A Literature Survey: Machine Learning techniques with application to Shakespeare's Plays

Mavromati Galateia
Electrical and Computer Engineering
University of Thessaly
gmavromati@uth.gr

***Abstract*— Shakespeare is one the most prestigious writer of all time. However, he is thought to have collaborated with other authors into writing some of his plays. Authorship attribution (AA) is the process of determining the writer of a document. The Shakespeare authorship question is an argument that a collaborator or another author wrote the works attributed to Shakespeare. This AA problem has troubled literal scholars since the middle of the 19th century. In this survey, different approaches of this problem are presented and analyzed using supervised and unsupervised machine learning techniques. Some of the supervised techniques are Support Vector Machine, Artificial Neural Networks and Decision trees, while unsupervised is hierarchical clustering. Some features extraction mechanisms are introduced such as stylometry, function words and n-grams. Different methods and techniques resulted in collision and inconsistency between the outputs. For instance, Shakespeare was proposed by Artificial Neural Network to be the rightful author of Titus Andronicus, while hierarchical clustering suggested that Shakespeare is not the writer of this play. The candidate authors who are thought to be Shakespeare's collaborators, based on this survey, are John Fletcher and Christopher Marlowe.**

I. INTRODUCTION

William Shakespeare is one of the most important playwrights of all time. It is not surprising the fact that his work interests not only litterateurs, actors, schools and scientists but also people from various social levels. Everyone has heard about Romeo and Juliet, Hamlet's famous phrase "To be or not to be" or "A Midsummer night dream". It is a common belief that in Elizabethan era (1558–1603) authors used to collaborate with each other. From times to times scholars embrace the belief that Shakespeare did not solely write some of his plays, because of his lack of education and aristocratic sensibility, but other authors were involved in the writing. On the other hand, most modern-day scholars reject this claim, arguing that there is strong evidence that Shakespeare wrote the plays and poems attributed to him. This conflict has led to an endless academic debate with unsettled answers. The question of identifying the real author of the Shakespearean works is known as "Shakespeare authorship question". There are probably about 80 candidates that are thought to have cooperated with Shakespeare or to be the rightful authors for some of his plays but the most popular are John Fletcher, Sir Francis Bacon, Christopher Marlowe and Thomas Kyd. Moreover, some unidentified plays has been attributed to Shakespeare but for various reasons the authorship is questionable. Some of them are Edward III, The Two Noble Kinsmen and Pericles, Prince of Tyre. This group of plays is called Shakespeare apocrypha and has troubled researchers into finding the real author of these plays.

Authorship attribution (AA) is the process of attempting to identify the true author of a document or text, given a collection of documents whose authorship is known. AA constructs a classification problem and machine learning techniques are used in order to solve it. It is important to mention that it differs from text classification because in AA beside the text context, writing style plays a significant part in classification.

Scholars use different techniques in identifying Shakespeare's authorship. Plechac [1] uses Support Vector Machine (SVM) as a classifier together with rhythmic types in order to identify writing style, while Merriam [10] chooses Multilayer Perceptron combined with function words. Boyd [20] uses Linear discriminant analysis (LDA), Decision Trees (DTs) and SVM together with 5 types of measures to recognize the authors writing signature, while Aljumily [23] uses hierarchical and non hierarchical clustering together with function words, bi-gram and tri-grams. These different classifications and dataset results in contrasting outcomes. The results given are inconsistent with each other and can not provide a final answer. For instance, Merriam supports that "Double Falsehood" was written by Fletcher, while Boyd strongly believes that it is Shakespeare's play.

The rest of the paper is organized as follows. In Chapter II supervised machine learning techniques are presented. In subsection A, traditional methods such as SVM tries to detect collaboration in "Henry VIII" between Shakespeare, Fletcher and Massinger. Moreover, LDA, DTs and SVM are trying to find collaboration in "Double Falsehood" based on psychological features. In subsection B, Artificial Neural Networks takes the wheel and is trying to find the who wrote various plays between Shakespeare and Fletcher, and Shakespeare and Marlowe. In Chapter III, unsupervised machine learning techniques are presented and more specifically hierarchical and non hierarchical linear and non linear clustering attempts to find which of the disputed plays are written by Shakespeare.

II. SUPERVISED MACHINE LEARNING

A. Traditional Machine Learning Techniques

Various traditional machine learning methods such as SVM, LDA and DTs are seeking to detect collaboration

between the plays attributed to Shakespeare "Henry VIII" and "Double Falsehood". The possible candidates are Shakespeare, Fletcher and Massinger for "Henry VIII" and Shakespeare, Fletcher and Theobald for "Double Falsehood".

*1) **Collaboration in "Henry VIII":*** In 1850, James Spedding proposed that Shakespeare did not solely write Henry VII but Fletcher involved in writing various scenes [2]. Many scholars supported this idea and was also proposed, later, that Massinger could also be involved in writing Henry VIII [3], [4]. In [1], this problem tries to be solved with machine learning methods in order to detect the involvement of Fletcher and Massinger.

Attribution of particular scenes. In the first method, an attribution of individual scenes of Henry VIII is performed using the Support Vector Machine as a classifier (see Appendix VI-A) .

In order to perform the attribution two other parameters should be taken into consideration: the frequencies of 500 most frequent rhythmic types, i.e. the bit string representing the distribution of stressed and unstressed syllables in a particular line, and the frequencies of 500 most frequent words as a feature set. As training samples, individual scenes of plays written by Shakespeare, Fletcher, and Massinger are used written in the same period as Henry VIII. The samples consists of:
- **Shakespeare**: The Tragedy of Coriolanus (5 scenes), The Tragedy of Cymbeline (27 scenes), The Winter's Tale (12 scenes), The Tempest (9 scenes)
- **Fletcher**: Valentinian (21 scenes), Monsieur Thomas (28 scenes), The Woman's Prize (23 scenes), Bonduca (18 scenes)
- **Massinger**: The Duke of Milan (10 scenes), The Unnatural Combat (11 scenes), The Renegado (25 scenes)

In order for the results to be more accurate some actions need to be performed. The training data is imbalanced because Shakespeare has 53 training samples, Fletcher 90 and Massinger 46. For the results to be unbiased the number of training samples per author was decided by random selection. To avoid overfitting which could be caused by testing the model on the scenes from the same play as it was trained on, the classification of the scenes of each play was accomplished by a model trained on the rest scenes of the remaining plays. The entire process was executed 30 times in order to get more representative results which resulted in 30 classifications of each scene. And lastly, cross-validations are performed not only of the combined models (500 words $\cup$ 500 rhythmic types), but also of the words-based models (500 words) and versification-based models (500 rhythmic types) alone. In Table I it can be seen that when using models based on the 500 most frequent rhythmic types there is high accuracy in Fletcher and Shakespeare recognition (between 0.97 to 1.00 with the exception of Valentinian) but lower accuracy in Massinger (between 0.81 to 0.88). With the word based models the accuracy yields between 0.95 and 1.00. With combined models the accuracy is much better, hence, combined models provide a reliable discrimination between the authors' styles.

		rythmic types	words	combination
Shakespeare	Corionalus	0.98	1	1
	Cymbeline	0.98	1	1
	Winter's Tale	0.99	1	1
	Tempest	0.97	1	1
Fletcher	Valentinian	0.84	0.95	0.96
	Monsieur Thomas	1.00	0.98	1
	Woman's Prize	0.98	1	1
	Bonduce	0.98	0.98	1
Massinger	Duke of Milan	0.81	0.99	0.99
	Unnatural Combat	0.83	1	1
	Renegado	0.88	1	1

TABLE I: Accuracy of authorship recognition provided by the models based on (1) 500 most frequent rhythmic types, (2) 500 most frequent words, (3) 1000- dimensional vectors combining features (1) and (2). The number gives the share of correctly classified scenes through all 30 iterations.

The next step was for the classifier to be applied to the individual scenes of Herny VIII. The conclusions derived from **Table II** are that Massinger did not take part in writing Henry VIII. The probability of collaboration with Fletcher and Shakespeare is extremely high because all 30 models agree that 7 of the scenes were written by Shakespeare and 5 of them were writter by Fletcher. Lastly, the results correspond to Spedding's proposals to a high extent with the exception of two scenes.

