Lecture Notes on Language and Literature (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/langl.2023.061319
ISSN 2523-5869 Vol. 6 Num. 13

Revolution of Shakespearean Plays' Genre Research: Exploring New Avenues through Machine Learning and Shapley Value Analysis

Chang Yuan[1,*]

[1]*Ocean University of China, Qingdao, China*
[*]*Corresponding author*

Keywords: Shakespeare, machine learning, Shapley value

Abstract: In the realm of literary research, the challenges of being confined to narrow niches and disconnected from broader contexts have been long-standing. In response, the integration of digital research methods into literary studies has emerged as a compelling area of exploration. Among the classic subjects of literary research, the classification of Shakespearean drama genres holds particular significance. In this paper, we present a case study focused on introducing a promising predictive and analytical method, which leverages Linear Discriminant Analysis (LDA) and the Shapley value. Our methodology begins by employing decision trees to reduce the dimensionality of textual data. Subsequently, a LDA based on Bayesian optimization algorithm is applied to predict the genres of texts. Finally, we utilize the Shapley value to analyze the important words within the texts and unveil their profound literary associations with respective genres. By adopting this approach, our research contributes to the widespread adoption and digital transformation of literary studies, thereby pioneering new avenues in Shakespearean drama research.

1. Introduction

Shakespeare's dramatic works stand as enduring classics in world literature. Renowned literary critic Bloom astutely posits that "The Shakespearean exuberance peoples a heterocosm of men and women, abounding with a secular blessing of more life on into a time without boundaries" [1]. This assertion elucidates the perennial relevance of Shakespeare's plays in literary criticism and research [2]. Even within the contemporary digital context, Shakespeare's dramas continue to emanate their literary charm, exemplifying an enduring and revitalized significance [3].

In academia, Shakespeare's dramatic works enjoy a vast foundation of research, with a particular focus on the language and genres of his plays [4]. Regarding language, Shakespeare's linguistic innovations manifest in various forms, such as word class conversion, blending of archaic and neologistic words, word coinages, metaphors, idioms, puns, parallelism, and the popularization of blank verse, all of which contribute to the eloquence exhibited by the characters in his works. Currently, research on Shakespearean dramatic language predominantly employs qualitative methods, such as literature reviews and close textual analysis [5]. While these approaches have yielded valuable insights, they do encounter certain limitations, such as low efficiency and an inability to reveal

quantitative relationships, leading to a relatively insular and niche landscape within literary studies.

The classification methods of dramatic genres have not remained uniformly consistent from ancient times to the present. In ancient Greece, drama was categorized into two main forms: comedy and tragedy. Aristotle, in his work Poetics, compared various dramatic genres and differentiated comedy and tragedy based on the emotional tone, the status of the protagonist, the type of language used, and the cathartic effect [6]. During the Renaissance period, theatrical works, especially tragedies, were also classified based on plot elements or symbols [7]. While Shakespeare did not explicitly distinguish his theatrical genres, contemporary scholars classify his plays into four categories: comedy, tragedy, history, and tragicomedy (also known as romance). All of the aforementioned classification methods in the past heavily relied on manual reading by experts and scholars, leading to lower efficiency and subjectivity in the process. Therefore, in the present day, employing novel methods to extract textual features from dramas and determining their genre attribution represent valuable and promising research topics.

In the current era of information technology, the digital transformation of literary studies has shown promising progress. The emergence of corpus stylistics based on electronic datasets has injected new vitality into literary research. For instance, Culpeper [8] employed database technology to conduct a semantic categorization study of key words in character dialogues in Romeo and Juliet. Similarly, Vickers [9] utilized database technology to construct a corpus and employed linguistic methods to identify Shakespeare's contributions to The Spanish Tragedy. Furthermore, in Shakespearean literary research, several foundational mathematical methods have been applied. Whissell [9] employed language emotion dictionaries and function methods to predict the genres of 23 plays. Their t-test analysis revealed that tragedies significantly applied more Active words. Papp-Zipernovszky et al. employed qualitative text analysis and linear regression to study Shakespeare's sonnets, uncovering that certain foreground types had the highest significance in readers' emotional evaluations [10]. However, most of the digital methods applied in Shakespearean research remain relatively superficial, failing to deeply unveil the quantitative relationships within Shakespearean literature, thus impeding the sustainable development of Shakespearean studies.

