Clustering Tasks and Decision Trees with Augustan Love Poets: Cohesion and Separation in Feature Importance Extraction*

Carlos Javier Nusch[1,2,*,†], Gimena del Rio Riande[3,†], Leticia Cecilia Cagnina[4,†], Marcelo Luis Errecalde[4,†] and Leandro Antonelli[5,6,†]

[1]PREBI, SEDICI, Universidad Nacional de La Plata, Argentina
[2]CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Argentina
[3]IIBICRIT, Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
[4]LIDIC, Facultad de Ciencias Físico Matemáticas y Naturales, Universidad Nacional de San Luis, Argentina
[5]LIFIA, Facultad de Informática, Universidad Nacional de La Plata, Argentina
[6]CAETI, Facultad de Tecnología Informática, Universidad Abierta Interamericana, Argentina

Abstract

This article extends various automatic text analysis tasks from previous works by applying natural language processing techniques to a corpus of Latin texts from the 1st century BC and 1st century AD. The motivation behind this work is to delve into and understand a historical literary trend revolving around the themes of love, spanning from antiquity through to the medieval period. The analyzed authors include Gaius Valerius Catullus, Albius Tibullus, and Sextus Propertius, representing the literary movement of the neoterics, and Publius Vergilius Maro and Marcus Annaeus Lucanus, epic poets with distinct styles, serving as control samples. Unlike previous works, various corrections were added to the preprocessing tasks, including improved word tokenization with enclitics and handling of orthographic variances. For the clustering tasks, the K-Means method and the Silhouette Score were used to determine the optimal cluster sizes. Using these optimal clusters as labels, decision trees were trained for each range of n-grams, aiming to identify features with the highest Information Gain and Information Gain Ratio. The trees were trained based on the criterion of Entropy, and calculations of Feature Importance were performed. In this study, we focused on detailing the classification results and features extracted by the decision trees, based on the best Silhouette scores obtained and the Information Gain. We examined whether the words or parts of words with classificatory potential identified in the process matched the findings from previous exploratory tasks performed using other techniques.

Keywords

Augustan love poets, Document Clustering, K Means, Silhouette Coefficient, Decision Trees, Feature Importance, Information Gain Ratio

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark
*Corresponding author.
†These authors contributed equally.
carlosnusch@prebi.unlp.edu.ar (C. J. Nusch)
https://prebi-sedici.unlp.edu.ar/personal/carlos-nusch/ (C. J. Nusch)
0000-0003-1715-4228 (C. J. Nusch); 0000-0002-8997-5415 (G. d. R. Riande); 0000-0001-7825-2927 (L. C. Cagnina); 0000-0001-5605-8963 (M. L. Errecalde); 0000-0003-1388-0337 (L. Antonelli)

1. Introduction

This study[1] builds on a master's thesis [1] examining C. S. Lewis' observations [2] on the influence of Courtly love and Occitan literature on 20th-century love imagery. Similarities in love themes, treatment of the beloved, and political and military terms were found between Occitan and 1st-century BC Latin poetry. This thesis aims to identify textual patterns linking ancient love themes to the *Religion of Love* in medieval Occitan poetry, using a comparative approach that combines close reading with computational methods [3, 4]. This article evaluates clustering techniques for differentiating love poems from other Latin poetry, identifying key lexical features. Previous work [5] explained the techniques, while here we focus on feature extraction and optimal Silhouette Score values.

1.1. State of the Art

Several authors have applied clustering to ancient texts. Bracco et al. [6] used K-means to detect literary genres in cuneiform texts, and Martins et al. [7] used k-Nearest Neighbors for author classification. Cantaluppi and Passarotti [8] studied Seneca's complete works, Cicero's orations, Jerome's Latin New Testament, and Aquinas' major works. Nagy [9] used multivariate analysis and clustering to examine rhyme in twelve classical Latin poets, identifying stylistic differences between genres and authors. In recent work, he applied UMAP and t-SNE to show stylistic distinctions between Ovid's *Heroides* and other works, and the authenticity of the *Epistula Sapphus* [10]. Forstall et al. [11, 12] compared lexical and rhythmic features at character and word n-gram levels with other 1st-century BC poets.

1.2. Problem Definition and Contributions

The previous work aimed to explore clustering techniques to distinguish love poems from other types of poetry and identify useful lexical characteristics for classification. The K-means algorithm [13] was used, and the optimal number of clusters was determined with the Silhouette Index [14], which measures group cohesion and separation. Since K-means, based on Euclidean distance, does not provide detailed feature extraction, decision trees [15] were used to complement this approach. This combination allowed for the indirect extraction of features, with metrics such as Importance, Information Gain, and Information Gain Ratio [16] identifying the most relevant features.

2. Research Methodology and Approach

2.1. Analysis Corpus and Used Editions

The corpus includes the complete works of Gaius Valerius Catullus [17], Albius Tibullus [18], and Sextus Propertius [19], representing love poetry, as well as all books from the *Aeneid* by

[1] An appendix with key tables is included after the references. A larger dataset is available: Nusch, C. (2024). Clustering Tasks and Decision Trees with Augustan love poets [Data set]. CHR2024, Aarhus, Denmark. Zenodo. https://doi.org/10.5281/zenodo.12682694.

Publius Vergilius Maro [20] and *Pharsalia* by Marcus Annaeus Lucanus [21] as control samples, focused on political, historical, and martial themes. The analysis reveals differences in the number of words and verses per poem among different authors and genres (Table 1). To address concerns about unbalanced datasets, we used relative frequency and separated the authors to reduce noise and bias from the larger epic texts. Two datasets were used: one with the Augustan love poets and Vergil, and another with the Augustan love poets and Lucan.

Table 1
Summary of works and word statistics from various Latin authors.

Author (Work)	Verses	Total Words	Unique Words	Avg. Words Poem/Canto
Catullus (Merrill, 1893)	2289	12912	5802	110.35
Tibullus (Müller, 1898)	1930	12368	5201	334.27
Propertius (Postgate, 1915)	4008	25450	9809	242.38
Lucanus *Pharsalia* (Weise, 1935)	8061	51215	14750	5121.5
Virgilius *Aeneid* (Greenough, 1900)	9896	63896	16616	5324.66

To construct the analysis corpus, resources from the Perseus Project digital library [22, 23] at Tufts University were used. The library contains 2,412 works in 3,192 editions and translations (1,639 in Greek, 636 in Latin) and a total of 69.7 million words. The texts, curated by specialists and shared under a CC BY SA 3.0 (US) license, are available in XML format. Additional resources include models for grammatical tagging and stopwords for Latin. The poems were harvested through web scraping using R, while Python libraries were employed for text analysis and mining.

The analysis explored character n-grams (2 to 7) and word n-grams (1 to 5), using the Bag of Words (BOW) method [24]. Three types of matrices were generated: the first based on raw frequency (using Scikit-learn's CountVectorizer), the second on relative frequency (with a custom function), and the third using the TF-IDF technique [24], which highlights important words by weighing their frequency relative to their rarity across the dataset. While CountVectorizer simply counts word occurrences, TF-IDF reduces the impact of common words, giving more weight to unique terms.