	Classification results			
	Shakespeare	Fletcher	Massinger	Spedding's Attribution
1.1	**30**	0	0	Shakespeare
1.2	**30**	0	0	Shakespeare
1.3	0	**30**	0	Fletcher
1.4	0	**30**	0	Fletcher
2.1	0	**20**	10	Fletcher
2.2	0	**30**	0	Fletcher
2.3	**30**	0	0	Shakespeare
2.4	**30**	0	0	Shakespeare
3.1	0	**30**	0	Fletcher
3.2	**30**	0	0	**Shakespeare/Fletcher**
4.1	**30**	0	0	**Fletcher**
4.2	0	**23**	7	Fletcher
5.1	**30**	0	0	Shakespeare
5.3	9	**21**	0	Fletcher
5.4	7	**23**	0	Fletcher
5.5	0	**30**	0	Fletcher

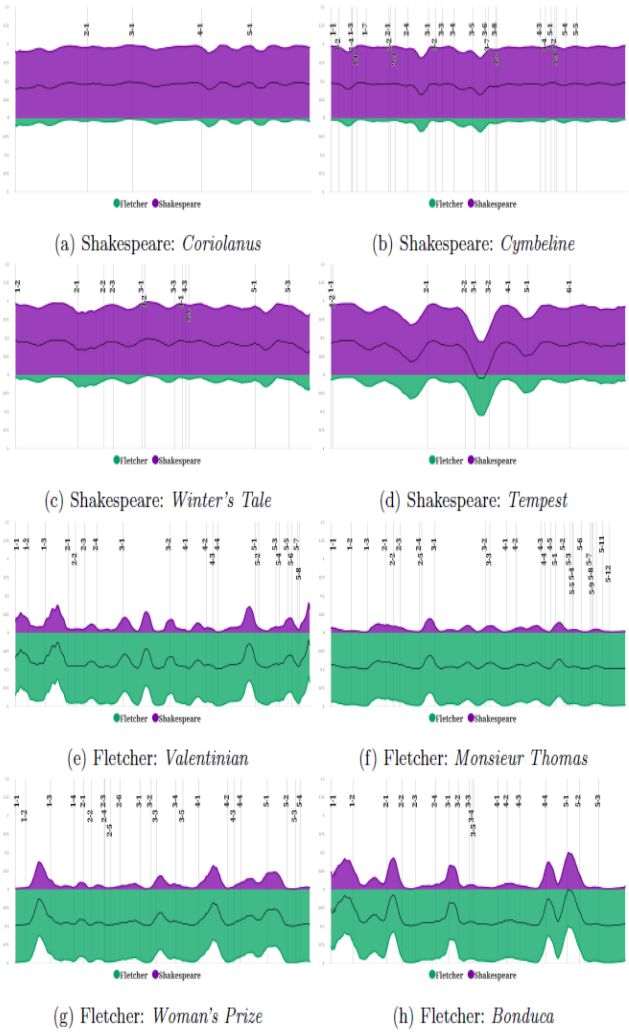
TABLE II: Classification of individual scenes of H8. The number indicates how many times out of 30 iterations the author has been predicted to a given scene. The highest value in each row is printed in bold. The rightmost column indicates to which author the scene is attributed by Spedding. Where Spedding differs from our results, a bold face is used.

Rolling Attribution

Many scholars suggested that the shift of authorization did not respect scene boundaries. Therefore, the attribution of particular scenes result should not be the final conclusion. In order to get a clearer picture, boundaries should not be taken into consideration. Rolling attribution provides a more reliable method. It was introduced by Maciej Eder [5] as a technique designed for cases involving mixed authorship. The general idea is that it group lines together where a group consists of d lines. The training set used are 4 plays

by Shakespeare and 4 plays by Fletcher. For each play 30 models are trained on the remaining data. The extreme lines of scenes are omitted and $d = 5$. The output of classification of each part is transformed to probability distribution using Platt's scaling [6].

The results of rolling attribution in these 8 plays can be seen in Figure 1 where each data point corresponds to a group of five lines and gives the mean probability of Shakespeare's and Fletcher's authorship. The values of Fletcher are displayed as negative in order to be distinct. It is obvious that the probability of an author having an involvement in the other author's plays is very low with the exception of Shakespeare's play "The Tempest" in the second act of scene 2 and in the first scene of act 5 in Fletcher's "Bonduca".



(a) Shakespeare: *Coriolanus* (b) Shakespeare: *Cymbeline*

(c) Shakespeare: *Winter's Tale* (d) Shakespeare: *Tempest*

(e) Fletcher: *Valentinian* (f) Fletcher: *Monsieur Thomas*

(g) Fletcher: *Woman's Prize* (h) Fletcher: *Bonduca*

Fig. 1: Rolling attribution of 4 plays by Shakespeare and 4 plays by Fletcher based on 500 most frequent rhythmic types and 500 most frequent words. Vertical lines indicate scene boundaries.

Again the following step is applying rolling attribution to Henry VIII. Based on Figure 2 it can be seen that all three sets of models indicate that the shift of authorship happened at the end of scene 1.2. The models indicate Fletcher's authorship in scenes 1.3, 1.4, 2.1, 2.2 and Rhythmic types suggests that the shift of authorship happened at the end of 2.2, while word-based models indicate that the shift happened before the end of the scene. Another shift happened in scene 2.4 giving Fletcher the "pen" to write scene 3.1 and 3.2 is attributed to both Shakespeare and Fletcher. Fletcher's style can be seen in scene 4.2 and the models disagree on the shift in that scene providing a result of mixed authorship.At the end of scene 5.1 Fletcher takes again authorship and writes scenes 5.2, 5.3, 5.4, 5.5 with a possibility of Shakespeare's involvement in 5.4 based on rhythmic types model.

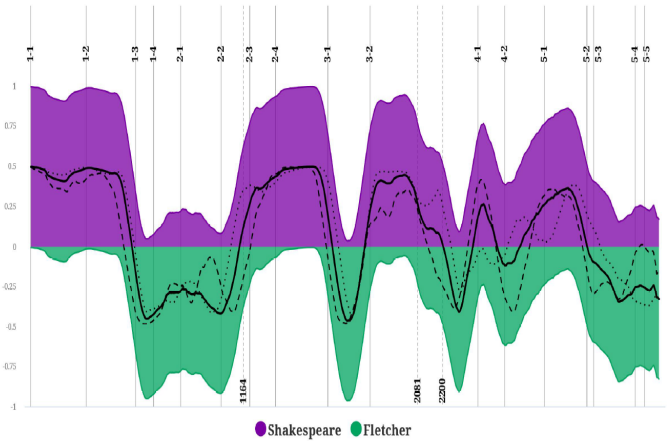


Fig. 2: Rolling attribution of H8 based on 500 most frequent rhythmic types and 500 most frequent words. Vertical lines indicate scene boundaries. Dashed line indicates results of rolling attribution based solely on 500 most frequent rhythmic types, dotted line indicates results of rolling attribution based solely on 500 most frequent words.

It can be highly supported that Henry VIII is a result of collaboration between John Fletcher and William Shakespeare, while the participation of Philip Massinger is rather unlikely. Moreover, from rolling attribution method it can be seen that each scene is a work from a single author.

*2) **Collaborations in "Double Falsehood"**:* Another authorship attribution emerges in [20] where the play of interest is Double Falsehood. In 1728, Lewis Theobald published a play titled Double Falsehood that was based on Shakespeare's lost plays that he had found. Psychological signature would help significantly in recognizing the author of a play. This research focuses in new techniques combining contemporary authorship identification (AID) methods with the psychology of language to infer who wrote Double Falsehood. The dataset used consists of Double Falsehood and 33 plays by Shakespeare, 9 by Fletcher and 12 by Theobald that are thought not to be collaborations. The text was preprocessed in order for Old English words to be converted into today's equivalent ones. In order to calculate the percentage of words of interest in each text that belonged in the measures presented below, text analysis programs, Linguistic Inquiry and Word Count (LIWC) [21] and RIOT Scan [22] were used. All plays in our corpus were quantified

with the five types of measures:

- Function words: consists of personal pronouns, impersonal pronouns, articles, prepositions, auxiliary verbs, conjunctions, negations, and high-frequency adverbs lacking direct referents. Average rate of function words in the current sample is 53.4%. Useful in order to identify unique psychological characteristics typical of each of the three candidate authors.
- Categorical-dynamic index (CDI): reveals psychological characteristics and how complex the thinking of an author is. Used together with average sentence length and the use of large words.
- LIWC Content words: The LIWC dictionary consists of more than 40 content categories. Based on that they were able to determine the rates of content word categories in the texts.
- Thematic signatures: the meaning-extraction method, or MEM was used to identify 13 broad themes. In the works of the three authors these themes were measured and a writing signature began to arise. Generally, the relative presence or absence of a given theme can be a useful cue to an individual's psychological characteristics.
- Low-base-rate tell: words that are used at a low rate can cue to authorship although they are not psychologically meaningful.

The classification techniques used are: linear discriminant analysis (LDA), decision trees (DTs) and support vector machines (SVMs).

Linear discriminant analysis: LDA operates as a subtype of factor analysis by using the linear combinations of variables in this specific application to form vectors, which are then used to differentiate authors. Ficher's classical LDA was used for the classification (see Appendix VI-B).

Decision trees: DTs differentiates authors on sequential chains of variables. Specifically, starts by searching for a language variable that can best distinguish between two or more authors. This process is then usually repeated until the information gained from additional variables is negligible or not necessary and it constructs a series of fulcrums that are considered one at a time, each one channeling an observation to a final authorship designation based on language. The J48 DT algorithm was used for the classification (see Appendix VI-C).

