Stylometric Classification of Medieval Greek Text Using Machine Learning Techniques

Nikolina Milioni
Research and Development
Uppsala University
Nikolina.Milioni.6178@student.uu.se

Abstract

Natural Language Processing highly focuses on classification problems. Regarding genre categorization, this can be a challenging task especially for low-resource languages and historical datasets. In this study, a Medieval Greek text classification task is performed, for which no tagged corpora yet exist. Thence, stylometric features and different machine learning algorithms are exploited. Both supervised and unsupervised learning are tested. The cross-validation predictive model performed with a Random classifier presents the highest accuracy (97.6%). The Random Forest model outperforms the Support Vector Machine classifier. The K-Means unsupervised model fails to clearly cluster the data in five different categories, which are epigrams, hymns, poems, historical and religious texts.

1 Introduction

High resource languages, such as English, are the main focus research area regarding Natural Language Processing (NLP). The provision of labeled and annotated corpora for these languages, along with the technological advent of Machine Learning (ML) and Neural Networks (NN), constantly benefits the linguistic research and addresses uprising challenges. However, this is not the case when it comes to low-resource modern languages or less-resourced languages, which the historical languages are considered to be (Piotrowski, 2012). When it comes to Greek, the number of tools, resources and research work on this language gradually increases (Papantoniou and Tzitzikas, 2020), however research on the historical versions of Greek language is either under-developed or not developed at all. Available resources mainly concern the Ancient Greek language (800-300 BC period) but not other variations, such as Koine Greek, Medieval Greek, Katharevousa (conservative form), Demotic or other dialects of these versions.

In regard to text classification in low resource languages, this process can be benefited by exploiting machine learning techniques along with stylometric features. In computational linguistics, stylometry is a quantitative analysis of linguistic features. These features can vary depending on the nature of the problem as there is no strict agreement on which exactly style markers should be applied on research (Ramyaa et al., 2004). These markers can vary between Lexical, Syntactical and Morphological features, such as mean sentence length, vocabulary richness, hapax legomena etc. This technique is highly implemented on authorship recognition, genre classification, gender classification, propaganda detection and writing style recognition.

The purpose of this study is to evaluate Machine Learning classification techniques on Medieval Greek documents, using stylometric features. Classification is the process of sorting collections into specific groups (Pustejovsky and Stubbs, 2012) and what makes automatic classification major is that most of the data and information remain unstructured. The performance of this task is supported by both Supervised and Unsupervised learning.

Medieval Greek is the written and spoken language between 700 AD – 1700 AD. It is quite similar to the Ancient Greek, as the polytonic system (orthography/diacritics) and most morphological and syntactic features remain. No tagged corpora exist so far for Medieval Greek, thus one goal of this project is to determine what information we can gain from categorized data without any further labeling. The implemented set of stylometric features is the one that Gianitsos et al. (2019) used as well. The reason for this option is to detect whether these features are equally successful both for Ancient and Medieval Greek, indicating great similarities between those two versions. As they state in their research, these features were selected based on three factors: no syntactic parsing on demand for these features, features applicable to the selected corpora and multifunctionality.

2 Related Work

No NLP research has been conducted with regards to text classification of Medieval Greek texts, to my knowledge. This project is heavily inspired and based on the research of Gianitsos et al. (2019), who developed a stylometric feature set for classical Ancient Greek text categorization which does not require syntactic parsing or tagging. The documents are categorized as prose or verse and the poems (verse) are classified as epic or drama. They implement a supervised machine learning method with the Random Forest classifier (RF). The results prove that Ancient Greek Literature can be classified by genre with high accuracy by implementing this heuristic approach.

Nevertheless, stylometric features and machine learning techniques have been carried out in various research on text classification topics, historical or not. Elahi and Muneer (2018) implemented stylometric features on an unsupervised learning task. Specifically, using K-Means Clustering, they attempted to identify and classify different writing styles within the same document. They opted for lexical features, vocabulary richness features and readability scores, 21 features in total. They merged two documents, a story and a research paper, and ran an experiment on those. The system successfully detected the two different writing styles of the merged document.

Genre detection has also been investigated in Classical Arabic texts using stylometric analysis (Al-Yahya, 2018). In this case, unsupervised clustering and supervised classification were implemented. The outcome of the study indicates that genre style signals in Classical Arabic texts can support automated genre detection.

Hettinger et al. (2015) focused on genre classification of German literature. They used machine learning methods to explore the performance of classification algorithms and also exploited stylometric features to enhance the genre classification process. They also added features that had not been used in the past, leading to an improved performance of German literature text classification.