Machine learning, a popular and advanced mathematical model of the present, aims to explore nonlinear quantitative relationships among data features [11]. Initially applied in natural language processing, machine learning has achieved remarkable breakthroughs. It has demonstrated high efficiency in text analysis and has been successfully applied in various domains, such as social network sentiment detection [12], intelligent speech dialogue [13], and sentiment analysis of literary works [14]. Moreover, researchers have further extended the use of machine learning methods in the field of literary studies, primarily to explore authorship attribution [15], literary metaphor detection [16], translation [17], and eye-tracking analysis [18]. Machine learning has significantly contributed to enabling scholars to draw intriguing conclusions in literary research. In the field of Shakespearean literature, machine learning methods have also found application. Plechac [19] utilized machine learning techniques and lexical analysis to determine authorship contributions in the poetic drama Henry VIII. Liu et al. [20] employed machine learning methods to mimic Shakespeare's writing style, promoting personalized cognitive learning experiences for learners. Moscato et al. [21] used machine learning techniques to study the mapping relationship between word frequencies and the premiere years of Shakespeare's plays, consequently identifying crucial keywords that determine a play's premiere year. In summary, machine learning methods hold great potential for significant advancements in literary research, particularly in the domain of Shakespearean studies.

To address the niche and insular issues in literary research mentioned above and to propel further development of Shakespearean dramatic genre studies, this study aims to make literary research more accessible and digitized, fostering the sustainable growth of Shakespearean studies. In this paper, we utilized a database of Shakespearean plays and employed machine learning methods to investigate

the intrinsic connections, particularly quantitative relationships, between text features and dramatic genres. By exploring the relationships between vocabulary characteristics and themes of plays within different genres, we have uncovered new perspectives for the study of Shakespearean dramas.

In our study, we employed LDA, a machine learning model, to construct an accurate model suitable for high-dimensional small sample datasets [22]. The features used in the model were frequency counts of carefully selected important words, which helped to restrict the feature dimensionality. The model's task was to predict the literary genre based on word frequencies. To address the challenge of hyperparameter tuning in LDA, we selected the Bayesian optimization algorithm as the optimization method. This choice facilitated effective parameter settings in the model. Furthermore, to gain deeper insights into the intrinsic connections between text and literary genres, we utilized the Shapley Value method to compute the relative importance of each significant word. This method, widely used in cooperative game theory to determine the contribution of each player, allowed us to reveal essential imagery and themes implicitly present in different genres of Shakespearean dramas.

Our contributions can be summarized as follows:

Introduction of a decision-tree-based feature selection method: We utilized a decision tree to perform multiple executions and selected 79 features as the feature subset.

Comparative evaluation of 10 state-of-the-art classification methods: We conducted extensive testing on both the training and testing datasets, demonstrating that LDA outperformed other methods in the domain of word frequency analysis, making it the most effective classification model.

Pioneering application of Shapley Value in Shakespearean drama research: For the first time, we applied the Shapley Value method to the study of Shakespearean dramas, using it to unveil the intrinsic connections between dramatic genres and text features by relating literary themes and imagery.

Our paper is structured into five main sections: The first section of our paper provides an introduction to the research background and the advancements in the digitization of Shakespearean drama. The second section elaborates on the dataset and model utilized in our study. In the third section, we present a preliminary presentation of our experimental results. The fourth section delves into a comprehensive analysis utilizing the Shapley value method to examine the intrinsic connections between text and genre in the classification of Shakespearean drama. The final section concludes our work, summarizing the key findings and contributions of our research.

2. Data and methods

2.1 Description of the dataset

In this study, we utilized a dataset named '181 early modern English plays: Transcriptions of early editions in TEI encoding' [23]. The dataset was obtained from the 'UC Irvine Machine Learning Repository', a publicly available database. It comprises 181 samples of Shakespearean plays, each containing information such as performance date, publication date, genre, and frequency of 51,256 significant words. For feature representation, we selected the word frequency of each word to reflect the intrinsic textual characteristics of different plays, and we considered the genre of each text as the target variable for prediction.

The dataset categorized the 181 Shakespearean plays into 20 different classes of texts. Some categories, such as 'Comedy,' consist of 70 samples, while others like 'Biblical Moral,' only have one sample, leading to severe class imbalance. To address this issue before the modeling stage, we conducted data preprocessing on the genre labels of the plays. Drawing inspiration from the work of Moscato et al. [21], we merged all categories into five major classes: ‘Comedy’, ‘History’, ‘Tragedy’, ‘Tragicomedy’, and ‘Other genres’. The specific merging criteria are presented in Table 1, and the

number of samples in each class is depicted in Figure 1.