Support vector machines: an SVM seeks to use combined inputs simultaneously as a functional way of discriminating between multiple authors. Moreover, it has the ability to combine all features into non-linear vectors, allowing for more nuanced differentiation between candidates. The Sequential Minimal Optimization (SMO) SVM was used for the classification (see Appendix VI-A).

After all the plays of the corpus were quantified by the five measures and classified using the techniques proposed above, Double Falsehood came to the foreground. The whole play analysis can be seen in Table III. The three authors had distinct stylistic psychological signatures along function-word dimensions and Shakespeare is the proposed author for Double Falsehood by all three models. It is noticable in 3a that Fletcher uses articles and prepositions in a low frequency. Quite the opposite does Theobald. Shakespeare showed a stylistic trend toward Fletcher. In CDI the authors had unique degrees of cognitive complexity. Theobald was the most complex of three, Shakespeare was in the middle and Fletcher had the most dynamism (3b). Again Shakespeare was found the winner of Double Falsehood. The LIWC content categories was found to bifurcate the classifiers as LDA voted for Theobald to be the rightful owner of Double Falsehood, while DT and SVM voted for Shakespeare. Fact that did not happen in Thematic singatures and low-base-rate tell as they both votes Shakespeare in all classifiers.

	LDA		J48 DT		SMO SVM	
Measure	Best candidate	p (%)	Best candidate	p (%)	Best candidate	p (%)
Function-word classes	Shakespeare	91.4	Shakespeare	96.8	Shakespeare	83.3
CDI, WPS, large words	Shakespeare	61.0	Shakespeare	93.1	Shakespeare	78.9
LIWC content categories	Theobald	97.3	Shakespeare	97.1	Shakespeare	75.4
Thematic signatures	Shakespeare	100	Shakespeare	97.1	Shakespeare	99.8
Low-base-rate tells	Shakespeare	83.8	Shakespeare	100	Shakespeare	97.1

TABLE III: Double Falsehood results for Each Language Measure, by Classification Technique.Whole play analysis.

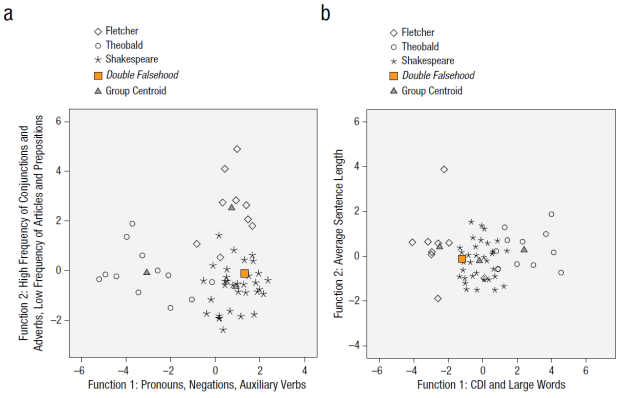


Fig. 3: Results for the linear discriminant analyses using (a) the eight classes of function words and (b) cognitive and stylistic complexity measures. The graphs plot the locations of individual plays and the group centroids for Fletcher, Theobald, and Shakespeare using the functions the models relied on to discriminate among the authors. CDI = categorical-dynamic index.

The next experiment was to break the play into acts. The results are shown in IV. Shakespeare has a dominant position in the first three acts. However, in the last two we observe that Fletcher has began to arise making his contribution apparent. Theobald's contribution is almost minor. This result hints that Shakespeare and Fletcher might have collaborated in the making of Double Falsehood but Shakespeare's stamp is much more crucial than Fletcher's.

B. Artificial Neural Networks

One of the authors that are thought to have cooperated and have shared authorship with Shakespeare, as proposed before, is John Fletcher. In [10] neural computing is involved to recognize the collaboration between the two authors on plays

	Act I		Act II		Act III		Act IV		Act V	
Classification and Measure	Best candidate	p (%)	Best candidate	p (%)	Best candidate	p (%)	Best candidate	p (%)	Best candidate	p (%)
LDA										
Function-word classes	Shakespeare	95.6	Shakespeare	88.7	Fletcher	54.6	Fletcher	71.4	Fletcher	82.3
CDI, WPS, large words	Shakespeare	66.6	Shakespeare	74.1	Shakespeare	50.8	Fletcher	64.4	Fletcher	63.3
LIWC content categories	Shakespeare	99.8	Theobald	93.2	Shakespeare	99.7	Fletcher	99.4	Shakespeare	71.6
Thematic signatures	Shakespeare	99.7	Shakespeare	96.2	Shakespeare	90.5	Fletcher	72.7	Fletcher	83.1
Low-base-rate tells	Shakespeare	43.5	Shakespeare	50.1	Shakespeare	62.1	Theobald	56.6	Shakespeare	45.7
J48 DT										
Function-word classes	Shakespeare	87.2	Shakespeare	87.2	Shakespeare	54.5	Shakespeare	54.5	Fletcher	96.0
CDI, WPS, large words	Shakespeare	69.3	Shakespeare	69.3	Shakespeare	69.3	Shakespeare	69.3	Fletcher	83.8
LIWC content categories	Shakespeare	93.4	Fletcher	61.9	Shakespeare	93.4	Fletcher	61.9	Fletcher	95.0
Thematic signatures	Shakespeare	92.9	Shakespeare	92.9	Shakespeare	92.9	Shakespeare	92.9	Shakespeare	92.9
Low-base-rate tells	Shakespeare	99.4	Shakespeare	99.4	Theobald	100	Fletcher	87.2	Shakespeare	99.4
SMO SVM										
Function-word classes	Shakespeare	98.5	Shakespeare	97.3	Shakespeare	67.0	Shakespeare	50.6	Fletcher	58.5
CDI, WPS, large words	Shakespeare	81.2	Shakespeare	79.7	Shakespeare	75.6	Shakespeare	57.2	Shakespeare	61.1
LIWC content categories	Shakespeare	95.2	Theobald	99.6	Shakespeare	96.3	Fletcher	97.3	Fletcher	69.7
Thematic signatures	Shakespeare	93.5	Shakespeare	68.7	Shakespeare	93.4	Fletcher	83.5	Fletcher	83.5
Low-base-rate tells	Shakespeare	88.4	Shakespeare	97.4	Theobald	45.2	Fletcher	38.9	Shakespeare	99.6

TABLE IV: Double Falsehood Results for Each Language Measure, by Act and Classification Technique

that are thought to be jointly written. Analytically, artificial neural networks use non-linear mathematical equations to successively develop meaningful relationships between input and output variables through a learning process. They are not disturbed by noisy data or non-linear correlations hence it is perfect for stylometric research. The neural network used in [10] is Multi-Layer Perceptron(MLP) which consists of an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP can easily handle the non-linear and interactive effects of explanatory variables. It is important, however, to find the correct amount of neurons used. If an MLP has too many inputs relatively to the number of training vectors, it will lose its ability to generalize to new data. If too few hidden units are given then it fails to capture all the features in the data, while too many leads to a failure to generalize. Both Merriam (1992) [12] and Horton (1987) [11] have studied the choice of discriminators in order for discriminators to be able to show reasonable stability across the corpus of an author's work and maintain their reliability when broken into acts which is very important in investigating collaborations. The discriminators chosen are 5 where Merriam's set consists of the ratios: did/(did+do), no/T-10, no/(no+ not), to the/to, upon/(on+upon) where T-10 is Taylor's ten function words (but, by,for, no, not, so,that, the, to, with) [17].

And Horton's set consists of ratios formed by dividing the total numbers of words in a sample by the number of occurrences of the following five function words: are, in, no, of, the.

The final parameters given in MLP was 5 neurons in input layer, 2 in output layer and after some testing they reached to the conclusion that 3 hidden units were sufficient enough to give cross validation accuracy around 90% and for misclassified vectors to be approximately equally divided between the two classifications, providing that the MLP is unbiased in its discrimination process. The complete MLP is shown in Figure 4. A training set was formed for each discriminator which consisted of $k = 100$ vectors, where each vector took the form $(ratio1,\ ratio2,\ ratio3,\ ratio4,\ ratio5,\ authorID)$. Each ratio was computed by word counts on 1,000-word samples from works that are verified not to be a collaborative work (Table V). At the exportation of each set of the ratios normalization was performed in order to give zero mean and unit standard deviation to make sure each discriminator contributes equally in the training process.

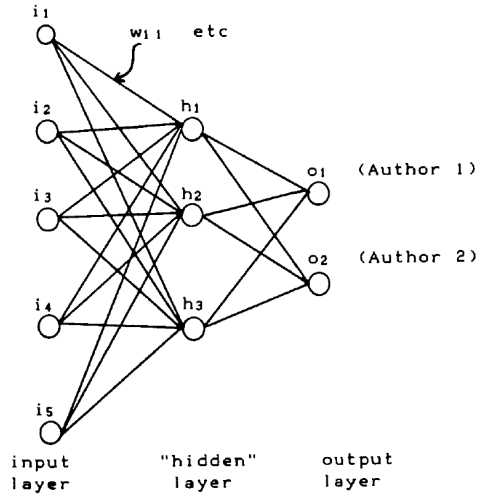


Fig. 4: Topology of a stylometric multi-layer perception for classifying works of two authors using five discriminators

The results from the training process are: Merriam-based network (MNN) achieved a cross validation accuracy of 90%, with the 10% misclassified being split into 6% Shakespeare classified as Fletcher, and 4% Fletcher classified as Shakespeare.The Horton-based network (HNN) achieved 96% cross-validation accuracy, with the both modes of misclassification lying at 2%.