2.2. Text Preprocessing Tasks

Before analysis, the text was cleaned by removing empty lines, sequences of spaces ("\n \n\n\n"), and editorial symbols for illegible gaps ("†"). Spanish quotation marks were replaced with English ones for tool compatibility, and punctuation was removed from character n-grams, as it was added by editors.

For stopwords[2], we used the Stopwords ISO [27] package for Latin, which we preferred over the Perseus Project version because it retains important words in elegiac poetry, such as *ego*, enabling the analysis of personal pronouns—a significant feature noted in previous works [1, 28].

To enhance tokenization, we added two procedures from The Classical Language Toolkit (CLTK) [29]: JVReplacer, to standardize spellings (e.g., *Iulius/Julius* and *uir/vir*), and LatinWordTokenizer, which helped identify enclitics (e.g., *-que*, *-ve*) and prevent incorrect tokenization.

[2]For a more detailed discussion of the complexity and variety of stopwords in Latin and other ancient languages, see A. Berra [25] and P.J. Burns [26].

3. Evaluation: The Clustering and Decision Trees as combined techniques

As explained previously [5], document clustering was performed using the K-means method and Silhouette scores to evaluate the best cluster configuration. The optimal number of clusters (k) was determined by testing k values from 2 to 20. Tests were conducted using fixed ranges of character n-grams (2 to 7) and word n-grams (1 to 5), with the Silhouette coefficient calculated for each k. The aim was to find both the best k and the most effective n-gram ranges for clustering.

Once the data was labeled, decision trees were trained using the entropy criterion to assess feature importance, with Information Gain (IG) and Information Gain Ratio (IGR) calculated.

4. Preliminary or Intermediate Results

Better Silhouette Scores were achieved using the raw frequency matrix with simple stopwords filtering (CountVectorizer with Stopwords), while the Relative Frequency and TF-IDF Matrices showed lower scores. TF-IDF scores were close to zero, indicating poor cluster separation and Relative Frequency values were around 0.5 for both datasets (Tables 2 and 3). The use of relative frequency significantly impacted the optimal number of clusters recommended by the K-means algorithm.

Table 2
Optimal clusters and corresponding Silhouette values for different ranges of n-grams (Corpora of Catullus, Tibullus, Propertius, and Vergilius).

	Raw Frequency		Relative Frequency		TF-IDF	
N-gram Type	**Clusters**	**Score**	**Clusters**	**Score**	**Clusters**	**Score**
Char 2-grams	2	0.94	7	0.19	7	0.16
Char 3-grams	2	0.93	2	0.534	3	0.07
Char 4-grams	2	0.904	2	0.536	3	0.031
Char 5-grams	2	0.85	5	0.446	2	0.008
Char 6-grams	2	0.801	5	0.443	15	0.008
Char 7-grams	2	0.76	5	0.44	15	0.004
Word 1-grams	2	0.802	2	0.55	2	0.013
Word 2-grams	2	0.718	2	0.52	18	0.001
Word 3-grams	2	0.719	2	0.54	19	0.007
Word 4-grams	2	0.72	2	0.56	10	0.007
Word 5-grams	2	0.721	2	0.58	19	0.007

Table 3
Optimal clusters and corresponding Silhouette values for different ranges of n-grams (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

	Raw Frequency		Relative Frequency		TF-IDF	
N-gram Type	**Clusters**	**Score**	**Clusters**	**Score**	**Clusters**	**Score**
Char 2-grams	2	0.94	2	0.19	2	0.15
Char 3-grams	2	0.92	3	0.53	3	0.074
Char 4-grams	2	0.89	2	0.53	4	0.031
Char 5-grams	2	0.85	5	0.44	2	0.009
Char 6-grams	2	0.801	2	0.53	4	0.004
Char 7-grams	2	0.79	2	0.55	18	0.004
Word 1-grams	2	0.802	2	0.52	2	0.011
Word 2-grams	2	0.74	2	0.38	18	0.001
Word 3-grams	3	0.74	3	0.54	17	0.003
Word 4-grams	3	0.75	2	0.56	15	0.008
Word 5-grams	3	0.75	2	0.58	19	0.007

The new tokenization and normalization process using CLTK modules had a noticeable impact. Regarding the most critical features for classifying clusters, results suggest that the methodology and resources should be reevaluated. In previous work, high Silhouette scores were observed with the frequency table, but feature importance metrics showed an uneven distribution, with one or two attributes dominating. While TF-IDF identified more features, the low Silhouette scores indicated poor classification. Despite better Silhouette scores, the same issue occurred with the relative frequency matrix.

4.1. Feature Extraction at Character N-Grams Level

As shown in Figure 1 the clustering task succeeded in separating Augustan love poets from epic poets. In the next page, Tables 4, and 5 show the feature extraction with this n-gram level.

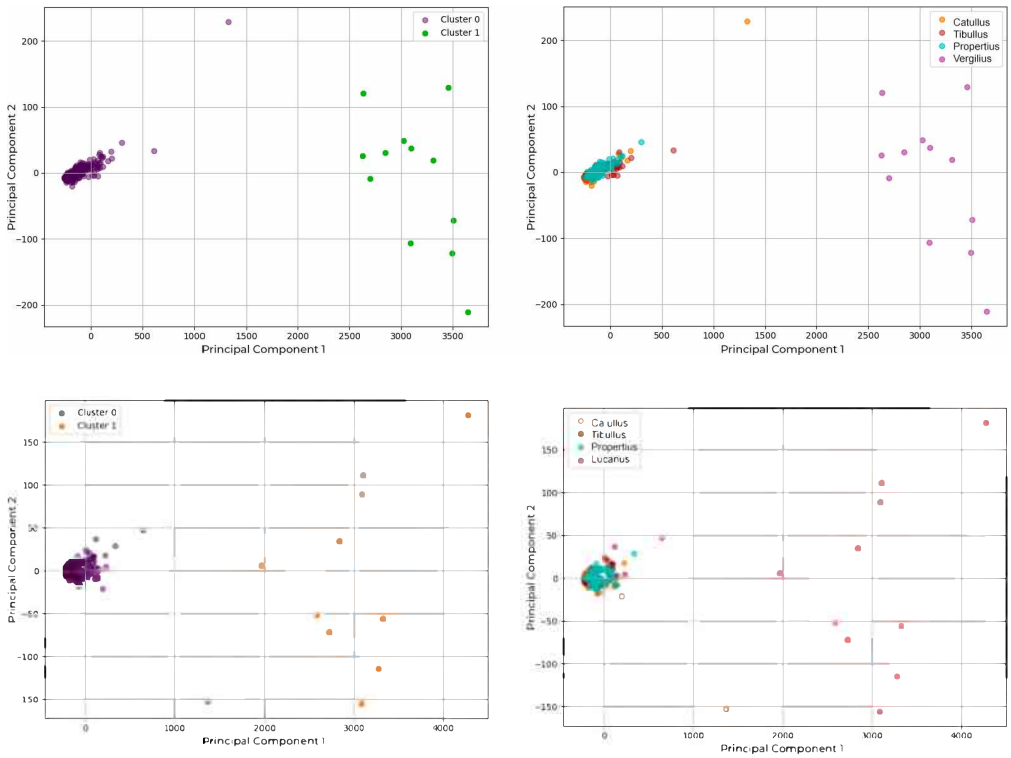

Figure 1: Scatter plot of clustering by K Means using a raw frequency matrix of 2 character n-grams indicating the clusters (left) and authors (right) with different colors.