Table 1: Criteria for Category Merging

Major classes	Sample Categories
Comedy	Classical Legend (Comedy)
	Comedy
	Domestic Comedy
	Romantic Comedy
History	Classical History
	Foreign History
	History
	Legendary History
	Pseudo-History
Tragedy	Tragedy
Tragicomedy	Tragicomedy
Other genres	Romance
	Biblical Moral
	Burlesque Romance
	Moral
	Pastoral
	Heroical Romance
	Domestic Drama
	Biblical Moral
	Burlesque Romance

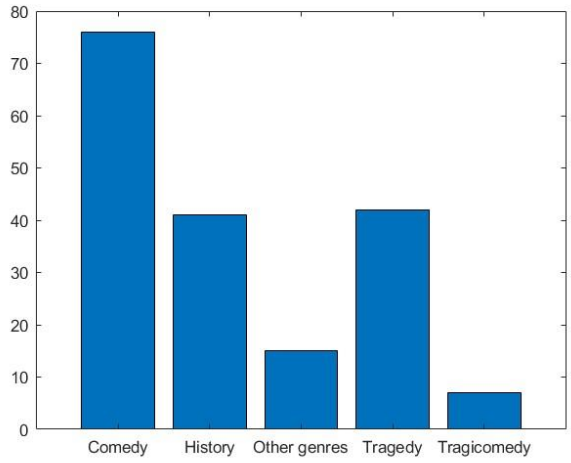

Figure 1: Distribution of Play Samples among Genre Categories

2.2 Selection of features

Given the dataset's high dimensionality with a substantial number of features and a limited number of samples, it becomes necessary to employ feature selection methods to reduce the feature space dimensionality and facilitate improved classification by the model [23]. Our objective is to identify a vector that represents the significance of each feature concerning classification and obtain a feature subset with the smallest possible size.

The selection of the feature subset for model prediction was determined using a decision tree. The decision tree, a renowned classification model, comprises a set of 'if-else' rules that can identify the significance of different features during classification [24]. By assessing the average change in node impurity before and after splitting based on a particular feature, the decision tree quantifies the importance of that feature. In this study, we utilized the commonly employed Gini index as a measure of node impurity, which is computed according to Equation 1. $p(j)$ represents the probability of class j at node i, and when a node i contains only one class, the Gini index is 0, signifying a state of purity. Consequently, the Gini index serves as an effective measure of impurity.

$$GiNi(i) = 1 - \sum_{j} p^{2}(j) \tag{1}$$

To ensure stability, we repetitively constructed 100 decision tree models on 80% random subsets. We selected word frequencies as features to be included in the subset if they had non-zero importance scores in any of the models. The specific features chosen are listed in Table 2. Overall, the application of the decision tree method for dimensionality reduction theoretically led to a substantial decrease in predictive difficulty for the dataset. Nevertheless, compared to the 181 samples, the dataset still retained a relatively large number of features.

Table 2: 79 selected important words and their frequencies

Word	Pct.	Word	Pct.	Word	Pct.	Word	Pct.
'the'	10	'lord'	10	'place'	10	'teach'	10
'and'	10	'well'	10	'we_royalplural_'	10	'ass'	30
'i'	30	'am'	10	'marry'	20	'sacred'	10
'a'	10	'there'	10	'power'	10	'richard'	10
'my'	10	'when'	10	'earth'	10	'kingdom'	10
'is'	10	'must'	10	'sword'	10	'spare'	10
'to_infinitive_'	10	'death'	30	'fool'	10	'vows'	10
'it'	10	'heart'	10	'fire'	10	'thence'	10
'to_preposition_'	20	'nor'	10	'same'	10	'moved'	10
'be'	10	'sweet'	10	'lives'	10	'eternal'	10
'but'	10	'made'	10	'majesty'	30	'girl'	10
'for_preposition_'	10	'TRUE'	10	'bed'	10	'flowers'	10
'as'	10	'done'	10	'fit'	10	'cruelty'	10
'that_relative_'	10	'still'	10	'lies'	10	'furious'	10
'shall'	10	'blood'	20	'england'	50	'sung'	10
'by_preposition_'	10	'live'	10	'english'	10	'nearest'	10
'that_demonstrative_'	20	'day'	10	'eat'	10	'roots'	10
'sir'	20	'faith'	10	'london'	10	'chances'	10
'thee'	10	'being'	10	'fools'	10	'befallen'	10
'they'	10	'therefore'	10	'wits'	10		