The first experiment was for HNN and MNN to classify the test set shown in Table V where each network also provided a measure of degree which attributed the class that the play belonged. The measured called Shakespearian

	Shakespeare	Fletcher
TRAINING SET	The Winter's Tale	The Chances
	Richard III	The Womans Prize
	Love's Labour's Lost	Bonduca
	A Midsummer Night's Dream	The Island Princess
	Henry IV	The Loyal Subject
	Henry V	Demetrius
	Julius Caesar	Enanthe
	As You Like It	
	Twelfth Night	
	Antony and Cleopatra	
TEST SET	All's Well that Ends Well	Valentinian
	Comedy of Errors	Monsieur Thomas
	Coriolanus	
	King John	
	Much Ado about Nothing	
	The Merchant of Venice	
	Richard II	
	Romeo and Julie	

TABLE V: Dataset used for MLP. Separated into training set and test set.

Characteristics Measure (SCM) defines as:

$$SCM = \frac{\Omega_s}{\Omega_s + \Omega_f}$$

where

- Ω_s = values of output from Shakespeare
- Ω_f = values of output from Fletcher

Hence, the stronger the output of the neuron from Shakespeare in contrast to Fletcher, the higher the SCM.

Play	Merriam SCM	Merriam Verdict	Horton SCM	Horton Verdict
All's Well that Ends Well	0.75	Shakespeare	0.71	Shakespeare
Comedy of Errors	0.74	Shakespeare	0.92	Shakespeare
Coriolanus	0.90	Shakespeare	0.91	Shakespeare
King John	0.84	Shakespeare	0.98	Shakespeare
Much Ado about Nothing	0.76	Shakespeare	0.91	Shakespeare
The Merchant of Venice	0.67	Shakespeare	0.97	Shakespeare
Richard II	0.81	Shakespeare	0.92	Shakespeare
Romeo and Julie	0.80	Shakespeare	0.87	Shakespeare
Valentinian	0.46	Fletcher	0.30	Fletcher
Monsieur Thomas	0.32	Fletcher	0.29	Fletcher

TABLE VI: Multi-layer perception results for core canon Shakespeare and Fletcher.

From Table VI it is clear that both HNN and MNN are able to classify correctly the plays to the rightful author. It can also be seen that despite the different sets of discriminators tha results of SCMs are similar and the paper showed that the correlation coefficient between the SCMs produced by the two MLPs is 0.894.

The next step is to check the behavior of the SCMs in the different acts of the plays. The plays used are Shakespeare's The Tempest and The Merry Wives of Windsor and Fletcher's Valentinian and Monsieur Thomas. The Merriam-based network reached accuracy of 65% as it misclassified Acts II and IV of the Tempest, Acts I and III of The Merry Wives of Windsor, Acts II and V of Valentinian, and Act IV of Monsieur Thomas. Horton-based network had a considerably better behavior achieving with an 85% accuracy as it was found to misclassify Acts III and IV from The Tempest and Act V from Valentinian. The overall result was that in whole plays classification the accuracy lied in 90% but when classifying separate Acts the accuracy fell significantly for MNN, while HNN remained pretty effective.

The interesting part of this survey is to apply the classification methods to plays that are thought to be collaborative work or anonymous plays that scholars tend to give credit to specific authors. The following plays at some time been linked to Shakespeare and Fletcher: The Double Falsehood, The London Prodigal, Henry VIII, and The Two Noble Kinsmen.

Play	Merriam SCM	Merriam Verdict	Horton SCM	Horton Verdict
Entire plays				
Double Falsehood	0.40	Fletcher	0.37	Fletcher
London Prodigal	0.31	Fletcher	0.30	Fletcher
Henry VIII	0.84	Shakespeare	0.94	Shakespeare
Two Noble Kinsmen	0.78	Shakespeare	0.65	Shakespeare

TABLE VII: Merriam and Horton MLP results for disputed plays

Based on Table VII it it can be seen that both SCMs state that anonymous play Double Falsehood as an entire play is Fletcher's work which agrees to [14]. However, if the play broke down in acts (Table VIII) then the SCM suggests that 3 out of 5 scenes are Shakespeare's construction. Given the greater statistical noise in the discriminators at the level of acts, it does not provide reliable information. The London Prodigal, also an anonymous play, has Fletcherian style except for Act I which has Shakespeare's influence. The results for Shakespeare's Henry VIII taken as either as an entire play or in acts indicate that is a solely Shakespearean play with a small exception to Act V that suggests a strongly Fletcherian contribution as before stated in [15]. Lastly, the Two Noble Kinsmen as an entire play support a Shakepearean attribution but with relatively low SCMs which indicates that there might be a collaboration. When broken in Acts it is clear that SCM attributed Acts I and V to Shakespeare, Acts II and III to Fletcher and the low percentage in Act IV shows that there might be a collaboration in the particular scene. Following the previous paper, in [13] the goal is to discriminate between the works of Shakespeare and his contemporary Christopher Marlowe. An extended study in [16] suggested that the best set of stylometric ratios are no/T-1O, (of x and)/of, so/T-1O, (the x and)/the, with/T-1O where x represents any word and T-10 is Taylor's ten function words (but, by, for, no, not, so, that, the, to, with ([17]).

There are again five inputs, three hidden units, and two outputs. The training set can be seen in Table IX and the test set consists of the remaining 26 Shakespeare's plays and Marlowe's plays Doctor Faustus (1616), Dido, Massacre at Paris, and The Jew of Malta. The Merriam-based Shakespeare-Marlowe network (MERMAR) in the training set achieved a cross-validation accuracy of 90%, with the 10% misclassified being split into 6% Shakespeare classified as Marlowe, and 4% Marlowe classified as Shakespeare. The signal strength from the two MERMAR outputs can be seen in Table X and the formula for the SCM is:

$$SCM = \frac{\Omega_s}{\Omega_s + \Omega_M}$$

Play by acts	Horton SCM	Horton Verdict
Double Falsehood		
Act I	0.66	Shakespeare
Act II	0.87	Shakespeare
Act III	0.29	Fletcher
Act IV	0.73	Shakespeare
Act V	0.29	Fletcher
London Prodigal		
Act I	0.89	Shakespeare
Act II	0.29	Fletcher
Act III	0.34	Fletcher
Act IV	0.28	Fletcher
Act V	0.30	Fletcher
Henry VIII		
Act I	0.98	Shakespeare
Act II	0.85	Shakespeare
Act III	0.97	Shakespeare
Act IV	1.00	Shakespeare
Act V	0.57	Shakespeare
Two Noble Kinsmen		
Act I	0.93	Shakespeare
Act II	0.30	Fletcher
Act III	0.32	Fletcher
Act IV	0.60	Shakespeare
Act V	0.91	Shakespeare

TABLE VIII: Horton MLP results for disputed plays broken in acts.

	Shakespeare	Marlowe
TRAINING SET	The Winter's Tale	Tamburlaine I
	Richard III	Tamburlaine II
	Love's Labour's Lost	and Edward II
	A Midsummer Night's Dream	The Island Princess
	Henry IV	The Loyal Subject
	Henry V	Demetrius
	Julius Caesar	Enanthe
	As You Like It	
	Twelfth Night	
	Antony and Cleopatra	

TABLE IX: Dataset used for MLP. Separated into training set and test set.

where Ω_s = values of output from Shakespeare and Ω_M = values of output from Marlowe. Hence, the stronger the output of the neuron from Shakespeare in contrast to Marlowe, the higher the SCM. It can be seen that the MLP gives authorship to Shakespeare for the play The Jew of Malta and Henry VI, Part 3 is given to Marlowe. The success rate reaches to 93%.