Table 4
Importance features (2 of character n-grams) using the raw frequency matrix method and Stopwords filtering (Corpora of Catullus, Tibullus, Propertius, and Vergilius).

Feature	Importance	Feature	IG	Feature	IGR
un	1	nl	0.1254	dg	0.5747
		uq	0.1226	aë	0.5463
		dg	0.1208	oï	0.5193
		gg	0.1205	ën	0.4929
		ms	0.1205	ïa	0.4929
		gm	0.1205	ez	0.4664
		dh	0.1181	aï	0.4664
		bt	0.1174	dh	0.4516
		yd	0.1129	ï	0.439
		dl	0.1128	eï	0.439
		nh	0.1128	ër	0.439
		ze	0.1128	mf	0.4105
		bn	0.1072	x	0.376
		my	0.1068	oö	0.3758
		ln	0.105	ë	0.3758
		rh	0.105	ïc	0.3758
		aë	0.1049	ön	0.3758
		df	0.1036	ïu	0.3758
		yt	0.0999	oë	0.3758
		yc	0.0999	gg	0.3724

Table 5
Importance features (2 of character n-grams) using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Feature	Importance	Feature	IG	Feature	IGR
g	1	ye	0.1464	ye	0.5942
		gm	0.1234	dh	0.4673
		dh	0.1151	mt	0.4094
		ya	0.1048	bf	0.3991
		by	0.1025	gm	0.3975
		xq	0.097	sf	0.3797
		ze	0.0938	dt	0.3708
		dt	0.0914	fc	0.3514
		oh	0.0909	fs	0.3514
		rh	0.0894	pc	0.3514
		oa	0.0879	nb	0.3514
		sn	0.087	dg	0.3514
		dq	0.0864	cp	0.3514
		yp	0.0864	sn	0.3306
		df	0.0858	bm	0.3287
		gy	0.0858	y	0.323
		ee	0.0858	mf	0.323
		yc	0.085	xq	0.3126
		sy	0.085	ms	0.2979
		lm	0.0836	ze	0.2883

In the case of relative frequency, after excluding Lucan, the best Silhouette Score at the character n-gram level was achieved with 4-grams. However, despite obtaining a relatively good score (0.53), the resulting classification did not meet expectations (Figure 2). The algorithm constructed two clusters, one of which contained only *Carmen 94*. A similar phenomenon occurred when excluding Vergil, but with the distinction that the isolated *Carmen* was *112*.

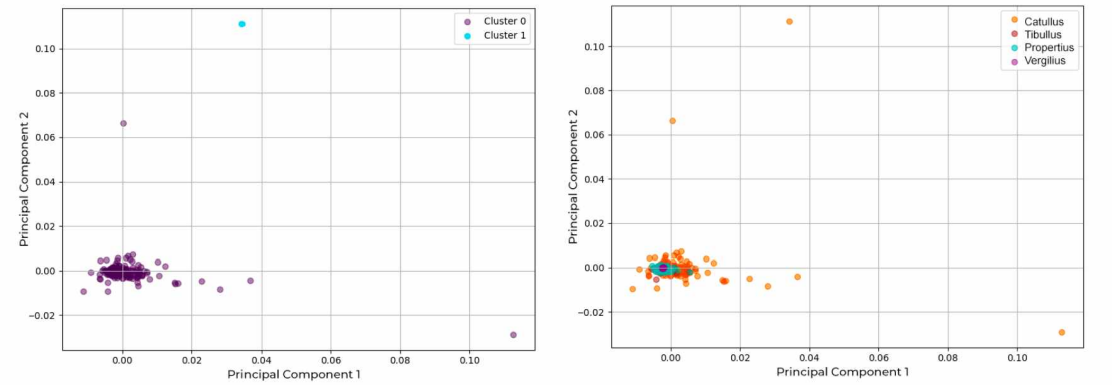

Figure 2: Scatter plot of clustering by K Means using a relative frequency matrix of 4 character n-grams indicating clusters (left) and authors (right) with different colors.

4.2. Feature Extraction at Word N-Grams Level

At the word n-gram level, similar to the character n-gram level, the best classification method was the raw frequency matrix. Although the relative frequency matrix also yielded good Silhouette scores, it consistently produced poor classifications, isolating only *Carmen 94* from the rest. At other n-gram levels, the *carmina* that were separated included *Carmina 14*, *82*, *85*, and *106*. All of these are relatively short, suggesting that the difference in length among the poems introduces internal variability in the corpus that hinders classification based on relative frequencies. The same task, when performed using raw frequencies, yielded excellent results, whether Lucan or Vergil was excluded from the analysis (Figure 3 and Tables 6, 7).

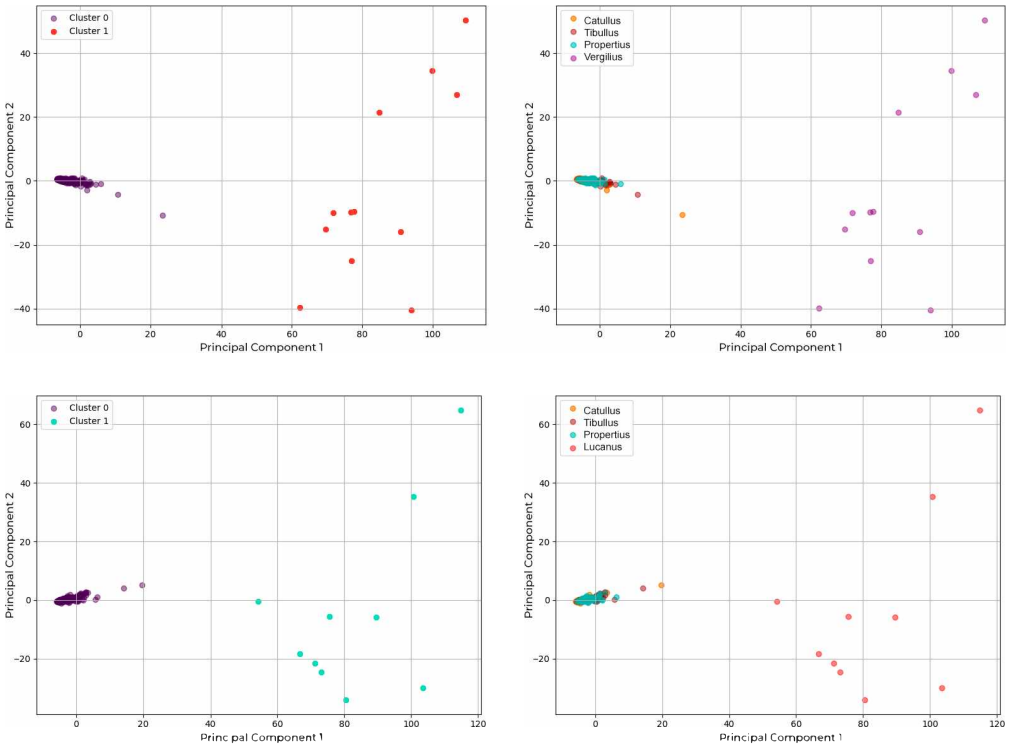

Figure 3: Scatter plot of clustering by K Means using a raw frequency Matrix of 1-word n-grams clusters (left) and authors (right) with different colors.