Kumar and Minz (2014) classified poems using ML methods. They tested K-nearest neighbor (KNN), Naive Bayesian (NB) and Support Vector Machine (SVM) with reduced features. For feature selection they opted for the Information Gain Ratio and the SVM model presented the best results.

Tizhoosh et al. (2008) investigated poem classification. They implemented five different approaches and for each one they exploited different poetic feature sets. In general, the results were significantly promising and some of the models superior against other. This study proves that when opting for proper features in relation to the data, the classifiers can present a great performance.

Ramyaa et al. (2004) investigated writing style detection using machine learning methods and stylometry. 21 style indicators were implemented on Victorian era texts from five authors, and two machine learning algorithms. Decision Trees achieved 82.4% accuracy and Neural Networks 88.2%. They suggested that by mixing those two methods, it could possibly lead to almost perfect classifications.

Simić and Jurada (2020) performed a task in authorship attribution based on stylometric features. This can be considered as a classification task. The implemented dataset consisted of 24 articles from five different topics. The articles were labelled either as high-quality press or yellow press. 23 features were selected, either part of speech based or traditional style markers, and four classifiers: AdaBoost, Random Forest, Multi-layer Perceptron and Extremely Randomized Trees. After detecting the ten most important features, they retrained and merged the above models. The combined classifier achieved an accuracy of 90%.

Lagutina et al. (2019) conducted a literature review on extraction and application of stylometric features around various topics. Regarding text classification by genre and sentiment (which is the aim of the current project), they analysed three different studies in the field. Balint et al. (2016), by using Discriminant Function Analysis, selected the eight most dominant features of an initial list of rhythm style markers (such as phonetic and metrical) and achieved an accuracy of 81% in a speech, essay, and newspaper article classification task. Amancio (2015) attempted to classify prose text as informative or imaginative. He implemented an adjacency network of words, stopwords, and bigrams, which presented a higher accuracy than simpler adjacency networks. Anchiêta et al. (2015) tested a 2000 smartphone review classification using word statistics, syntactic features and content specific features. The SVM classifier yielded the best results: an 82.75% accuracy with these features.

3 Experimental Setup

3.1 Dataset

The original dataset consists of 919 Greek medieval texts from the 4^{th} until the 16^{th} century A.D and includes 3.4 million words. It can be retrieved by the CLARIN:EL [1], the National Infrastructure for Language Resources and Technologies in Greece, which supports researchers of Language Studies, Digital Humanities and Social Sciences, Language Technology and relevant fields. The open-source documents are divided in 7 categories: religious (565), poetical (73), literary (5), political (4), historical (74), hymns (36) and epigrams (164). Due to size variation, 5 text categories are implemented in this project: religious, poetical, historical (documents with historical events), hymns and epigrams. Moreover, not all documents within the latter categories are used to perform the classification task because a number of those contain text without the Greek diacritics (polytonic orthography). For reproducibility reasons, specific documents are included in the train and test set, which are listed in Appendix A, along with the number of the selected documents.

3.2 Feature Selection

The selected style markers are the same that Gianitsos et al. (2019) used in their research, apart from one, which was the variance of sentence length. I was doubtful about how exactly the variance was calculated, so I decided to exclude this feature. They extracted those features based on similar studies in English language and some that only relate to Ancient Greek, as not all words have an explicit equivalent in other languages, especially when the structures of the languages differ significantly. The 22 stylometric features are included in Table 1. Most of these features are either function words, non-content words or inflected forms. Since a syntactic parser is yet to be developed for Medieval Greek, stylometric features can prove to be valuable for such tasks. Apart from the mean sentence length, the mean length of relative clauses, the interrogative sentences and clauses of purpose, the rest of the features are calculated and normalized per total number of tokens in the corresponding document. Appendix B presents the detailed feature list, as presented by Gianitsos et al. (2019).

