Attribution of Contested and Anonymous Ancient Greek Works

Sarah Beller and James Spicer
{sfbeller, jspicer}@stanford.edu

December 12, 2014

Abstract

Authorship attribution has been a persistent problem in the Classical genre, as texts that reach us from antiquity are often corrupted, edited, or forged over the thousands of years since their initial production. Scholars have worked on identifying writers' stylistic differences in an attempt to distinguish genuine texts from fakes, and to attribute an author to previously anonymous works. Increasing computing power allows the derivation of more complex features, giving us new information about each author's linguistic signature and writing style. Our system is able to accurately predict the author of a complete anonymous work, as well as many text fragments that currently have contested authorship. We experimented with using semantic and lexical features, and explored both discriminative and generative classification algorithms. Our highest-performing system achieved an attribution accuracy of 85.7%.

1 Prior Work

Scholars have discussed means of determining authorship since antiquity. Rigorous modern authorship attribution studies began in earnest in the nineteenth century, when it was described as 'stylometry.' The famous study of the Federalist Papers by Mosteller and Wallace in 1964 publicized the field, and the advent of modern computing has increased the scope of research. Early studies were hampered by computational limitations, and common algorithms tended to overfit data when feature dimensionality became too large [2]. The development of faster computers as well as new machine learning algorithms allowed researchers to overcome this issue, since newer classifiers were better able to deal with higher dimensions. Recent work has combined lexical and syntactic measures, leading to promising initial results [3].

2 Data

The entirety of the Classical corpus is digitized and available online through the Perseus Digital Library [1] as XML files. We use a selection of 69 of the most influential texts that ranged in age from the 8^{th} century BCE to the 2^{nd} century CE. The 63 works with known authorship make up the training set, and the 6 works with either unknown or contested authorship make up the test set. The texts' genres include epic poetry, prose, tragedy, comedy, and history.

3 Features & Preprocessing

Beginning with the ancient Greek words, we implement our own data processing to increase model accuracy by reducing feature dimensionality. We remove accents and stem words to their root by removing noun and verb endings. Proper nouns including character names and place names are ignored so that our models are as non-subject-specific as possible.

To get baseline accuracies against which further work can be compared, we first train our classifiers only on word frequencies ('bag-of-words'). We then derive ten other lexical features, including words per line, syllables per word, and the frequency of various parts of speech such as prepositions, particles, and *hapax legomena* (words that appear only once in the entire classical corpus). The focus on words' context rather than their meaning isolates the work's writing style rather than topic, which according to Morton [5] leads to superior discrimination between authors writing in the same culture and language.

4 Models

We implemented four different classification algorithms: Naive Bayes, Support Vector Machines (SVMs), K Nearest Neighbors (KNN), and Decision Trees. All were trained on the training matrix X ($> 10,000$ features $\times$ 63 works) and corresponding class label vector Y (63 author labels). To improve performance, we assign each author a unique number so that the models are manipulating integers rather than strings.

Naive Bayes

Naive Bayes is a generative model that assumes all features of a data point are independent. For a class C and features F:
$\log p(C|F) = \log[p(C)] + \sum_{i=1}^{n} \log p(F_i|C)$
It is less prone to overfitting than other models, which is important for this project due to the relatively small dataset. In general, generative models excel with little data.

SVM

SVMs are discriminative models that map data points into two separate categories that are as widely separated as possible. We use the many auxiliary binary models created by LibSVM to give a pseudo-multiclass SVM model. We assume that authorship categorization is linearly separable. SVM Optimimization problem:
$\max_a W(\alpha) = \sum_{i=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$
s.t. $\alpha_i \geq 0, i = 1, ..., m$ & $\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$
SVMs are used because they work well with high dimensions. Furthermore, author detection with SVMs on full word forms has been shown to be remarkably robust, even if the author wrote about different topics [6].

KNN

KNN is a discriminative model that weighs the label of each training point according to how closely it matches the query point. To find k nearest neighbors of data point $X_n^{(i)}$: Choose 0 if $\sum_{i=1}^{k} Y_n^{(i)} \leq k/2$ and 1 if $\sum_{i=1}^{k} Y_n^{(i)} \geq k/2$. We used KNN because it performs well with evenly-distributed, continuous variables, so is suited to our dataset where works are spread between a relatively large number of authors.