Table 6
Most important features at the level of 1-word n-grams ranked by Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Vergilius).

Raw Frequency Word (1, 1) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
urbem	1	aeneas	0.1813	teucrum	0.6931
		ingentem	0.1813	ingens	0.6931
		ingens	0.1813	ingentem	0.6931
		teucrum	0.1813	aeneas	0.6931
		fatis	0.1683	pius	0.6413
		omnipotens	0.1683	aethere	0.6413
		late	0.1683	ast	0.6413
		ignem	0.1683	fatur	0.6413
		auras	0.1683	diuom	0.6413
		iamque	0.1601	socios	0.6413
		ea	0.1601	clamore	0.6413
		genitor	0.1601	teucros	0.6413
		terram	0.1601	visu	0.6413
		talibus	0.1601	ignem	0.6061
		equidem	0.1601	omnipotens	0.6061
		diuom	0.1571	fatis	0.6061
		pius	0.1571	late	0.6061
		visu	0.1571	auras	0.6061
		teucros	0.1571	teucri	0.6056
		ast	0.1571	regem	0.6056
		aeneas	0.1571	ast	0.6056

Table 7
Most important features at the level of 1 word n-grams according to Importance, Information Gain, Information Gain Ratio using the raw matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Raw Frequency word (1, 1) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
pectora	1	populos	0.1589	populis	0.6931
		scelerum	0.1589	scelerum	0.6931
		populis	0.1589	bellorum	0.6931
		mundo	0.1589	uiscera	0.6931
		bellorum	0.1589	exit	0.6931
		exit	0.1589	populos	0.6931
		uiscera	0.1589	mundo	0.6931
		senatus	0.1464	caussa	0.636
		ciuilia	0.1464	nocentes	0.636
		nefas	0.1464	ciuilibus	0.636
		fatis	0.1464	coelo	0.636
		bellum	0.1388	libye	0.636
		milite	0.1388	robore	0.636
		ducis	0.1388	superi	0.636
		ciuilibus	0.1345	adhuc	0.636
		fauces	0.1345	ciuile	0.636
		robore	0.1345	fauces	0.636
		adhuc	0.1345	coeli	0.636
		caussa	0.1345	malorum	0.636
		libye	0.1345	potuere	0.5968

In the following table and figure, it can be observed that the use of relative frequency brings forth personal pronouns, terms previously associated with love poetry in earlier studies. However, these should be disregarded when obtained through this methodology, as the classifications achieved with them were quite poor, as can be seen in Table 8 and Figure 4.

Table 8
Most important features at the level of 1 word n-grams according to Importance, Information Gain, Information Gain Ratio using the relative frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Vergilius).

Relative Frequency word (1, 1) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
legit	1	moechatur	0.0244	moechatur	0.6931
		olera	0.0244	olera	0.6931
		olla	0.0244	olla	0.6931
		mentula	0.0151	mentula	0.114
		dicunt	0.0132	dicunt	0.0689
		legit	0.0124	legit	0.0542
		certe	0.009	certe	0.0209
		ipsa	0.0061	ipsa	0.0087
		mihi	0.003	mihi	0.0031
		tibi	0.003	tibi	0.003
		tu	0.0024	tu	0.0024
		nunc	0.0021	nunc	0.0022
		quid	0.0019	quid	0.002
		ne	0.0019	ne	0.002
		ego	0.0018	ego	0.0019
		mea	0.0018	mea	0.0019
		esse	0.0018	esse	0.0019
		iam	0.0018	iam	0.0018
		illa	0.0016	illa	0.0017
		nam	0.0015	nam	0.0017

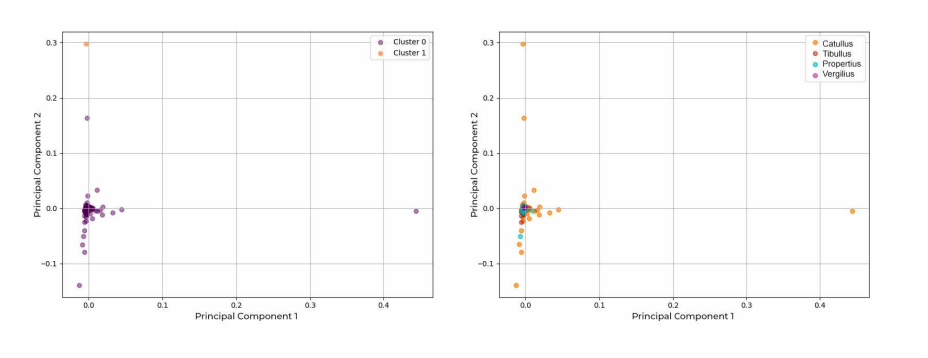

Figure 4: Scatter plot of clustering by K Means using a relative frequency matrix of 1-word n-grams indicating clusters (left) and authors (right) with different colors

Both datasets, whether excluding Lucan or Vergil, showed identical performance in author classification. Character n-grams (2 to 6) and single-word n-grams effectively separated epic authors from Augustan love poets using the raw frequency matrix, with Silhouette Scores above 0.8[3]. Lower scores led to suboptimal classifications, where one cluster contained only a single book (e.g., Book X or XII of the *Aeneid* or Book IX of *Pharsalia*).

Assessing the relevance of specific character n-grams for classification remains challenging, requiring a more detailed stylistic investigation. In summary, document grouping was effective, though feature-level techniques did not always highlight typical elegiac terms. The terms extracted via decision trees for Augustan love poets and Vergil predominantly reflected epic, mythical, and martial language[4].

[3]For more details on the extracted terms, see Appendix A and B.

[4]Please note that with the English quotation marks we have attempted to indicate the spaces before or after the words, in cases where it corresponds to the character n-gram.

4.3. Data from the Corpora of Catullus, Tibullus, Propertius, and Vergil:

- ***5-character n-grams:*** *'sub'* (low), *eucri* and *eucru* (part of *Teucri*), *fatus* (spoke), *fatur* (speaks), *auras* (breezes), *eneas* (Aeneas)
- ***6-character n-grams:*** teucru (Trojan), aeneas (Aeneas), *'fatur'* (speaks), *'fatus'* (spoke), *'fatis'* (fates), *ipoten* and *mnipot* (from *omnipotens*, presumably attributed to Jupiter), *clamor* (shout)
- ***1 word n-grams:*** *urbem* (city), *aeneas* (Aeneas), *teucrum* (Trojan), *ingentem* (huge), *omnipotens* (almighty), *aether* (ether/sky), *pius* (pious), *iamque* (and now), *socius* (ally), *clamore* (shout), *finis* (end), *fatis* (fates), *ignem* (fire), *auris* (from *auris*, ear or *aurum*, gold), *caelum* (sky), *genitor* (father), *hostis* (enemy), *terram* (land), *bellum* (war), *dux* (leader), *uisus* (vision).