2.3 Bayesian optimization-based LDA

LDA is a classical machine learning method that aims to achieve efficient sample classification by finding an optimal projection direction, mapping data from a high-dimensional space to a lower-dimensional space [25]. The optimal projection direction in LDA is defined as the vector that maximizes the distance between classes while minimizing the within-class variance, as illustrated in

Algorithm 1. N_i represents the number of samples in class i and C denotes the total number of classes:

Algorithm 1
Step 1: Calculate Intra-class mean vector u_i and population mean vector u
Step 2: Calculate within-class scatter matrix $S_W = \sum_{i=1}^{C}\sum_{j=1}^{N_i}(x_j - u_i)*(x_j - u_i)^T$
Step 3: Calculate between-class scatter matrix $S_B = \sum_{i=1}^{C} N_i(u_i - u)*(u_i - u)^T$
Step 4: Solving the optimization problem $J(\mathrm{w}) = \frac{\lvert \mathrm{J_B} \rvert}{\lvert \mathrm{J_w} \rvert} = \frac{\lvert w^{\mathrm{T}} \mathrm{S_B} w \rvert}{\lvert w^{\mathrm{T}} \mathrm{s_W} w \rvert}$
Step 5: Calculate projection vector $w = eig(S_W{}^{-1} * S_B)$
Step 6: Classify samples $y = x * w$

In practical applications, this method often faces challenges related to the difficulty of estimating within-class scatter matrices and their potential non-invertibility issues [26]. To enhance the estimability of the within-class scatter matrices, we adopted an approach inspired by the 'fitcdisr' function algorithm in Matlab 2023a, assuming that all classes share the same within-class scatter matrix. Furthermore, to address potential issues of non-invertibility, we employed regularization techniques to estimate the within-class scatter matrix, as outlined in Equation 2. In this equation, γ represents the regularization coefficient, and $\hat{\Sigma}$ denotes the covariance matrix of features.

$$\hat{\Sigma}_{\gamma} = (1-\gamma)\hat{\Sigma} + \gamma \mathrm{diag}(\hat{\Sigma}) \tag{2}$$

Bayesian optimization algorithm is an iterative optimization method used for optimizing black-box functions [27]. Its objective is to maximize or minimize the value of the objective function by selecting the next sample point at each iteration, without explicitly inferring the mathematical expression of the objective function. Due to considering uncertainty at each iteration, it can discover global optima even with a limited number of sample points, thereby avoiding being trapped in local optima [28]. Given its capability to handle uncertainty and efficiently explore the search space, we employ the Bayesian optimization algorithm to optimize the regularization coefficient of matrices in LDA. By doing so, we aim to enhance the effectiveness of LDA by finding an optimal regularization coefficient that leads to improved classification accuracy.

3. Experimental results

3.1 Experimental Model Setup

All of our experiments were conducted on a computer equipped with an Intel(R) Core(TM) i5-8300H CPU, 16.0 GB of RAM, running on a 64-bit Windows 11 operating system. The experimental environment was set up using Matlab r2023a.

To assess the performance of the machine learning algorithms, we employed a dataset consisting of 181 Shakespearean plays. We utilized 79 important word frequencies as features for genre classification, which exhibited nonzero importance in the classification of play genres. These features were used to predict the genres corresponding to the texts of the plays, with the aim of advancing the digital research of Shakespearean play genres.

To ensure reproducibility, we adopted 12 common machine learning classification methods, all implemented in Matlab R2023a. However, some models within Matlab's built-in functions do not

support multi-class classification. As a result, for these models, we applied Error Correcting Output Codes (ECOC) for adaptation. The dataset was randomly split into an 80:20 ratio, where 80% was used for model training and 20% for testing the predictive performance of each model. Each model underwent 100 repetitions. The classification methods used are presented in Table 3.