	Stylometric Features
	Pronouns and non-content adjectives
1	ἄλλος (other)
2	αὐτός (self/him,her,it)
3	demonstrative pronouns
4	selected indefinite pronouns
5	personal pronouns
6	reflexive pronouns
	Conjunctions and particles
7	conjunctions
8	μέν (indeed)
9	particles
	Subordinate clauses
10	circumstantial markers
11	conditional markers
12	ἵνα (where/in order that)
13	ὅπως (how/in order that
14	sentences with relative pronouns
15	temporal and causal markers
16	ὥστε not preceded by ἤ (so/as to)
17	mean length of relative clauses
	Miscellaneous
18	interrogative sentences
19	superlatives
20	sentences with ὦ exclamations
21	ὡς (how/that/so that/since)
22	mean sentence length

Table 1: Style markers for Medieval Greek

3.3 Pre-processing

At the outset, the retrieved documents were in .doc format and had to be converted into .txt format. In this first Medieval Greek text classification, a random split is implemented for the supervised learning; 80% of the documents of each category is used for the training process and 20% for the evaluation (Appendix A). Concerning the epigram documents, those lacked a diacritic sign, the *iota* subscript, which is most times used as an inflectional suffix and located only under three long vowels: α (alpha), ω (omega) and η (eta). So, instead of ᾳ, ῃ, and ῳ, these were written as αι, ηι

[1] https://inventory.clarin.gr/corpus/890

and ωι where the *iota* subscript is moved next to the vowel. This detail is changed for a more accurate feature implementation.

3.4 Supervised Learning

For all classification models, python 3.7.4 and scikit-learn library are implemented. Grid Search hyperparameter tuning is applied on all supervised models. The detailed parameters are listed in Appendix E. Moreover, for every classifier (RF, SVM, K-Means) one baseline is built based on features from the TF-IDF technique, so that the stylometric feature models could be compared to.

The first supervised method tested in this multi class classification task is a Random Forest Classifier. In order to investigate the potential quality of the classifier and to study the feature importance, a cross validation process was necessary. Similarly to Gianitsos et al. (2019), a (repeated) Stratified 5-fold validation is implemented so that the classes with the most data are not over-represented due to equal weight assignment. The 5-fold cross validation is repeated 10 times in order to improve the estimated performance (50 models are trained in total) and for generalization purposes. For the overall feature importance extraction, the random forest built-in method is exploited (Gini coefficient). After cross validation, a Random Forest model is trained. Accuracy, weighted F1 scores and feature importance are also calculated.

For the next experiment, a Support Vector Machine (SVM) is tested. Since this is not a linear SVM, the built in method for feature importance extraction cannot be used, thus the Permutation Importance technique from scikit-learn is implemented.

3.5 Unsupervised Learning

The final model is a K-Means Clustering one. The purpose of this experiment is to get an idea on how well this algorithm performs, meaning how accurately the clusters are generated, and what information it might reveal about data on which no training is preceded. Principal Component Analysis (PCA) is used in two ways: on the one hand, to simply visualize the high dimensional data and on the other hand to reduce data dimensionality before model fitting, and investigate whether this improves the models' performance. Finally, the elbow method technique is implemented to predict the number of clusters that should be generated depending on the nature of our features.

4 Results

Table 2 presents the cross-validation results. Concerning one sample of cross-validation trial (fold 1-5), the mean accuracy is 97.1% and the standard deviation 1.5. The weighted F1 score is 97.3% and its standard deviation 1.4. After 10 trials, the (overall) mean accuracy and the standard deviation are 97.6% and 1 correspondingly and the (overall) weighted F1 scores present a 97.7% accuracy and 1 standard deviation. The five most important features (with Gini importance > 0.05) are listed in table 3.

	Accuracy (%)	**Weighted F1 (%)**
Fold 1	96.8	97.1
Fold 2	99.3	99.3
Fold 3	95.5	95.6
Fold 4	95.5	95.8
Fold 5	98.7	98.5
Mean	97.1	97.3
SD	1.5	1.4
Overall	97.6	97.7
SD	1	1

Table 2: Cross Validation evaluation with Random Forest

Feature	**Gini**
Indefinite pronouns	0.153
Reflexive pronouns	0.138
ἵνα	0.086
Interrogative sentences	0.085
αὐτός	0.082
ὦ exclamation	0.061
Inferential sentences	0.051

Table 3: Cross Validation best features

The accuracy of the trained Random Forest classifier reached an 87% accuracy. Poems are highly misclassified whereas epigrams, historical and religious texts are classified with high accuracy. The best features are presented in table 4 according to Gini importance. Both the predicted models and the Random Forest outperformed the TF-IDF baseline, which reached an accuracy of 81%.

The Random Forest classifier outperforms the SVM classifier, which presents a 71% accuracy. A standard scaler is also implemented to improve the performance of the SVM classifier. After tun-

Feature	**Gini**
Indefinite pronouns	0.162
ἵνα	0.11
αὐτός	0.102
Reflexive pronouns	0.085
Interrogative sentences	0.074
ὦ exclamation	0.066

Table 4: Random Forest Gini feature importance

ing the hyperparameters, the TF-IDF baseline has a better performance when it comes to accuracy (77%). Table 5 lists the best features for the SVM model according to scikit-learn's feature selection module.