Table X is graphically displayed in Figure 5 where the diagonial line represents the Shakespearean singal being equal to Marlowe's, hence the values above the line are classified as Shakespearean and below as Marlovian. The wrong attribution of The Jew of Malta and Henry VI,Part 3 could be because of the 10% classification error. However, because the data was less noisy and larger the expectations

Play	Marlowe signal	Shakespeare signal	SCM	MERMAR verdict
SHAKESPEARE				
IH6	0.097	0.947	0.91	S
2H4	0.058	1.022	0.95	S
2H6	0.318	0.681	0.68	S
3H6	0.601	0.409	0.41	M
ADO	- 0.051	0.994	1.05	S
AWW	-0.003	0.984	1.00	S
CE	0.229	0.768	0.77	S
COR	0.102	0.923	0.90	S
CYM	- 0.043	1 077	1 04	S
H8	0.025	0.965	0.98	S
HAM	0.006	0.941	0.99	S
JN	0.399	0.659	0.62	S
KL	-0.026	0.977	1 03	S
MAC	0.211	0.799	0.79	S
MM	0.153	0.844	0.85	S
MV	0.129	0.889	0.87	S
MWW	0.076	0.931	0.93	S
OTH	0.188	0.793	0.81	S
R2	0.095	0.955	0.91	S
ROM	0.110	0.930	0.89	S
TGV	- 0.021	1 011	1.02	S
TIM	- 0.001	1.039	1.00	S
TITUS	0.388	0.674	0.64	S
TMP	- 0.030	1.043	1.03	S
TRO	- 0.025	0.972	1.03	S
TS	- 0.043	1 009	1.05	S
MARLOWE				
Dido	0.720	0.376	0.34	M
Faust	0.777	0.231	0.23	M
Jew Malta	- 0.104	1.036	1 11	S
Massac	0.850	0.125	0.13	M
ANONYMOUS				
EDW3	0.494	0.587	0.54	S
Conten	0.841	0.091	0.10	M
Tragedy	0.806	0.148	0.16	M

TABLE X: Results for the rest of the Shakespeare and Marlowe canons

for MERMAR was to have >90% reliability when used on entire plays. Some scholars state that The Jew of Malta was not Marlowe's which could justify the low signal, placing it as an outlier. Another thought is that the play is a production of mixed authorship and that's why the network, by putting in it in Shakespeare's play, declared against Marlowe's authorship. Based on [18] Henry VI,Part 3 is Shakespeare's revision of Marlowe's play The True Tragedy of Richard Duke of York and it possesses a Marlovian character.

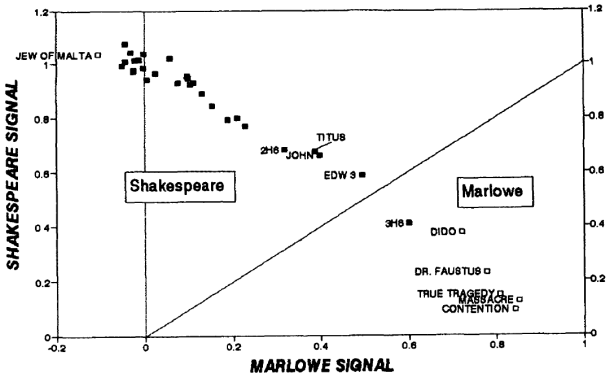


Fig. 5: Classifications of Shakespeare (closed squares) and Marlowe (open squares) plays not previously seen by MERMAR

MERMAR was called to classify three anonymous plays: Edward III, The First Part of the Contention and The True Tragedy. In Edward III the classification showed that

Shakespeare is the author which agrees to many scholars but the SCM is pretty low ($SCM = 0.54$). This indicates an influence of Marlowe as stated before in [19]. Tucker Brooke considered both The Contention and the Octavo The True Tragedy to be Marlovian, views supported by MERMAR.

As stated before Henry VI,part 3 is the Marlovian The True Tragedy with Shakespearean additions. Hence when subtracting Shakespeare's addition it produces a play let's call it 3H6(Tragedy) and when adding the revised scenes it's called 3H6(non-Tragedy). In Figure 6 the results of MERMAR are displayed and it is noticeable that Tucker Brooke's hypothesis is confirmed but of course it needs further consideration.

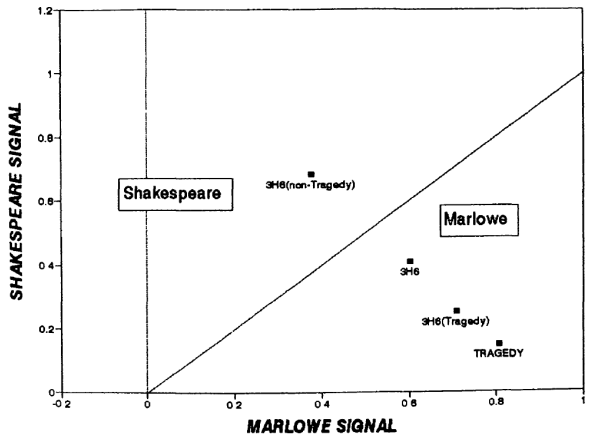


Fig. 6: MERMAR findings for Tucker Brooke's hypothesis

III. Unsupervised Machine Learning

The authorship attribution in [23] was encounter with Hierarchical and Non-Hierarchical Linear and Non-Linear Clustering Methods. Cluster analysis aims to detect and graphically to reveal structures or patterns in the distribution of data items, variables, or texts, in n-dimensional space, where n is the number of variables used to describe an author's style. The authors involved are Sir Francis Bacon, William Shakespeare, John Fletcher, Christopher Marlowe and Thomas Kyd. Specifically, the corpus used consisted of:

- 9 belong to Sir Francis Bacon
- 6 works to William Shakespeare(five history plays and one tragedy)
- 7 works to John Fletcher (tragic-comedies)
- 7 works to Christopher Marlowe (five tragic-histories and two tragedies)
- 4 works to Thomas Kyd (tragedies),
- 9 disputed works (six history plays and three tragedies).

After some research, it was found that the the most suitable and reliable stylistic criteria are function words, word n-grams, and character n-grams assuming that these don't change much from edition to edition. For the purpose of the analysis 135 function words (e.g. the, of, i, you), 100 word bi-grams (e.g and now- ,and with-, and so), and 24930 letter tri-grams (e.g. sti- uch- our- thr- men) were examined. In order to represent the data mathematically Vector Space Model (VSM) was used. The matrices generated was of dimensions: 42×135 D_{FW}, 42×100 D_{bigram}, and 42×24930 $D_{trigram}$.Where:

- Each of the 42 rows of D_{FW} represents a function word frequency profile for a corresponding text and each of the 135 columns represents a different function word.
- Each of the 42 rows of D_{bigram} represents a word bi-gram frequency profile for a corresponding text and each of the 100 columns represents a different word bi-gram.
- Each of the 42 rows of $D_{trigram}$ represents a character triple-gram frequency profile for a corresponding text and each of the 24930 columns represents a different character tri-gram.

Prior to clustering analyzing some actions needed to be taken. The first was dimensionality reduction. The variance of each variable was measured in order to identify and keep the most significant ones. The variance of a set of variable values is the average deviation of those values from their mean:

$$v = \frac{\sum_{i=1\ldots n}(x_i - \mu)^2}{n}$$

The final dimensions are 42×60 D_{FW}, 42×30 D_{bigram}, and 42×40 $D_{trigram}$. The next action implemented was length normalization. Some clusters grouped the vectors based on length and not on fuction word frequency profiles. This caused skewness and unreliability. The mathematical function that solved this problem for each row is:

$$M_i = M_i \frac{\mu}{length(C_i)}$$

where, M_i = the count for a given variable, μ = the mean document length and $length(C_i)$ = the total number of frequency counts occurring in that row vector.A linear hierarchical cluster analysis is applied on a 42×60 D_{FW} that is dimensionality-reduced and length-normalized. The result of this is shown in Figure 7 , where clustering by text length is removed.

Two reasons lead the researchers to apply hierarchical and non-hierarchical linear and non-linear clustering methods. The first is the difficulty they encountered to determine the presence of non-linearity in high dimensional data. The fact that hierarchical methods are linear made it necessary to include non-linear methods in order for the results to be more reliable. The second is that different methods or classes provide better results. Hence, in addition to hierarchical method they used non-hierarchical method.

Hierarchical cluster analysis forms clusters iteratively, by successively joining or splitting groups. Provides more information than non-hierarchical as it identifies the main clusters and constituency relations relative to one another as well as their internal structures. It is a three stage procedure. The first step is the construction of a one-dimensional symmetric matrix of proximity. The second is the examination of the proximity matrix in order to determine whether or not a non-random structure actually exists in D_{FW}, D_{bigram} and

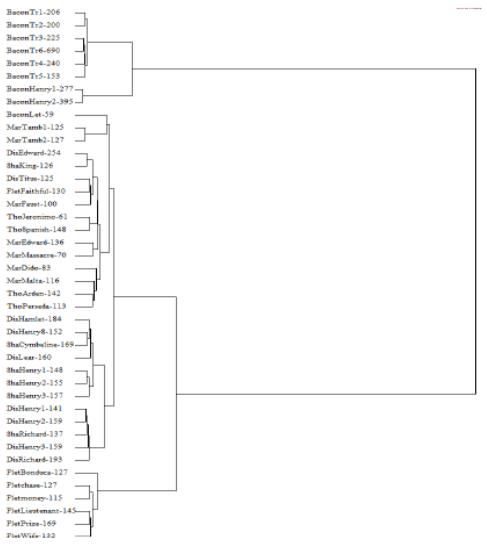


Fig. 7: A linear hierarchical cluster analysis of a 42 × 60 dimensionality-reduced and length-normalized D_{FW} using Squared Euclidean distance and an increase in sum of squares