Decision Trees

Decision trees are discriminative models that create a tree data structure with class label 'leaves' and feature 'branches'. They perform well with large datasets with a high number of both features and class labels.

Discriminant Analysis

We attempted to implement both LDA and QDA, but the two algorithms struggled with the high ($> 10,000$) dimensions required by our program.

5 Results

Due to the small size of the dataset, we used leave-one-out cross validation on the training set:

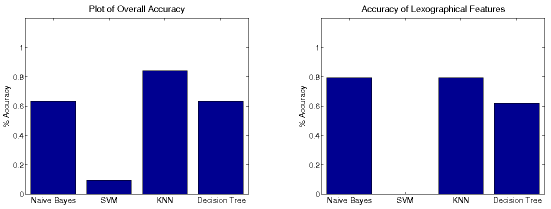
Model	Baseline (%)	Best (%)
Naive Bayes	60.32	85.71
SVM	9.52	0
KNN	80.95	84.13
Decision trees	63.49	58.73

Table 1: Accuracy of baseline and best models.

We also used our most accurate models to predict authorship on contested and anonymous works (see Table 2).

6 Discussion

There are several difficulties inherent in authorship attribution based on writing style. Our data is skewed by class imbalance: different authors have different numbers of surviving texts,



(a) Accuracy on entire feature set

(b) Accuracy on best feature set

and some works are longer than others. Figure 3 shows this variation.

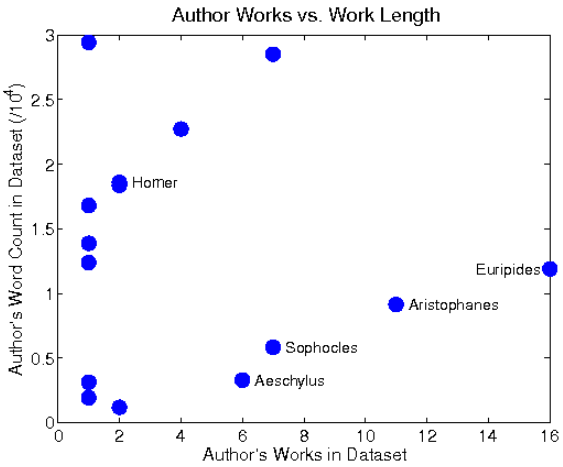


Figure 3: Number of works compared to total words, by author.

Another issue is the lack of consistent writing style throughout longer works, including plays in which different characters are expected to speak in different ways. Our most significant problem was balancing our sparse data with the high dimensionality of our feature set and high number of class labels. It was this issue that resulted in the poor performance of SVMs, which is primarily a binary classifier.

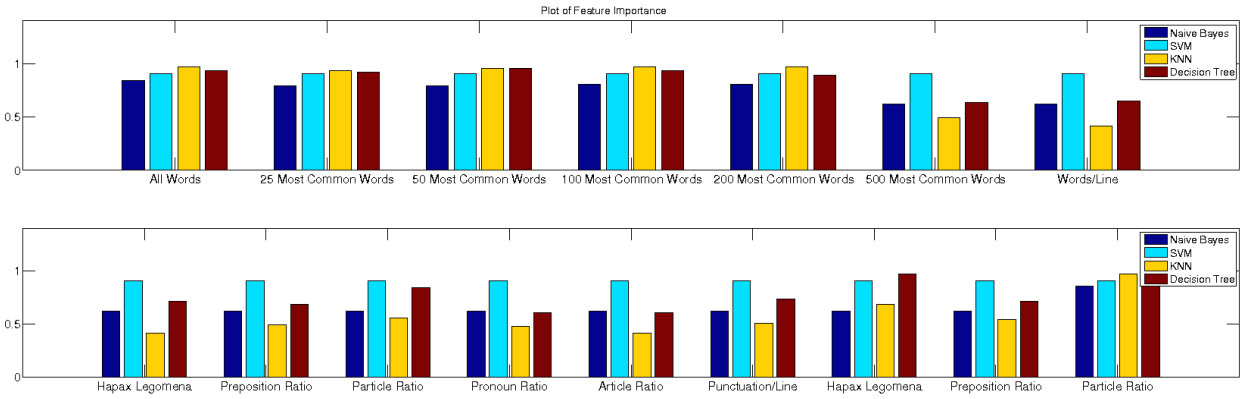


Figure 1: Performance for each lexical feature on its own.