Feature	**Importance Score**
Sentence mean length	0.143
Reflexive pronouns	0.050
ὦ exclamation	0.036
Circumstantial markers	0.026

Table 5: Permutation importance in SVM style markers

In the Unsupervised learning, the K-Means clustering without PCA implementation, does not present significant results according to the homogeneity score and evaluation metrics. The performance is low and the results remain insignificant even when PCA is applied before fitting. The baseline performs better, but still the data is highly misclassified (42% accuracy and 0.39 homogeneity score). Once again, PCA does not improve the quality of the classifier, as the accuracy is decreased to 19%, however the homogeneity score is slightly improved (0.40). The elbow method always suggests 3 or 4 number of clusters regardless the exploited features (figures in Appendix C).

5 Discussion and Future Work

5.1 ML Perspective

The Random Forest classifier performs better than the SVM classifier. After looking closely at the clustering output, we could argue that RF presents a better performance because it is less influenced by outliers (a.k.a. datapoints that lie far away from the rest observations). Also, the RF algorithm can discard features that are not useful during the task. After 10 trials of cross validation, the accuracy and weighted F1 scores are higher and standard deviation is lower than the corresponding results of the first trial only. Thus, the estimated performance is improved which also indicates that there is noise in the data and different k-fold splits induce different output.

Model	**Ac (%)**	**WF1 (%)**
RF Baseline (Tf-Idf)	81	77
RF (style)	87	85
SVM Baseline (Tf-Idf)	77	76
SVM (style)	71	64
KMeans Baseline(Tf-Idf)	43	42
KMeans Baseline + PCA	19	27
KMeans (style)	24	14
KMeans (style) + PCA	20	0.7

Table 6: Models' performance

Model	**Homogeneity**
KMeans Baseline(Tf-Idf)	0.39
KMeans Baseline + PCA	0.40
KMeans (style)	0.02
KMeans (style) + PCA	0.02

Table 7: Homogeneity scores

Concerning SVM, this classifier depends on the decision boundary to label the documents. By studying the clustering plot (Appendix D), we can suspect that the boundaries are likely to be quite unclear in many cases, even if we train the classifier. This proves that the dataset is complex and definitely not lineary separable. The SVM classifier presents higher performances in high dimensionalities, which clearly explains the fact that the SVM baseline, in which the TF-IDF features were exploited, performs better than the stylometric SVM model (the number of style markers is 22). Concerning the parameteres in the SVM, a Radial Kernel or Radial Basis Function (instead of Linear) was more appropriate in order to deal with overlapping data. The overlap is also indicated by the homogeneity score in the unsupervised learning. Homogeneity score is calculated based on Shannon's entropy and perfect homogeneity is equal to 1. This is the case when all data points that belong to a specific cluster share the same label.

All tested K-Means models present a very low homogeneity score. K-Means is a centroid clustering algorithm that exploits the Euclidean distance to find the distance between the centers. Before data fitting, this number (number of clusters) must be chosen, which is not demanding in this case since we know the number of groups. However, the proper number of clusters is tested with the Elbow Method, which calculates the Within Cluster Sum of Square (WCSS), a monotonously decreasing function. Ten values of k are used and k is picked at the point where WCSS becomes more stable after a fast decline. Appendix C depicts the elbow method with TF-IDF and stylometric features. The former has a less smooth curve than the latter. In both cases though, the method decides 4 as the optimal number of clusters. It could also be 3 in the stylometric model, however I believe that the leap between 3 and 4 is still important to be discarded. The elbow method fails to predict the correct number of clusters either because, given the provided features, this is not the correct algorithm for the problem or more data processing (like standardization) is on demand. Given that there is one very big category (religious texts), this affects the model's performance since K-Means enhances the weight in bigger classes. K-Means also affects overlapping in a negative way. If the clusters do not have a spherical shape, like in the output of the study (Appendix D), it means that K-Means cannot handle the structure of the data. In future research, the Silhouette method could be tested to predict the number of clusters.

Concerning the unsupervised learning, further research can be conducted by implementing more methods or different algorithms. By standardizing the data we can boost the model's performance, as by this method we can achieve a standard deviation of 1 and a zero mean. Standardization improved the SVM classifier after all. A density-based clustering algorithm could be ideal for this kind of data (like DBSCAN) since the data has a strange geometrical shape and is imbalanced, or a distribution-based clustering algorithm (like GMM) which can handle overlapped clusters and implements probabilistic approaches.