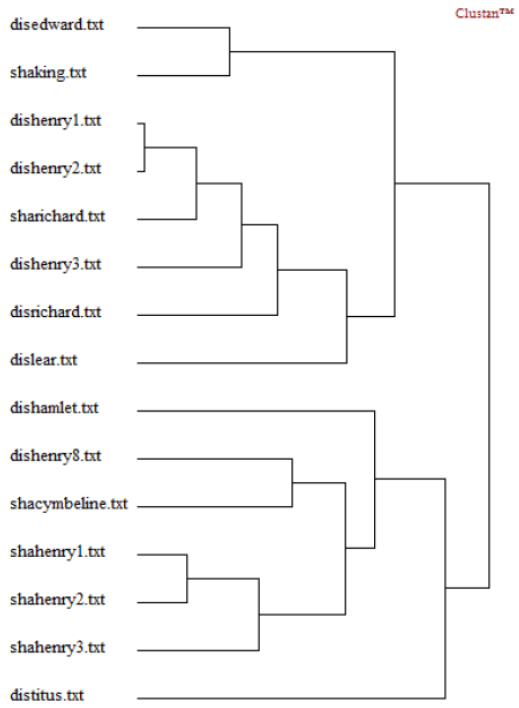


Fig. 8: The hierarchical cluster analysis of Shakespeare D_{FW} using Product-Moment correlation and Mean Proximity

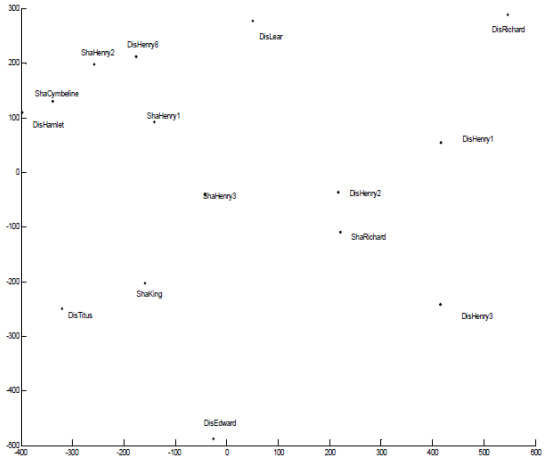


Fig. 9: PCA of Shakespeare D_{FW}.

$D_{trigram}$. The last step is the generation of clusters based on the proximity matrix by selecting the best clustering method results about the constituency structure of the forty two text matrix row vectors. This method was Mean Proximity: the averages of the within-cluster correlations/distances were maximized for all cluster comparisons.The result of the hierarchical analysis of D_{FW} for Shakespeare's Function Words, Word Bi-Grams, and Character Triple-Grams can be seen in Figure 8 where fifteen texts are grouped into eleven clusters according to the similarities of frequency vector profiles. It is remarkable that Disputed—History of Henry VI, Part I (dishenry1), Disputed—History of Henry VI, Part II (dishenry2) and Disputed—History of Henry VI, Part III (dishenry3) are not clustered with Shakespeare's works in and only one (sub)cluster and that disputed plays The Tragedy of Hamlet, Prince of Denmark (DisHamlet), The Tragedy of King Lear (DisLear), and Titus Andronicus DisTitus are well separated from the other Shakespeare's works.

Principal Component Analysis (PCA) is a non hierarchical linear method that re-described the 42 texts in terms of a number of variables, such that most of the variability in the original variables was retained. PCA is a four stage procedure which consists of the construction of a symmetric proximity matrix for distances among vectors, the construction of an orthogonal basis for the covariance matrix, the selection of dimensions and the projection into m-dimensional space. Figure 9 shows similar results with hierarchical clustering.

SOM-U: The unified distance matrix or U-matrix is a representation of Self-Organizing Map (SOM) that calculates the nonlinear distances between data vectors and is presented with different colorings. SOM U-Matrix generates graphical representations in two-dimensional space.The analysis was a two-stage process. The first is the training of SOM by loading all the vectors comprising D_{FW}, D_{bigram} and $D_{trigram}$ into the input space and the second is the generation of the two-dimensional representation of the D_{FW}, D_{bigram} and $D_{trigram}$ on the map. The U-matrix representation of SOM output used the Euclidean distances between connection vectors to find cluster boundaries stored in a new matrix U_{DFW}, $U_{Dbigram}$ and $U_{Dtrigram}$. Dark coloring between the vectors corresponds to a large distance and, thus, represents a gap between the values in the input space. A light coloring is the boundaries between clusters or the vectors, indicating that the vectors are close to each other in the input space. Light areas represent clusters and dark areas cluster separators. In

Figure 10 the light pink-vintage areas shown in the SOM map are the regions where the texts are topologically close, that is where they cluster, and the dark pink-vintage are where they topologically far apart. Again the results are similar as mentioned before.

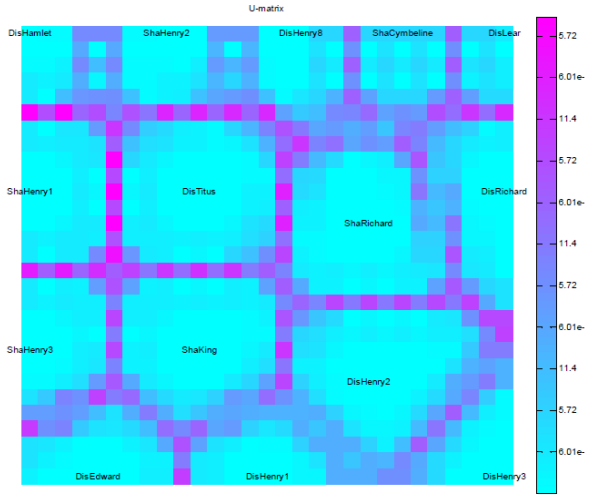


Fig. 10: SOM U-matrix of Shakespeare D_{FW}.

Voronoi diagram is a partition of a plane into regions close to each of a given set of objects. These regions are called cells, which surround each vector. The partition of a manifold surface into areas surrounding vectors is a tessellation. The analysis was in a three-stage process. The first step was the construction of a 2-dimensional Voronoi plot for a set of vectors in D_{FW}, D_{bigram} and $D_{trigram}$. The second was the construction of Delaunay Triangulation (Voronoi map) on the same 2-dimensional plot. The third step was the computation of the Voronoi map to obtain a 2-dimensional topology of the Voronoi map for the set of vectors in D_{FW}, D_{bigram} and $D_{trigram}$. The results for Shakespeare's D_{FW} are shown in Figure 11 which suggests that Shakespeare is not the author of all the works traditionally attributed to him.

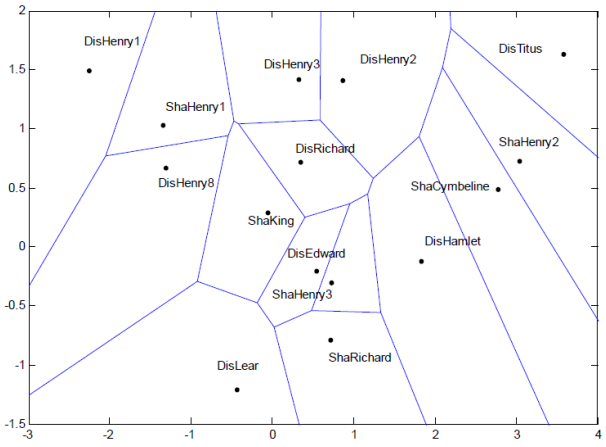


Fig. 11: Voronoi map of Shakespeare D_{FW}.

Some analysis of the frequencies of usage of Shakespeare and other Authors' function words, bi-grams and tri-grams took place with the same techniques proposed before. The fuction words results (Figure 12) showed similarity between Shakespeare's and disputed plays: Disputed—King Edward (DisEdward) and William Shakespeare—History of King John (ShaKing); Disputed—History of Henry VIII (DisHenry8), Disputed—The Tragedy of King Lear (DisLear) and William Shakespeare—Cymbeline, King of Britain (ShaCymbeline); Disputed—History of Henry VI, Part I (DisHenry1), Disputed—History of Henry VI, Part II (DisHenry2), and William Shakespeare—History of Richard II (ShaRichard) indicating that Shakespeare could be their author. Close similarity also showed between Marlowe's works and two of the disputed works: Disputed—History of Henry VI, Part III (DisHenry3) and Christopher Marlowe—Tamburlaine, Part 1 (MarTamb1); Christopher Marlowe—Tamburlaine, Part 2 (MarTamb2), Christopher Marlowe—Edward II (MarEdward), and Christopher Marlowe—Massacre at Paris (MarMassacre); and between Fletcher's The Faithful Shepherdess (FletFaithful) and Disputed—Titus Andronicus (DisTitus) on the other. However it is worth noticing that that Disputed—The Tragedy of Hamlet, Prince of Denmark (DisHamlet) and Disputed—Richard III (DisRichard) are not placed nearer to any of Shakespeare's works.