A similar phenomenon occurs with the terms obtained from the grouping of the Augustan love poets with Lucan, where words referring to the political causes of the Civil War predominate, emphasizing the crimes committed by the different factions and the physical consequences on the bodies of the Roman soldiers and citizens [30]:

4.4. Data from the Corpora of Catullus, Tibullus, Propertius, and Lucan:

- ***5-character n-grams:*** *aesar* (from Caesar), *aussa* (cause), *scera* (part of *viscera*, viscera), *pulos* (from *populos*, peoples), *elero* (part of *scelero*, referring to crimes), *coeli* (of *'coel'*, from heaven), *libye* (Libya), *gladi* (part of *gladium*, sword), *adhu* and *adhuc* (from until now)
- ***6-character n-grams:*** *'osque'* (composed most likely by the accusative plural ending of the second declension combined with the enclitic -que), *pulos* and *opulos* (peoples), *libye* (Libya), *scera* and *iscera* (from *viscera*, entrails), *bellor* and *elloru* (as part of *bellorum*, of the wars), *elerum*, *'lerum'* (from *scelerum*, crime), *caussa* and *'causs'* (cause), *'gladi'* (as part of *gladium* on its different forms, the sword), *'coeli'* (sky).
- ***1 word n-grams:*** *pectora* (chests), *populos* (peoples), *scelerum* (crimes), *bellorum* (wars), *senatus* (senate), *ciuilia* (civil), *nocentes* (from *nocens*, guilty or harmful), *caussa* (cause), *fatis* (fates), *coelo* (sky), *mundus* (world), *caeli* (sky), *diui* (gods), *libye* (Libya).

In previous work, we found that the Silhouette method consistently recommended two clusters for 2-character n-grams and three clusters for 1-word n-grams, regardless of the technique used. Scatter plots aligned with the stylistic distribution reported by Forstall et al. [12], who used SVM to analyze Catullus' influence on Paul the Deacon's poetry[5].

However, in this instance, whether due to the new preprocessing corrections or the novel comparison methods employed in separating Vergil and Lucan, the clustering tasks were not always accurate. While some correct distributions of the authors can be detected in the scatter plot space, the algorithm's non-human interpretation results in an unclear cluster classification, grouping Tibullus, Propertius, and Vergil against Catullus (Figure 5).

[5]For a colored version of the scatter plot, see Coffe et al. [31].

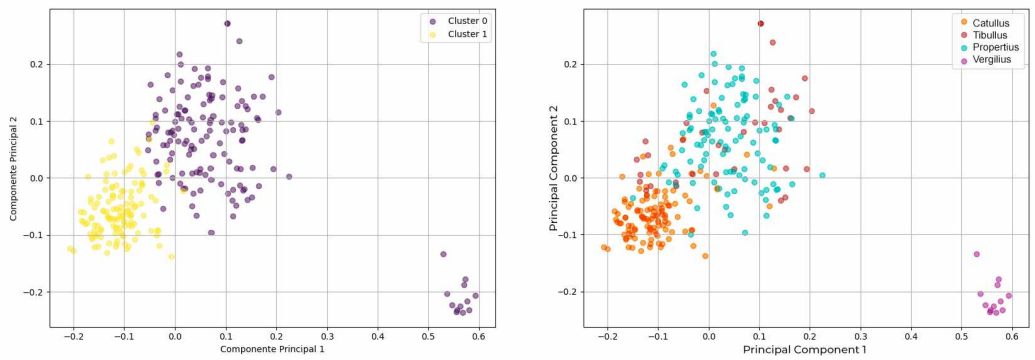

Figure 5: Scatter plot of clustering by K Means using a TF IDF matrix of 1-word n-grams indicating clusters (left) and authors (right) with different colors.

5. Conclusions and Learned Lessons

This article highlights the need to reevaluate methodologies and resources. Positive clustering results were obtained, especially with raw frequencies and n-grams, with Silhouette Scores above 0.8. Preprocessing steps and CLTK modules for tokenization and normalization significantly impacted the results, emphasizing the importance of tailored tools for ancient texts.

Relative frequency and two datasets reduced noise and bias from the epic authors, aiming to balance the text sets. However, even with more balanced data, raw frequencies provided better clustering results than relative frequency and TF-IDF matrices. As in the previous study, uneven feature importance and variable performance across n-gram levels and matrices suggest further refinement is needed for consistent results.

N-grams from Vergil's and Lucan's works show a dominance of political, historical, and war-related terms. This suggests that, despite efforts to balance the datasets, the lexical characteristics of epic poetry still influence classification. The identified n-grams reflect the epic and mythical focus of these authors, contrasting with the love and personal themes of Catullus, Tibullus, and Propertius. The variability in document length—regular in *Pharsalia* and *Aeneid*, but variable in the Augustan love poets—affects results. Additionally, the internal variability among the Augustan love poets' corpora also affects classification. We could experiment by partitioning Catullu's work into polymetric poems, *carmina maiora*, and epigrams or elegiac couplets, and run separate analyses, or intervene in the *corpus catullianum* by removing non-amorous themed poems. However, this is complex, as thematic boundaries are not clear-cut.

This exploratory analysis requires further refinement of other techniques such as variable ranges of character and word n-grams (only fixed ranges were used in this study), other similarity measures such as Jaccard, Cosine, or Soft Cosine, or clustering methods like Gaussian Mixture Models, DBSCAN, or hierarchical clustering. Future research could apply normalization

techniques such as L1 or Z-scaler, and phenomena like collocations and co-occurrences, which were not applied in this study. A close reading of clusters based on relative frequency also offers promise.

As for the representation of the documents, there is a need to explore techniques with embeddings like those developed by Burns et al., Bamman et al., and Johnson et al. [29, 32, 33, 34].

It should also be noted that the terms obtained by the Decision Tree technique are words with classification power for that dataset, not necessarily the most typical of one type of poetry or another, as there may be important words for both genres penalized by the metrics of Importance, Information Gain, or TF-IDF. The unequal size of the poems also contributed to the clarity of classification in raw counts, indirectly transferring poem length as a classification criterion. Similarly, in decision trees, the feature split points reflected the same pattern, with epic poem features having much higher frequencies, clearly impacting the results.

Finally, it is important to briefly consider the implications of applying computational and Distant Reading techniques alongside hypotheses or educated guesses from Close Reading. Frequency counting, for instance, is used here to model documents, but humans do not speak to be counted. Otherwise, Catullus would simply have repeated the name *Lesbia*, and his love would have been understood without the effort of creating poetry. Fortunately, language is far more abstract and complex, and computational methods are only beginning to reveal its intricacies. This issue has resurfaced with criticisms, such as those by Noam Chomsky, against generative models [35]. It is true that the human mind performs language tasks in a highly elegant manner and acquires a language exposed to a much smaller number of data than those handled by Large Language Models (LLMs). LLMs are tools developed for other tasks that did not originally seek to emulate the human mind [36, 37]. But it is also true that one must yield to the evidence of the successful results obtained with the use of these techniques and their undeniable capacity to facilitate all kinds of tasks. It's essential to acknowledge both the limits and strengths of computational tools, recognizing that Distant Reading offers a different scale of analysis—rooted not just in methodology but in changes in how information is produced, accessed, and analyzed in the digital age [38]. Despite criticisms [39], Digital Humanities methodologies hold great promise for studying language-rich subjects that balance aesthetic and rhythmic elements like refrains, alliterations, and anaphoras, presenting a unique challenge for modern analytical techniques.