For the supervised learning, a probabilistic model (like Naive Bayes Classifier) could be tested to evaluate its performance on this classification task. Investigating the probability of events on this dataset, it might present accurate results. Finally, the TF-IDF and stylometric features could be combined to test whether this merged set of features would be more efficient for this problem.

5.2 Data Perspective

The implemented stylometric features combined with RF can classify prose text with high accuracy. All historical texts (16), almost all religious texts (92/94) and most epigrams (28/31) are correctly labelled. One religious text is classified as historical and one as a religious hymn. All hymns are misclassified. However, the dataset consists of a limited amount of texts and only 2 hymns are included in the test set and 10 in training set. Thus, in this case it is quite unclear whether the style markers are inappropriate to identify religious hymns or more data is required for this category. Also, most epigrams (28), which are usually considered as verse text, are mainly categorized correctly, apart from 3 mislabelled as poems. Poems are highly misclassified as historical texts (9), epigrams (1) and religious (1) and only one poem is assigned the correct label. According to the results, the recall is extremely high when it comes to prose texts, as only 2/108 are misclassified. Gianitsos et al. (2019) presented similar results in the binary classification they implemented (prose vs verse). In this study, the data is imbalanced when it comes to quantity since the amount of religious texts is about six times the amount of historical texts. However, since the historical texts (which are widely less) are all correctly labeled, this imbalance does not seem to affect the problem in this case. This imbalance though seems to affect the task when using the SVM classifier. All hymns, most historical texts (13/16) and most poems (9/14) are mislabeled as religious texts. Also, no poem is correctly labeled which, along with the RF results, indicates that probably new style markers should be appended in the feature list that fit better on this category.

RF model with TF-IDF features labels correctly religious texts and epigrams, but significantly fails in other categories. The 80% accuracy has only be achieved due to the great number of religious text labeling and most of the epigrams. Again, hymns and poems present the lowest precision and recall indicating, along with the previous results, that this data include a lot of noise, thus a greater quantity is probably needed in the training process. Similarly, SVM with TF-IDF features classifies epigrams and religious texts quite accurately, but more historical texts are labelled as religious and hymns are completed mislabelled. However, most poems are correctly classified this time (9/14).

In all case, no classifier managed to categorize correctly any of the hymns of the test set. All models apart from the SVM with style markers, label correctly 28/31 epigrams. Only the RF with style markers labels correctly all historical texts. The SVM with TF-IDF features corresponds better to poem category. RF labels correctly the religious texts (92/94) either with style markers or TF-IDF features.

Concerning feature importance, there is a significant overlap between cross-validation and the trained RF classifier. The common most important features are the indefinite pronouns, reflexive pronouns, ἵνα ,αὐτός, interrogative sentences and ὦ exclamation. These findings are quite different than what Gianitsos et al. (2019) reported. Only reflexive pronouns and αὐτός are in common. Also, more features (6 in total) present Gini importance over 0.05 on the Medieval texts than on the Ancient Greek dataset (5 in total), meaning that these features are more important and applicable on medieval texts than on ancient. The importance score on the SVM classifier also reveals only two features that are in common with the RF classifier (reflexive pronouns and ὦ exclamation).

So far, hymns and poems seem to be the most problematic categories. Thus, what is expected from the unsupervised classification is a not so clear clustering result. The elbow method suggests four clusters, which means that there is an invisible category, not recognised by any algorithm. By decreasing the feature dimensionality before model fitting, it affects negatively the models' performance, especially in the baseline method, meaning that the features are quite important for this process. It is only though with TF-IDF and PCA that some hymns and poems are correctly classified and in all clustering processes, hymns and poems achieve a 0% F1, recall and precision. When using the style markers, all epigrams are correctly classified, however 747/783 texts are classified as epigrams. According to the baseline (TF-IDF) a lot of religious texts are labelled as such, however most of them are misclassified as poems. Overall, the results indicate that there is a correlation between religious texts and poems, epigrams and poems, historical and religious, hymns and epigrams/religious texts.

6 Conclusion

In this study, the performance of different machine learning algorithms and stylometric features is tested on a classification problem of Medieval Greek texts. Supervised and unsupervised learning are both implemented. The Random Forest reaches an accuracy of 87%, overcoming the Support Vector Machine and their corresponding baselines, which were generated based on TF-IDF features. The repeated stratified 5-fold cross validation of RF presents a 97.6% accuracy. After standardization, the SVM reaches a 71% using the style markers as features. The unsupervised method fails to form the correct clusters as the performance levels are significantly low. The most important style markers are quite dissimilar between RF and SVM.