Work	**Naive Bayes**	**KNN**	**Decision Tree**
Prometheus Bound	Euripides	Euripides	Aeschylus
Iphigenia at Aulis	Euripides	Euripides	Euripides
Phoenissae	Euripides	Euripides	Sophocles
Rhesus	Aeschylus	Euripides	Aeschylus
The Shield of Heracles	Hesiod	Hesiod	Aeschylus
Homeric Hymns	Pindar	Pindar	Hesiod

Table 2: Authorship predictions on unknown and disputed works.

7 Conclusions

We noticed a significant performance improvement when we trained on only the most common words in the data, resulting in our top accuracy when we combined the 100 most common words with our derived lexical features (See Table 1). In fact, this combination proved more accurate than when we included trained on all words. This is due to fact that the rarest words are often topic specific, and so will hinder prediction rather than help it.

As Figure 1 shows, the lexical features we derived proved remarkably accurate even by themselves. The ratios of various parts of speech (prepositions, particles, articles, etc.) performed particularly well, perhaps because they are most indicactive of a particular writer's style.

The predictions on the anonymous and contested works lead to interesting conclusions:

Prometheus Bound is usually attributed to Aeschylus, although modern scholars are divided on the play's authenticity due to non-Aeschylean meter, style, and portrayal of Zeus. Our 2-1 split against Aeschylus is not too unsurprising, however, as Euripides and Aeschylus share the same genre and many stylistic features.

Iphigenia at Aulis, *Phoenissae*, and *Rhesus* are all attributed to Euripides, but have had their authorship called into question due

to their stylistic differences from the rest of Euripides' work. Our classifiers' split on each one reflects the academic view that the style is certainly Euripidean, but with many foreign elements introduced.

The *Shield of Heracles* was viewed as an imitation of Hesiod's epic poetry as early as the Hellenistic period; at times the text even copies directly from the *Iliad*. Modern scholars are not in consensus, but the common view is of the work as being in the style of Hesiod, a view reflected by our classifiers' 2-1 result.

Lastly, the *Homeric Hymns* are a set of hymns to Greek gods, ascribed to Homer from antiquity. It is agreed that the poems imitate Homer's style but were written centuries after his death. It is a notable success for our program that none of the classifiers ascribed *Homeric Hymns* to Homer despite its eponymous similarity to Homer's works.

8 Future Work

Our next steps for the project include using a feature selection algorithm to reduce the dimensionality of our feature set. Although a reliable POS tagger does not exist for Ancient Greek, the Perseus Project has a small dependency treebank for a small number of texts, which we will use to develop a more robust and less context-specific system. Additionally, we plan to refine and split several of our lexical features from broader categorizations into more specific features. On a larger scale, we will also use the system to predict authorship for unclassified fragments of larger works, as well as the larger works themselves.

References

[1] www.perseus.tufts.edu

[2] Stamatatos, Efstathios (2009), "A survey of modern authorship attribution methods." *Journal of the American Society for Information Science and Technology* 60: 538-556.

[3] Stamatatos, Efstathios, Fakotakis, Nikos, and Kokkinakis, George (2001), "Computer-based Authorship Attribution without Lexical Measures." *Computers and the Humanities* 35: 193-214.

[4] Michaelson, S. and Morton, A.Q. (1972), "The New Stylometry: A One-Word Test of Authorship for Greek Writers." *The Classical Quarterly*, New Series 22.1: 89-102.

[5] Morton, A.Q. (1965), "The Authorship of Greek Prose." *Journal of the Royal Statistical Society*, Series A (General), 128.2: 169-233.

[6] Dietrich, Joachim, Kindermann, Jrg, Leopold, Edda and Paass, Gerhard (2003), "Authorship Attribution with Support Vector Machines." *Applied Intelligence* 19: 109-123.