6. Acknowledgments

I sincerely thank Dr. Kyle P. Johnson, Director of AI at Morgan, Lewis and Bockius LLP, and Dr. Patrick J. Burns from the Institute for the Study of the Ancient World, NYU, for their kind and insightful responses to my inquiries on tokenization and the use of CLTK and LatinCy libraries. I also extend my gratitude to Professor Benjamin Nagy from the Institute of the Polish Language, Polish Academy of Sciences (IJP PAN), Krakow, for his expert advice on correcting verse counts based on authorized editions, which greatly enhanced the accuracy of this text analysis.

References

[1] C. J. Nusch, Las Edades del Amor: una propuesta para el proyecto Aetates Amoris destinado a la poesía amorosa, Tesis, Universidad Nacional de Educación a Distancia, España, 2021. URL: http://sedici.unlp.edu.ar/handle/10915/125629. doi:10.35537/10915/125629.

[2] C. S. Lewis, La alegoría del amor: un estudio sobre tradición medieval, 2015 ed., Encuentro, Madrid, 1936.

[3] S. Ramsay, Reading Machines: Toward and Algorithmic Criticism, Urbana, 2011.

[4] F. Moretti, Distant Reading, Verso, 2013. Google-Books-ID: Sh9uNQEACAAJ.

[5] C. J. Nusch, G. del Rio Riande, L. C. C. Cagnina, M. L. Errecalde, L. Antonelli, Initial Explorations for Document Clustering Tasks in Latin Elegiac Poets, in: Decisioning, Pereira, Colombia, 2024.

[6] Giovanni Bracco, Silvio Migliori, Giorgio Mencuccini, Daniela Alderuccio, Giovanni Ponti, Data mining tools and GRID infrastructure for Assyriology text analysis (an Old-Babylonian situation studied through text analysis and data mining tools), in: RAI-Rencontre Assyriologique Internationale- Private and State in the Ancient Near East, Belgium, 2013, p. [No pages].

[7] A. Martins, C. Grácio, C. Teixeira, I. Pimenta Rodrigues, J. L. G. Zapata, L. Ferreira, Historia Augusta authorship: an approach based on Measurements of Complex Networks, Applied Network Science 6 (2021) 1–23. URL: https://appliednetsci.springeropen.com/articles/10.1007/s41109-021-00390-7. doi:10.1007/s41109-021-00390-7, number: 1 Publisher: SpringerOpen.

[8] G. Cantaluppi, M. Passarotti, Clustering the Corpus of Seneca: A Lexical-Based Approach, in: M. Carpita, E. Brentari, E. M. Qannari (Eds.), Advances in Latent Variables: Methods, Models and Applications, Springer International Publishing, Cham, 2015, pp. 13–25. URL: https://doi.org/10.1007/10104_2014_6. doi:10.1007/10104_2014_6.

[9] B. Nagy, Rhyme in classical Latin poetry: Stylistic or stochastic?, Digital Scholarship in the Humanities 37 (2022) 1097–1118. URL: https://doi.org/10.1093/llc/fqab105. doi:10.1093/llc/fqab105.

[10] B. Nagy, Some stylometric remarks on Ovid's Heroides and the Epistula Sapphus, Digital Scholarship in the Humanities 38 (2023) 1183–1199. URL: https://doi.org/10.1093/llc/fqac098. doi:10.1093/llc/fqac098.

[11] C. W. Forstall, S. L. Jacobson, W. J. Scheirer, Evidence of intertextuality: investigating Paul the Deacon's Angustae Vitae, Literary and Linguistic Computing 26 (2011) 285–296. URL: https://doi.org/10.1093/llc/fqr029. doi:10.1093/llc/fqr029.

[12] C. W. Forstall, W. Scheirer, A Statistical Stylistic Study of Latin Elegiac Couplets, in: 2010 Chicago Colloquium on Digital Humanities and Computer Science, 2010, p. [No pages]. URL: https://www.semanticscholar.org/paper/A-Statistical-Stylistic-Study-of-Latin-Elegiac-Forstall-Scheirer/e3caac9ec4ee16baac70ed94808dca57dff48a2d.

[13] S. Lloyd, Least squares quantization in PCM, IEEE Transactions on Information Theory 28 (1982) 129–137. URL: https://ieeexplore.ieee.org/document/1056489. doi:10.1109/TIT.1982.1056489, conference Name: IEEE Transactions on Information Theory.

[14] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of

cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257. doi:10.1016/0377-0427(87)90125-7.
[15] J. R. Quinlan, Induction of decision trees, Machine Learning 1 (1986) 81–106. URL: https://doi.org/10.1007/BF00116251. doi:10.1007/BF00116251.
[16] C. E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–423. URL: https://ieeexplore.ieee.org/document/6773024. doi:10.1002/j.1538-7305.1948.tb01338.x, conference Name: The Bell System Technical Journal.
[17] E. T. Merrill, Catullus; edited by Elmer Truesdell Merrill, Boston Ginn, 1893. URL: http://archive.org/details/catulluseditedby00catuuoft.
[18] L. Müller, Sex. Propertii Elegiae, Teubner, Leipzig, 1898. Google-Books-ID: _lKc8LSDfZYC.
[19] J. P. Postgate, Tibulli aliorumque carminum libri tres, Scriptorum classicorum bibliotheca Oxoniensis, Oxford, 1915. Google-Books-ID: qal45dBDIbEC.
[20] J. B. Greenough, The Bucolics, AEneid, and Georgics of Virgil, Ginn, Boston, 1900. OCLC: 51863711.
[21] C. H. Weise, Pharsaliae Libri X. M. Annaeus Lucanus., G. Bassus, Leipzig, 1935.
[22] [No author], Perseus Digital Library Homepage, [No date]. URL: https://www.perseus.tufts.edu/hopper/.
[23] L. M. Cerrato, R. F. Chavez, Perseus Classics Collection: An Overview, [No date]. URL: https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0053.
[24] K. Spärck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation 28 (1972) 11–21. URL: https://doi.org/10.1108/eb026526. doi:10.1108/eb026526, publisher: MCB UP Ltd.
[25] A. Berra, Ancient Greek and Latin Stopwords, 2024. URL: https://github.com/aurelberra/stopwords, original-date: 2017-10-07T21:49:36Z.
[26] P. J. Burns, Constructing Stoplists for Historical Languages, Digital Classics Online (2018) 4–20. URL: https://journals.ub.uni-heidelberg.de/index.php/dco/article/view/52124. doi:10.11588/dco.2018.2.52124.
[27] [No author], Stopwords ISO, [No date]. URL: https://github.com/stopwords-iso/stopwords-iso/blob/master/README.md.
[28] C. J. Nusch, Una breve exploración de la terminología amorosa en los corpora catullianum, tibullianum y propertianum con métodos y herramientas computacionales: etiquetado gramatical, lemas, bigramas y co-apariciones, Revista de Humanidades Digitales 9 (2024) 1–40. URL: https://revistas.uned.es/index.php/RHD/article/view/38680. doi:10.5944/rhd.vol.9.2024.38680.
[29] K. P. Johnson, P. J. Burns, J. Stewart, T. Cook, C. Besnier, W. J. B. Mattingly, The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages, in: H. Ji, J. C. Park, R. Xia (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 20–29. URL: https://aclanthology.org/2021.acl-demo.3. doi:10.18653/v1/2021.acl-demo.3.
[30] M. M. Vizzotti, De la tragedia de Séneca a la épica de Lucano: estrategias de representación

de los paradigmas filosóficos y literarios, Tesis, Universidad Nacional de La Plata, 2014. URL: http://sedici.unlp.edu.ar/handle/10915/34410. doi:10.35537/10915/34410.