Further research should be conducted on this task. Both feature methods (style and TF-IDF) could be merged to test their efficiency in conjunction. The dataset should be furtherly studied for potential noise identification and more algorithms could be furtherly tested. The experiments could be conducted with a more balanced data amount, e.g. same amount of data in the training set to avoid imbalance issues and provide a better result interpretation. To address the misclassification issue in hymn and poem categories, rhythm style markers could be tested, like Balint et al. (2016) did. The selected features of this study are mainly lexical features, but vocabulary richness features could also be added in the list (such as hapax legomena) and test the performance. Finally, the style marker models could be retrained by only using the most important stylometric features and by adding new lexical stylometric features which might be more appropriate for this Medieval Greek classification problem.

References

Maha Al-Yahya. 2018. Stylometric analysis of classical arabic texts for genre detection. *The Electronic Library*.

Diego Raphael Amancio. 2015. A complex network approach to stylometry. *PloS one*, 10(8):e0136076.

Rafael T Anchiêta, Francisco Assis Ricarte Neto, Rogério Figueiredo de Sousa, and

Raimundo Santos Moura. 2015. Using stylometric features for sentiment classification. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 189–200. Springer.

Mihaela Balint, Mihai Dascalu, and Stefan Trausan-Matu. 2016. Classifying written texts through rhythmic features. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 121–129. Springer.

Hassaan Elahi and Haris Muneer. 2018. Identifying different writing styles in a document intrinsically using stylometric analysis. *The complete code and detailed documentation is available on the attached Github Link: https://github. com/harismuneer/Writing-Styles-Classification-Using-Stylometric-Analysis*.

Efthimios Gianitsos, Thomas Bolt, Pramit Chaudhuri, and Joseph Dexter. 2019. Stylometric classification of ancient greek literary texts by genre. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 52–60.

Lena Hettinger, Martin Becker, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2015. Genre classification on german novels. In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 249–253. IEEE.

Vipin Kumar and Sonajharia Minz. 2014. Poem classification using machine learning approach. In *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*, pages 675–682. Springer.

Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and PG Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195. IEEE.

Katerina Papantoniou and Yannis Tzitzikas. 2020. Nlp for the greek language: A brief survey. In *11th Hellenic Conference on Artificial Intelligence*, pages 101–109.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis lectures on human language technologies*, 5(2):1–157.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications.* " O'Reilly Media, Inc.".

Congzhou He Ramyaa, Khaled Rasheed, and Congzhou He. 2004. Using machine learning techniques for stylometry. In *Proceedings of International Conference on Machine Learning*.

Ilija Simić and Mauro Jurada. 2020. Stylometry based article classification and paper fingerprinting.

Hamid R Tizhoosh, Farhang Sahba, and Rozita Dara. 2008. Poetic features for poem recognition: A comparative study. *Journal of Pattern Recognition Research*, 3(1):24–39.

Appendices

A Train- Test sets

Epigrams (155)
Train: epigramma_0001 - epigramma_0112, epigramma_0114 - epigramma_0124, epigramma_0126
Test: epigramma_0127 - epigramma_0128, epigramma_0131 - epigramma_0135, epigramma_0137 - epigramma_0152, epigramma_0156 - epigramma_0161, epigramma_0163 - epigramma_0164

Hymns (12)
Train: religious_hymn_0001 - religious_hymn_0006, religious_hymn_0013 - religious_hymn_0014, religious_hymn_0016, religious_hymn_0034
Test: religious_hymn_0035, religious_hymn_0036

Historical (72)
Train: history_0001 - history_0024, history_0026 - history_0049, history_0051 - history_0058
Test: history_0059 - history_0074

Poems (72)
Train: poetry_0002, poetry_0005,
poetry_0008 - poetry_0011,
poetry_0013 - poetry_0064
Test: poetry_0065 - poetry_0078

Religious (472)
Train: religious_0001 - religious_0020
religious_0022 - religious_0049
religious_0138 - religious_0144
religious_0146 - religious_0247
religious_0249 - religious_0469
Test: religious_0470 - religious_0563

B Stylometric Features

- ἄλλος (allos, “other”) is computed by counting all inflected forms of ἄλλος, -η, -ο.

- αὐτός (autos, “self” or “him/her/it”) is computed by counting all inflected forms of αὐτός, -ή, -ό.

