisy Odyssey segments came from books 8 and 11, which are not obvious candidates for interpolations/later additions in the way that books 2 or 10 of the Iliad are. That being said, we identified several lines within each noisy Odyssey chunk that are unusual linguistically or are otherwise suspected of interpolations/corruption, so perhaps further (and more granular) experimentation is needed. Another explanation for the model not identifying much noise or any minor clusters within the Odyssey is that the Odyssey simply has less internal variation than the Iliad. This hypothesis could also be tested in further computational analyses: perhaps the Odyssey might be more accurately modeled as made up of fewer clusters of distinct authorship than the Iliad.

VI. References

References

[1] Nagy, Gregory. Homeric Questions. 1st ed. University of Texas Press, 1996. http://nrs.harvard.edu/urn-3:hul.ebook:CHS_Nagy.Homeric_Questions.1996

[2] Classical Language Toolkit. "The Classical Language Toolkit." Accessed December 3, 2024. http://cltk.org.

[3] Scikit-learn. "Clustering Text Documents Using k-Means." Scikit-learn 1.5.0 Documentation. Accessed December 5, 2024. https://scikit-learn.org/1.5/auto_examples/text/plot_document_clustering.html.

[4] Scikit-learn. "sklearn.cluster.DBSCAN: Perform DBSCAN Clustering from Vector Arrays or Distance Matrices." Scikit-learn Documentation. Accessed December 16, 2024. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html.

[5] Sandell, R., and C. Bozzone. "One or Many Homers? Using Quantitative Authorship Analysis to Study the Homeric Question." Academia.edu. Accessed November 1, 2024. https://www.academia.edu/34546901/One_or_Many_Homers_Using_Quantitative_Authorship_Analysis_to_Study_the_Homeric_Question.

[6] Mosteller, Frederick, and David L. Wallace. "Inference in an Authorship Problem." Journal of the American Statistical Association 58, no. 302 (June 1, 1963): 275–309. doi:10.2307/2283270.

[7] Keselj, Vlado, Fuchun Peng, Nick Cercone and Calvin Thomas. "N-gram-based Author Profiles for Authorship Attribution." Proc. of the Conference Pacific Association for Computational Linguistics (2003).

[8] Manousakis, N., and E. Stamatatos. "Devising Rhesus: A Strange 'Collaboration' between Aeschylus and Euripides." Digital Scholarship in the Humanities 33, no. 2 (January 1, 2018): 347-361–361. doi:10.1093/llc/fqx021.