[31] N. Coffee, J. Gawley, C. Forstall, W. Scheirer, D. Scheirer, J. Corso, B. Parks, Modelling the Interpretation of Literary Allusion with Machine Learning Techniques Journal of Digital Humanities, Journal of Digital Humanities (2013) 478–479. URL: https://journalofdigitalhumanities.org/3-1/modelling-the-interpretation-of-literary-allusion-with-machine-learning-techniques/.

[32] P. J. Burns, LatinCy: Synthetic Trained Pipelines for Latin NLP, 2023. URL: http://arxiv.org/abs/2305.04365. doi:10.48550/arXiv.2305.04365, arXiv:2305.04365 [cs].

[33] D. Bamman, P. J. Burns, Latin BERT: A Contextual Language Model for Classical Philology, 2020. URL: http://arxiv.org/abs/2009.10053. doi:10.48550/arXiv.2009.10053, arXiv:2009.10053 [cs].

[34] P. J. Burns, Building a Text Analysis Pipeline for Classical Languages, in: Building a Text Analysis Pipeline for Classical Languages, De Gruyter Saur, 2019, pp. 159–176. URL: https://www.degruyter.com/document/doi/10.1515/9783110599572-010/html. doi:10.1515/9783110599572-010.

[35] N. Chomsky, I. Roberts, J. Watumull, Noam Chomsky: The False Promise of ChatGPT, The New York Times (2023). URL: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

[36] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. URL: http://arxiv.org/abs/1810.04805. doi:10.48550/arXiv.1810.04805, arXiv:1810.04805 [cs].

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, I. Polosukhin, Attention is All you Need, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017, p. [No pages]. URL: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[38] Ricardo Pimenta, De Narciso ao mundo-imagem: por uma urgência de uma perspectiva crítica sobre a cena informacional contemporânea, in: Ciência da Informação : sociedade, crítica e inovação, Rio de Janeiro, 2022, p. [No pages].

[39] T. Brennan, The Digital-Humanities Bust, 2017. URL: https://www.chronicle.com/article/the-digital-humanities-bust/, section: The Review.

Appendix: Extra Data from the Corpora of Catullus, Tibullus, Propertius, Vergilius and Lucanus

Table 9
Most important features at the level of character 2 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method and Stopwords filtering (Corpora of Catullus, Tibullus, Propertius, and Vergilius).

Raw Frequency char (2, 2) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
un	1	nl	0.1254	dg	0.5747
		uq	0.1226	aë	0.5463
		dg	0.1208	oï	0.5193
		gg	0.1205	ën	0.4929
		ms	0.1205	ïa	0.4929
		gm	0.1205	ez	0.4664
		dh	0.1181	aï	0.4664
		bt	0.1174	dh	0.4516
		yd	0.1129	ï	0.439
		dl	0.1128	eï	0.439
		nh	0.1128	ër	0.439
		ze	0.1128	mf	0.4105
		bn	0.1072	x	0.376
		my	0.1068	oö	0.3758
		ln	0.105	ë	0.3758
		rh	0.105	ïc	0.3758
		aë	0.1049	ön	0.3758
		df	0.1036	ïu	0.3758
		yt	0.0999	oë	0.3758
		yc	0.0999	gg	0.3724

Table 10
Most important features at the level of character 3 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Raw Frequency char (3, 3) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
aba	1	ipo	0.1601	ffa	0.6056
		om	0.1601	dfa	0.6056
		ffa	0.1571	adg	0.5747
		dsu	0.1536	dhu	0.5747
		oln	0.1536	dgn	0.5747
		gii	0.1481	xsc	0.5747
		ols	0.1444	ciq	0.5747
		teb	0.1433	ybr	0.5747
		tto	0.1433	axu	0.5747
		uom	0.139	hyb	0.5747
		toq	0.139	ols	0.5521
		xce	0.139	yde	0.5463
		dfa	0.138	aë	0.5463
		rex	0.1365	b p	0.5463
		bii	0.1352	amd	0.5463
		teu	0.1352	moq	0.5463
		nip	0.1352	mdu	0.5463
		roq	0.1316	ipo	0.5458
		euc	0.1316	om	0.5458
		bum	0.1316	giq	0.5193

Table 11
Most important features at the level of character 4 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Raw Frequency char (4, 4) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
pri	1	temq	0.1813	iuom	0.6931
		iuom	0.1813	ucru	0.6931
		ucru	0.1813	temq	0.6931
		boru	0.1813	ubib	0.6931
		ubib	0.1813	boru	0.6931
		sumq	0.1683	nimb	0.6413
		mman	0.1683	m ef	0.6413
		lumq	0.1683	effa	0.6413
		mnip	0.1683	ffat	0.6413
		ipot	0.1683	ast	0.6413
		nipo	0.1683	cios	0.6413
		bibu	0.1683	lumq	0.6061
		ea	0.1601	ipot	0.6061
		iisq	0.1601	mman	0.6061
		cri	0.1601	mnip	0.6061
		dani	0.1601	nipo	0.6061
		teuc	0.1601	bibu	0.6061
		scun	0.1601	sumq	0.6061
		ttol	0.1601	adfa	0.6056
		eucr	0.1601	anid	0.6056

Table 12
Most important features at the level of character 5 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Raw Frequency char (5, 5) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
sub	1	ntemq	0.1813	eucru	0.6931
		eucri	0.1813	borum	0.6931
		eucru	0.1813	ntemq	0.6931
		fatus	0.1813	fatur	0.6931
		imman	0.1813	iuom	0.6931
		fatur	0.1813	imman	0.6931
		borum	0.1813	ucrum	0.6931
		iuom	0.1813	eucri	0.6931
		temqu	0.1813	e teu	0.6931
		e teu	0.1813	temqu	0.6931
		ucrum	0.1813	fatus	0.6931
		mnipo	0.1683	cios	0.6413
		auras	0.1683	clamo	0.6413
		sumqu	0.1683	anch	0.6413
		fatis	0.1683	effa	0.6413
		nipot	0.1683	lamor	0.6413
		nt ac	0.1683	ocios	0.6413
		omnip	0.1683	effat	0.6413
		eneas	0.1683	undam	0.6413
		ipote	0.1683	m eff	0.6413

Table 13
Most important features at the level of character 6 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Raw Frequency char (6, 6) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
tisque	1	teucru	0.1813	ngente	0.6931
		aeneas	0.1813	teucri	0.6931
		fatur	0.1813	aeneas	0.6931
		ntemqu	0.1813	ucrum	0.6931
		e teuc	0.1813	e teuc	0.6931
		eucrum	0.1813	teucru	0.6931
		ngente	0.1813	borum	0.6931
		imman	0.1813	imman	0.6931
		ucrum	0.1813	fatus	0.6931
		borum	0.1813	ntemqu	0.6931
		fatus	0.1813	eucrum	0.6931
		teucri	0.1813	fatur	0.6931
		temque	0.1813	temque	0.6931
		fatis	0.1683	m regi	0.6413
		auras	0.1683	a fatu	0.6413
		eneas	0.1683	pius a	0.6413
		auras	0.1683	a teuc	0.6413
		ipoten	0.1683	uisu	0.6413
		omnip	0.1683	clamor	0.6413
		mnipot	0.1683	e ora	0.6413

Table 14
Most important features at the level of word 1 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Vergilius).