- Demonstrative pronouns are computed by counting all inflected forms of the three Greek demonstrative pronouns οὗτος, αὕτη, τοῦτο (houtos, haute, touto, “this”), ὅδε, ἥδε, τόδε (hode, hede, tode, “this”), and ἐκεῖνος, ἐκεῖνη, ἐκεῖνο (ekeinos, ekeine, ekeino, “that”).

- Selected indefinite pronouns are computedby counting all inflected forms of τις, τις, τι (tis, tis, ti, “any”) in non-interrogative sentences. Interrogative sentences are excluded because the Greek interrogative pronoun (τίς) is often identical in form to the indefinite pronoun.

- Personal pronouns are computed by counting all inflected forms of the pronouns ἐγώ (ego, “I”) and σύ (su, “you”).

- Reflexive pronouns are computed by counting all inflected forms of ἐμαυτοῦ (emautou, “he himself”).

- Conjunctions are computed by counting all instances of the common conjunctions τε, τ′ (te or t, “and”), καί, καὶ (kai, “and”), ἀλλά, ἀλλὰ (alla, “but”), καίτοι (kaitoi, “and indeed”), οὐδέ, οὐδὲ, οὐδ ′ (oude or oud, “and not”), μηδέ, μηδὲ, μηδ′ (mede or med, “and not”), οὔτε, οὔτ ′ (oute or out, “and not”), μήτε, μήτ′ (mete or met, “and not”), and ἤ, ἢ (e, “or”).

- μέν (men, “indeed”) is computed by counting all instances of μέν and μὲν.

- Particles are computed by counting all instances of ἄν, ἂν (an, a particle used to express uncertainty or possibility), ἆρα (ara, “then”), γέ, γ′ (ge or g, “at least”),δ′, δέ, δὲ (d or de, “but”), δή, δὴ(de, “indeed”), ἕως (heos, “until”), κ′, κε, κέ, κὲ, κέν, κὲν, κεν (k, ke, ken, a particle used to express uncertainty or possibility), μά(ma, used in oaths and affirmations, “by”), μέν, μὲν (men, “indeed”), μέντοι (mentoi, “however”),μὴν, μήν (men, “truly”),μῶν (mon, “surely not”), νύ, νὺ, νυ (nu, “now”), οὖν (oun, “so”), περ (per, an intensifying particle, “very”), πω (po, “yet”), and τοι (toi, “let me tell you”).

- Circumstantial markers are computed by counting all instances of ἔπειτα, ἔπειτ′ (epeita or epeit, “then”), ὅμως (homos, “all the same”), ὁμῶς (homos, “equally”), καίπερ (kaiper, “although”), and ἅτε, ἅτ′ (hate or hat, “seeing that”).

- Conditional markers are computed by counting all instances of εἰ,εἴ,εἶ, ἐάν, and ἐὰν(ei, ei, ei, ean, ean, all translated “if”).

- ἵνα (hina, an adverb of place often translated “where” or a conjunction indicating purpose often translated “in order that”) is computed by counting all instances of ἵνα and ἵν′ (hin).

- ὅπως (hopos, an adverb of manner often translated “how” or a conjunction indicating purpose often translated “in order that”) is computed by counting all instances of ὅπως.

- Fraction of sentences with a relative clause is determined by counting sentences that have one or more of the inflected forms of the Greek relative pronouns ὅς, ἥ, ὅ (hos, he, ho, “who” or “which”).

- Temporal and causal markers are computed by counting all instances of μέχρι (mekri, “until”), ἕως(heos, “until”), πρίν (prin, “before”), ἐπεί (epei, “when”), ἐπειδή (epeide, “after” or “since”), ἐπειδάν (epeiden, “whenever”), ὅτε (hote, “when”), and ὅταν (hotan, “whenever”).

- ὥστε (hoste, a conjunction used to indicate a result, “so as to”) not preceded by ἤ is calculated by counting all instances of ὥστε not immediately preceded by ἤ. This limitation is imposed to exclude instances in which ὥστε is part of a comparative phrase.

- The mean length of relative clauses is determined by counting the number of characters between each relative pronoun and the next punctuation mark.

- Interrogative sentences are computed by counting all instances of “;” (the Greek question mark).

- Regular superlatives adjectives are computed by counting all instances of -τατος, -τάτου, -τάτῳ, -τατον, -τατοι, -τάτων, -τάτοις, -τάτους, -τάτη, -τάτης, -τάτῃ, -τάτην, -τάταις, -τάτας, -τατα, -τατά, and τατε at word end. One inflected form, -ταται, is excluded so as to avoid confusion with the Homeric third person singular middle/passive indicative verb ending -αται. This method does not detect certain irregular superlatives, such as ἄριστος (aristos, “best”) or πρῶτος (protos, “first”), which would be significantly harder to disambiguate from non-superlative forms.