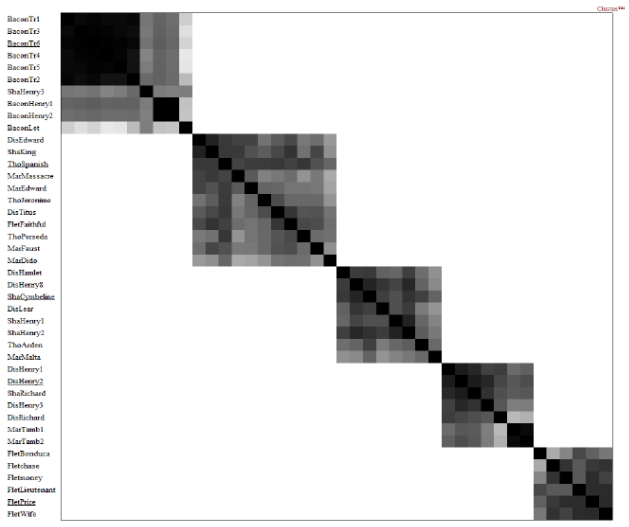


Fig. 12: An assessment of clustering tendency test for D_{FW}.

However there were some inconsistencies in the hierarchical and non-hierarchical linear and non-linear analyses due to the type of data structure each method captures. For example, in hierarchical analysis DisRichard is sub-clustered on its own, but is close to MarTamb1 while in PCA DisRichard is placed close to MarDido, in SOM to MarEdward and MarMassacre, and in Voronoi DisRichard is placed on its own. The significant note that one should keep in this analysis is that Shakespeare's plays are not clustered with the nine disputed ones, in particular DisHenry3 and DisTitus, but with another author or a collaborator. The bi-

gram analysis couldn't show good similarities in the methods apart from the clustering of DisLear with Fletcher's works. Therefore the only result suggested is that DisLear is not excluded from the possibility of having another author's or a collaborator's style. In the tri-gram analysis the results are close to the heatmap presented before: some of the disputed texts are not close enough to Shakespeare's works, but are close to the works by the other authors, in particular Marlowe and Fletcher. The general conclusions are that Shakespeare did not write the disputed plays traditionally attributed to him but it was not possible to indicate the rightful author or collaborator.

IV. Future work

In generally theatrical plays there are some parameters that prevent the accuracy of the results being extremely high. One of the most important reason is the "hidden text". The characters most of the time never means what they say. This in fact is the role of an actor: to find what the characters thoughts could be and what he/she truly wanted to say. If our topic and dataset was Ancient Greek plays our job would be easy and pretty accurate. This is because in Ancient Greece the actors used to wear masks during the performance, therefore the playwrights had to make sure that the audience would be able to acknowledge the sentiments of the character. To accomplish that the characters expressed and explained exactly how they were feeling during the different actions that took place. But this was not the case in Elizabethan theatre. Shakespeare's plays have plenty of irony, deceit, untruthful heroes and betrayal which a computer is having a hard time detecting. Moreover, Shakespeare's characters are thought to be multifaceted and have a variety of sentiments.

In [7] the author tried to approach the sentiments using a lexicon . The first experiment was to determine whether sentiment could be able to distinguish Shakespeare's comedies from tragedies. The AFINN [8] word list for sentiment analysis was chosen which contains 2477 English words, labeling each with a valence, an integer between -5 (most negative words) and +5 (most positive) without modifying the modern words to Shakespearean dialect. The method is to calculate average word valence (AWV) by summing the valence values for all words in the play and then dividing by the number of the play's words in AFINN. In Figure 13 there are 2 important point than needs to be emphasized. Titus Andronicus has the lowest AWV of the tragedies (0.40), which is a success since the play is widely considered to be Shakespeare's bloodiest and most violent work. The comedy of errors although it was considered a comedy it has $AWN = -0.03$. This is because the characters' lives seem to be in ruin throughout most of this play, The Comedy of Errors thinks it is a tragedy, but the timing resolves the action in blissfully comedic ways.

The next experiment consists of knowlege-based sentiment analysis to track the emotional trajectories of interpersonal relationships rather than of a whole text or an isolated

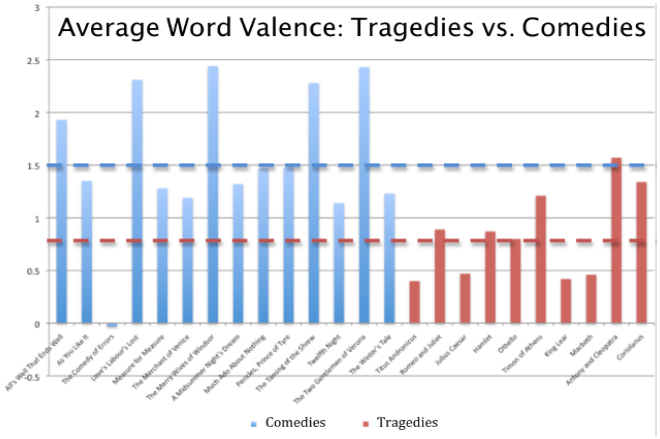


Fig. 13: For each play, the valence of each word was summed and then divided by the number of words in both the word list (AFINN) and the respective play. The result is the average word valence (AWV), which seems to discriminate well between tragedies and comedies. The blue line marks the average AWV for comedies and the red line represents the same value for tragedies.

character. As supported in [9] a character-to-character sentiment was computed by summing the valence values over each instance of continuous speech and then assumed that sentiment was directed towards the character that spoke immediately before the current speaker. Of course this is not exactly correct as characters come and go in various scenes and the characters talking can change through the play. The results however was pretty satisfying as face-to-face dialogue produced a strong enough signal to generate sentiment rankings.

In Table XI it can be seen that Claudius ($sum = -27$) correctly collect the most negative sentiment from Hamlet and this is because he thinks that he murder his father. On the contrary, Gertrude ($sum = 24$) collects a very positive sentiment which is absurd because Hamlet thought she was involved in the murder of King Hamlet.

Character	Hamlet's Sentiment Valence Sum
Guildenstern	31
Polonius	25
Gertrude	24
Horatio	12
Ghost	8
Marcellus	7
Osric	7
Bernardo	2
Laertes	-1
Ophelia	-5
Rosencrantz	-12
Claudius	-27

TABLE XI: The characters in Hamlet are ranked by Hamlet's sentiment towards them.

In order to clarify the relationship between mother and son in Figure 14 it can be seen that there is an extremely change of emotions in line 2,250 which corresponds to Act III scene IV. In this scene Hamlet confronts his mother and understands that his mother was not involved in the

murdering so his sentiment jump from -1 to 22. On the contrary Gertrude realises that Hamlet killed an innocent man and is becoming mad. Therefore her sentiment changes negatively from 1 to -19.

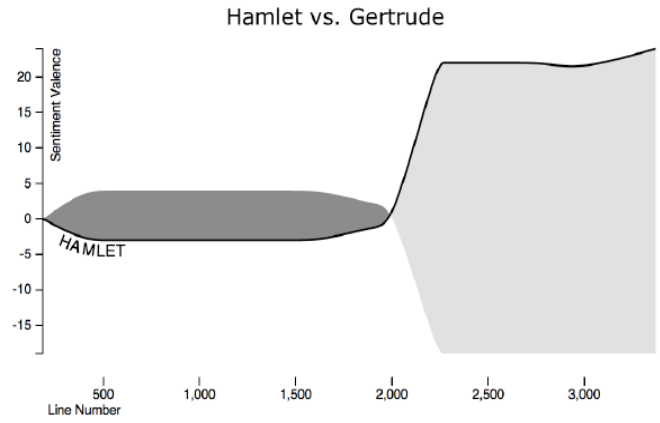


Fig. 14: The above chart tracks how Gertrude's and Hamlet's sentiment towards one another changes over the course of the play. Hamlet's sentiment for Gertrude is denoted by the black line, and Gertrude's for Hamlet is marked by the opposite boundary of the dark/light gray area.

It would be interesting to confirm these results using machine learning techniques as proposed in [24]. Another project could be to find the thematics that appear in his plays and where a play contributes to the writing of another play. Moreover, it is considered that Shakespeare was inspired by Ancient Greek plays. Hence, the similarity between the Shakespeare's plays and Ancient Greek plays could provide information on the influence the latter had in the author's thinking and imagination.

V. Conclusions

The true authorship of Shakespeare's plays has troubled scholars. Rhythmic types, function words or n-grams provide a significant help in detecting a writing style. Then machine learning techniques are used applying these styles to find whether a play was a creation of collaboration or it was written by a singular author. In this survey, the machine learning methods presented were Multilayer Perceptron, Linear discriminant analysis (LDA), Decision Trees (DTs), SVM and clustering. It is generally believed that authorship attribution is an issue that provide uncertain results. In Table XII the different outcomes are presented that derived from entire plays. It is noticeable that with different techniques and functions the results diverges. In Henry VI, part 3 we could say that opinions coincide. Merriam who used MLP together with function words supports that it was written by Marlowe, while Aljumily who used clustering together with function words, bi-grams and n-grams believes that it is not written by Shakespeare. However, other from that example the two scholars seem to disagree as Merriam states that Hamlet, King Lear and Titus Andronicus and Edward III were written by Shakespeare, while Aljumily states the opposite. Again no cohesion is visible in Henry VIII as Merriam suggests that it was written by Shakespeare, Aljumily believed that Shakespeare is not the author, while Plechac is in the middle suggesting that it is a collaboration between Shakespeare and Fletcher.