Raw Frequency word (1, 1) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
urbem	1	aeneas	0.1813	teucrum	0.6931
		ingentem	0.1813	ingens	0.6931
		ingens	0.1813	ingentem	0.6931
		teucrum	0.1813	aeneas	0.6931
		fatis	0.1683	pius	0.6413
		omnipotens	0.1683	aethere	0.6413
		late	0.1683	ast	0.6413
		ignem	0.1683	fatur	0.6413
		auras	0.1683	diuom	0.6413
		iamque	0.1601	socios	0.6413
		ea	0.1601	clamore	0.6413
		genitor	0.1601	teucros	0.6413
		terram	0.1601	uisu	0.6413
		talibus	0.1601	ignem	0.6061
		equidem	0.1601	omnipotens	0.6061
		diuom	0.1571	fatis	0.6061
		pius	0.1571	late	0.6061
		uisu	0.1571	auras	0.6061
		teucros	0.1571	teucri	0.6056
		ast	0.1571	regem	0.6056

Table 15
Most important features at the level of char 2 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Raw Frequency char (2, 2) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
g	1	ye	0.1464	ye	0.5942
		gm	0.1234	dh	0.4673
		dh	0.1151	mt	0.4094
		ya	0.1048	bf	0.3991
		by	0.1025	gm	0.3975
		xq	0.097	sf	0.3797
		ze	0.0938	dt	0.3708
		dt	0.0914	fc	0.3514
		oh	0.0909	fs	0.3514
		rh	0.0894	pc	0.3514
		oa	0.0879	nb	0.3514
		sn	0.087	dg	0.3514
		dq	0.0864	cp	0.3514
		yp	0.0864	sn	0.3306
		df	0.0858	bm	0.3287
		gy	0.0858	y	0.323
		ee	0.0858	mf	0.323
		yc	0.085	xq	0.3126
		sy	0.085	ms	0.2979
		lm	0.0836	ze	0.2883

Table 16
Most important features at the level of char 3 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Raw Frequency char (3, 3) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
te	1	oer	0.1589	bye	0.6931
		bye	0.1589	oer	0.6931
		rct	0.1464	dhu	0.636
		obo	0.1388	ye	0.636
		ax	0.1388	rct	0.5942
		ye	0.1345	b p	0.5627
		dhu	0.1345	giq	0.5627
		nfu	0.1328	gme	0.5341
		agm	0.1277	emt	0.5309
		teb	0.1234	bmo	0.5309
		rut	0.1234	eer	0.5309
		gme	0.1224	xir	0.5309
		toq	0.1195	ax	0.5275
		xce	0.1195	obo	0.5275
		x n	0.1195	nny	0.5
		lsu	0.1195	al	0.5
		pei	0.116	mto	0.5
		axe	0.116	bif	0.5
		aux	0.116	efo	0.5
		saq	0.116	gad	0.5

Table 17
Most important features at the level of char 4 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Raw Frequency char (4, 4) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
ra n	1	ibye	0.1589	auss	0.6931
		auss	0.1589	coel	0.6931
		coel	0.1589	ibye	0.6931
		glad	0.1589	glad	0.6931
		sena	0.1464	oelo	0.636
		s rh	0.1464	bye	0.636
		iuil	0.1464	dhuc	0.636
		leru	0.1464	rtib	0.636
		efas	0.1464	adhu	0.636
		moto	0.1464	auce	0.636
		susq	0.1464	cesp	0.5968
		fauc	0.1464	moes	0.5968
		tebr	0.1464	xcus	0.5968
		rcto	0.1464	suor	0.5968
		lumq	0.1464	tors	0.5968
		arct	0.1464	otue	0.5968
		mpul	0.1388	adau	0.5968
		robo	0.1388	gulo	0.5968
		uile	0.1388	nfan	0.5968
		obor	0.1388	mpag	0.5968

Table 18
Most important features at the level of char 5 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Raw Frequency char (5, 5) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
aesar	1	aussa	0.1589	causs	0.6931
		lerum	0.1589	scera	0.6931
		causs	0.1589	eleru	0.6931
		scera	0.1589	pulos	0.6931
		coeli	0.1589	coeli	0.6931
		libye	0.1589	coel	0.6931
		pulos	0.1589	lerum	0.6931
		gladi	0.1589	libye	0.6931
		eleru	0.1589	aussa	0.6931
		coel	0.1589	gladi	0.6931
		glad	0.1589	glad	0.6931
		ellor	0.1589	ellor	0.6931
		arcto	0.1464	oeli	0.636
		peri	0.1464	oelo	0.636
		fatis	0.1464	i dam	0.636
		ciuil	0.1464	oties	0.636
		tent	0.1464	ic fa	0.636
		nefas	0.1464	obore	0.636
		fauc	0.1464	adhuc	0.636
		susqu	0.1464	iscri	0.636

Table 19
Most important features at the level of char 6 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Raw Frequency char (6, 6) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
osque	1	pulos	0.1589	elerum	0.6931
		scera	0.1589	ssere	0.6931
		lerum	0.1589	iscera	0.6931
		opulos	0.1589	celeru	0.6931
		bellor	0.1589	s phar	0.6931
		causs	0.1589	gladi	0.6931
		ssere	0.1589	m popu	0.6931
		s phar	0.1589	pulos	0.6931
		elerum	0.1589	lerum	0.6931
		exit	0.1589	elloru	0.6931
		lia be	0.1589	bellor	0.6931
		caussa	0.1589	caussa	0.6931
		iscera	0.1589	opulos	0.6931
		libye	0.1589	libye	0.6931
		gladi	0.1589	coeli	0.6931
		m popu	0.1589	scera	0.6931
		coeli	0.1589	causs	0.6931
		celeru	0.1589	lia be	0.6931
		elloru	0.1589	exit	0.6931
		us for	0.1464	unctas	0.636

Table 20
Most important features at the level of word 1 n-grams according to Importance, Information Gain, Information Gain Ratio using the raw frequency matrix method (Corpora of Catullus, Tibullus, Propertius, and Lucanus).

Raw Frequency word (1, 1) n-grams					
Feature	**Importance**	**Feature**	**IG**	**Feature**	**IGR**
pectora	1	populos	0.1589	populis	0.6931
		scelerum	0.1589	scelerum	0.6931
		populis	0.1589	bellorum	0.6931
		mundo	0.1589	uiscera	0.6931
		bellorum	0.1589	exit	0.6931
		exit	0.1589	populos	0.6931
		uiscera	0.1589	mundo	0.6931
		senatus	0.1464	caussa	0.636
		ciuilia	0.1464	nocentes	0.636
		nefas	0.1464	ciuilibus	0.636
		fatis	0.1464	coelo	0.636
		bellum	0.1388	libye	0.636
		milite	0.1388	robore	0.636
		ducis	0.1388	superi	0.636
		ciuilibus	0.1345	adhuc	0.636
		fauces	0.1345	ciuile	0.636
		robore	0.1345	fauces	0.636
		adhuc	0.1345	coeli	0.636
		caussa	0.1345	malorum	0.636
		libye	0.1345	potuere	0.5968