- Sentences with ὦ exclamations is determined by identifying sentences that have at least one instance of ὦ (o, “O”), a Greek exclamation.

- ὡς(hos, an adverb of manner often translated “how” or a conjunction often translated as “that,” “so that,” or “since,” among several other possibilities) is computed by counting all instances of ὡς.

C Elbow Method

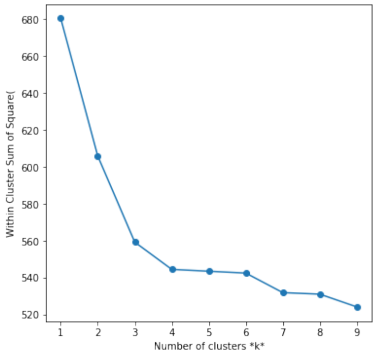

Figure 1: Elbow Method-TF-IDF

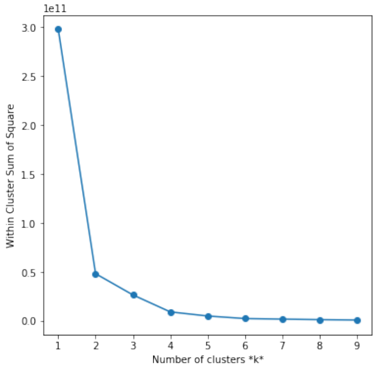

Figure 2: Elbow Method-Stylometric

D Clustering Visualization

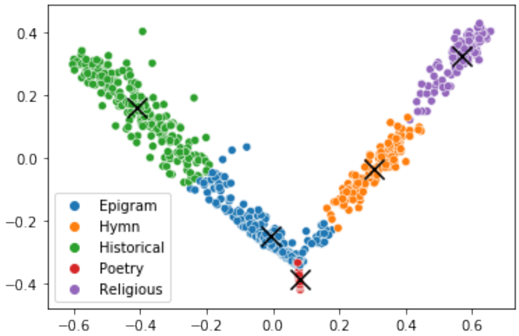

Figure 3: Clustering with TF-IDF

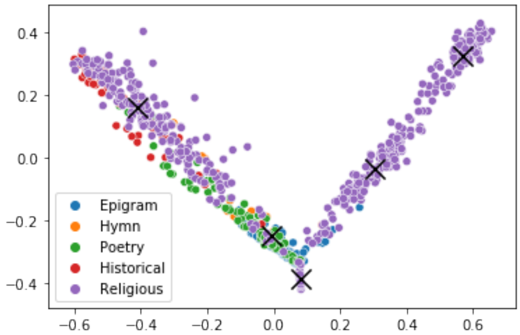

Figure 4: Clustering with TF-IDF - Ground Truth

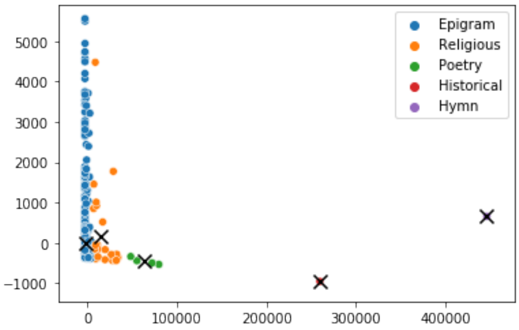

Figure 5: Clustering with style markers

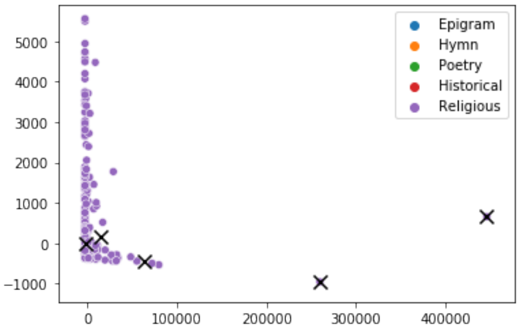

Figure 6: Clustering with style markers - Ground Truth

E Model Hyperparameters

Random Forest: n_estimators = 200, n_jobs = 1, random_state = 0
Stratified Cross Validation: n_splits=5, n_repeats=10,random_state=0
Support Vector Machine + Style markers: C = 10, kernel = 'rbf', gamma = 0.11 + Standard Scaler
Support Vector Machine Baseline (TF-IDF): C=1000, gamma=0.0