Entire play	Merriam	Plechac	Boyd	Aljumily
Henry VIII	S	S,F	-	NS
Double Falsehood	F	-	S	-
Two Noble Kinsmen	S*	-	-	-
Edward II	S**	-	-	NS
Titus Andronicus	S	-	-	NS
Henry VI, Part 3	M	-	-	NS
Hamlet	S	-	-	NS
King Lear	S	-	-	NS
London Prodigal	F	-	-	-

*TABLE XII: Authorship Attribution results based on entire plays. S denotes attribution to Shakespeare, F to Fletcher, M to Marlowe and NS not to Shakespeare. Where * is shown means with a little contribution of Fletcher, ** of Marlowe and *** of Shakespeare.*

Table XIII shows the results based on plays that were broken into acts. We can observe that there is more consistency in the outcomes as at least Double Falsehood is thought to be a collaboration between Shakespeare and Fletcher with the latter having equal or less contribution in the play. It is important to point out that we can not have a final conclusion on Shakespeare plays as there is a collision in the results of the different scholars.

Play by acts	Merriam	Plechac	Boyd
Henry VIII	S	S,F	-
Double Falsehood	S,F	-	S*
Two Noble Kinsmen	S*	-	-
Edward II	S**	-	-
London Prodigal	F***	-	-

*TABLE XIII: Authorship Attribution results based on plays broken in acts. S denotes attribution to Shakespeare, F to Fletcher, M to Marlowe and NS not to Shakespeare. Where * is shown means with a little contribution of Fletcher, ** of Marlowe and *** of Shakespeare.*

REFERENCES

[1] Plecháč, Petr. "Relative contributions of Shakespeare and Fletcher in Henry VIII: An Analysis Based on Most Frequent Words and Most Frequent Rhythmic Patterns." arXiv preprint arXiv:1911.05652 (2019).

[2] Spedding, J, Who wrote Shakespeare's Henry VIII," The Gentleman's Magazine,pp. 115123, 1850.

[3] Fleay, F G, Mr. Boyle's theory as to Henry VIII"," Athenum, vol. 2994, p. 355,1885.

[4] Oliphant, E H C, The works of Beaumont and Fletcher," Englische Studien,vol. 15, pp. 321360, 1891.

[5] Eder, Maciej, "Rolling stylometry," Digital Scholarship in the Humanities, vol. 31, no. 3, pp. 457–469, 2016

[6] Platt, J, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," Advances in Large Margin Classifiers, vol. 10, no. 3, pp. 61–74, 1999.

[7] Nalisnick, Eric T., and Henry S. Baird. "Extracting sentiment networks from Shakespeare's plays." 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013.

[8] F. A° . Nielsen, "Afinn," Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, March 2011. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?6010

[9] E. Nalisnick and H. Baird, "Character-to-character sentiment analysis in shakespeare's plays," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Short Papers. Association for Computational Linguistics, 2013.

[10] Matthews, Robert AJ, and Thomas VN Merriam. "Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher." Literary and Linguistic computing 8.4 (1993): 203-209.

[11] Horton, T. B. (1987). Doctoral thesis, University of Edinburgh.

[12] Merriam, T. V. N. (1992). Doctoral thesis, University of London.

[13] Merriam, Thomas VN, and Robert AJ Matthews. "Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe." Literary and Linguistic Computing 9.1 (1994): 1-6.

[14] Metz, G. H. (ed.) (1989). Sources of Four Plays Ascribed to Shakespeare. University of Missouri Press, Columbia.

[15] Hoy, C. (1956). The Shares of Fletcher and His Collaborators in the Beaumont and Fletcher Canon (VII), Studies in Bibliography, 15: 129-46.

[16] Merriam, T V. N. (1993). Marlowe's Hand in Edward III, Literary and Linguistic Computing, 8.2: 59-7

[17] Taylor, G. (1987). The Canon and Chronology of Shakespeare's Plays. In William Shakespeare. A Textual Companion Clarendon Press, Oxford

[18] Tucker Brooke, C. F (1912). The Authorship of the Second and Third Parts of 'King Henry VI'. Transactions of the Connecticut Academy of Arts and Sciences, 47: 145-211

[19] Robertson, J M. (1924). An Introduction to the Study of the Shakespeare Canon. Routledge, London

[20] Boyd, Ryan L., and James W. Pennebaker. "Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis." Psychological science 26.5 (2015): 570-582.

[21] Pennebaker, J. W., Booth, R. J., and Francis, M. E. (2007). Linguistic Inquiry and Word Count (LIWC2007): A computerized text analysis program [Computer software]. Austin, TX: LIWC.net

[22] Boyd, R. L. (2014b). RIOT Scan: Recursive Inspection of Text Scanner (Version 1.8.3) [Software].

[23] Aljumily, Refat. "Hierarchical and non-hierarchical linear and non-linear clustering methods to "Shakespeare Authorship Question"." Social Sciences 4.3 (2015): 758-799.

[24] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, 2016, pp. 628-632, doi: 10.1109/ICATCCT.2016.7912076.

[25] Subasi, Abdulhamit. Practical Machine Learning for Data Analysis Using Python. Academic Press, 2020.

VI. APPENDIX

A. Support Vector Machine

SVM can be used as a classification or regression method, maintaining in both cases the main concept that characterizes the algorithm which is to draw a hyperplane of maximum margin. In the classification case, the task is pretty simple: find the hyperplane that separates the classes and has maximum margin, which means maximum distance from the closest point of each class, which is where the support vectors are drawn. For new data, the algorithm decides the class by comparing its position relative to the support vectors.

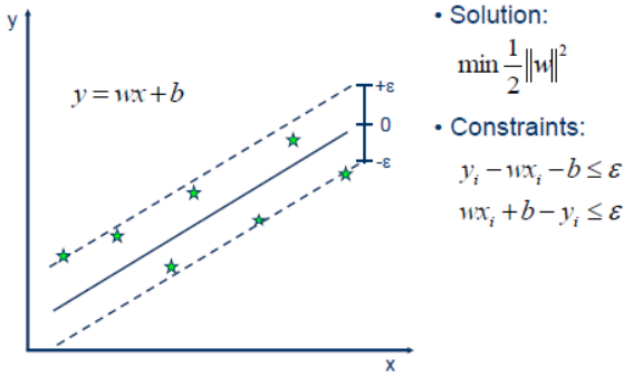


Fig. 15: Support Vector Machine

Finally, one can add a trade-off parameter to the objective function:

$$\min \ \frac{1}{2}||w||^2 + C\sum|\xi_i|,$$

where ξ denotes the deviation from the support vector, and tune the hyperparameter C to match the data's requirements and achieve higher performance. A higher C value means less tolerance for the points inside the margin in classification and outside the margin in regression.

B. LDA

As proposed in [25], LDA is a classifier used to find a linear combination of features, which separates two or more classes of data. The succeeding combination can be used as a linear classifier. In LDA, the classes are expected to be normally distributed also it can be utilized for both dimension reduction and classification. In a two-class dataset, the a priori probabilities for class 1 and class 2 are p_1 and p_2; the class means and overall mean are μ_1, μ_2, and μ; and the class variances are cov_1 and cov_2 respectively.

$$\mu = p_1 \times \mu_1 + p_2 \times \mu_2$$

Then, within-class and between-class scatters are used to represent the needed criteria for class separability. The scatter measures for a multiclass situation are calculated as:

$$S_w = \sum_{j=1}^{C} p_j x cov_j$$

where C refers to the number of classes and

$$cov_j = (x_j - \mu_j)(x_j - \mu_j)^T$$

Then, the aim is to find a discriminant plane to maximize the ratio of between-class to within-class scatters (variances):

$$J_{LDA} = \frac{wS_b w^T}{wS_w w^T}$$

C. Decision Trees

Decision trees are built by recursively splitting the data according to a criterion involving its features that minimizes an error function. In order to avoid correlation between trees, each tree is constructed using a random subsample of the original dataset and a random subset of all features as split criteria. They have high performance variance depending on the characteristics of the dataset and are susceptible to overfitting , thus need careful hyperparameter tuning. An example of a decision tree is shown in Figure 16.

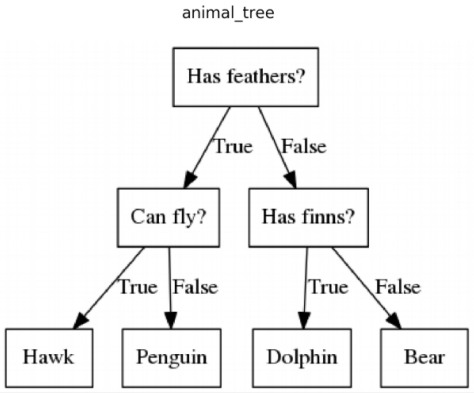


Fig. 16: Decision Tree.