Table 3: All models we used

Model	Function in Matlab	hypyparameter settings
LDA(L-DIS)	fitcdisr	OptimizeHyperparameters=auto
Kernel Regression(K-REG)	fitcecoc	OptimizeHyperparameters=auto Learners=kernel
K-Nearest Neighbor(KNN)	fitcknn	OptimizeHyperparameters=auto
Linear Regression(L-REG)	fitcecoc	OptimizeHyperparameters=auto Learners=linear
Support Vector Machine(SVM)	fitcecoc	OptimizeHyperparameters=auto Learners=linear
Decision Tree(TREE)	fitctree	OptimizeHyperparameters=auto
Bagging-Discriminate(B-DIS)	fitcensemble	OptimizeHyperparameters=auto Method=Bag Learners=discriminant
Subspace-KNN(S-KNN)	fitcensemble	OptimizeHyperparameters=auto Method=Subspace Learners=knn
Random Forest(RF)	fitcensemble	OptimizeHyperparameters=auto Method=Bag Learners=tree
Adaboost-Discriminate (A-DIS)	fitcensemble	OptimizeHyperparameters=auto Method=AdaBoostM2 Learners=discriminant
Adaboost-Decision Tree (A-TREE)	fitcensemble	OptimizeHyperparameters=auto Method=AdaBoostM2 Learners=tree
BP Neural Network(BPNET)	fitcnet	OptimizeHyperparameters=auto

3.2 Results of experiments

Accuracy is utilized as the evaluation metric for comparing the performance of regression methods. This metric assesses the model's performance by calculating the proportion of correctly predicted instances relative to the total number of instances. The specific formula for computing accuracy, as represented by Equation 3, is as follows:

$$Accuracy = N_{correct} / N_{total} \quad (3)$$

Table 4 reports the descriptive statistics of the classifiers' accuracy in 100 runs. It can be observed that B-DIS and A-DIS achieved the highest average accuracy of 1.00 on the training data. L-DIS, L-REG, and BPNET also achieved close accuracies on the training set. K-REG exhibited an average accuracy above 0.85, ranking as the third-best model in terms of average accuracy on the training set. Additionally, K-REG, B-DIS, and A-DIS emerged as the models with the best median accuracy,

reaching 100% accuracy on the training set. Regarding the volatility of prediction accuracy, A-DIS demonstrated the most stable results, with a standard deviation of 0.00, implying it achieved 100% accuracy consistently across all predictions. In contrast, SVM and S-KNN exhibited higher fluctuations. Overall, considering both the magnitude and stability of accuracy, B-DIS and A-DIS performed the best on the training set. This suggests that the DIS method is effective in capturing the inherent relationship between text features and categories in text datasets with a large number of features and limited samples. Furthermore, the ensemble learning approach has further enhanced the accuracy of these models.

On the testing set, L-DIS achieved the best average and median accuracy, with BPNET as the second-best model. Remarkably, L-DIS exhibited outstanding performance in prediction volatility on the testing set, with a standard deviation of only 0.01. The second-best model in prediction volatility on the testing set was TREE. It is notable that K-REG, which performed well on the training set, showed poorer performance on the testing set, indicating the potential issue of overfitting for kernel methods when handling high-dimensional data. This also highlights the advantage of the DIS model in handling datasets with high-dimensional features and limited sample sizes, effectively avoiding overfitting and fully exploring the inherent relationship between word frequencies and text genres. Moreover, ensemble learning increased the risk of overfitting, as both Bagging and AdaBoost methods led to noticeable overfitting, thereby limiting the generalization ability of the DIS model. Additionally, the Subspace method constrained the prediction accuracy of KNN.

Table 4: Descriptive statistics of the model prediction accuracy

Model	Training set accuracy			Testing set accuracy		
	Mean	Median	Std	Mean	Median	Std
L-DIS	0.93	0.94	0.02	0.79	0.78	0.01
K-REG	0.87	1.00	0.23	0.44	0.41	0.08
KNN	0.77	0.76	0.07	0.68	0.68	0.04
L-REG	0.96	0.99	0.07	0.62	0.62	0.04
SVM	0.70	0.81	0.31	0.52	0.57	0.18
TREE	0.71	0.72	0.03	0.65	0.65	0.02
B-DIS	1.00	1.00	0.00	0.61	0.59	0.05
S-KNN	0.61	0.51	0.22	0.43	0.41	0.07
RF	0.75	0.77	0.19	0.64	0.68	0.11
A-DIS	1.00	1.00	0.00	0.62	0.62	0.01
A-TREE	0.77	0.79	0.19	0.64	0.68	0.11
BPNET	0.96	1.00	0.07	0.73	0.76	0.05

In addition to Table 4, we also presented box plots for the training and testing set classification accuracy obtained from the 100 runs in Figures 2 and 3. In the training set box plot, SVM exhibited the widest distribution range of prediction accuracy, followed by S-KNN, corresponding to their characteristic of high prediction volatility. In the testing set box plot, SVM also demonstrated the largest distribution range of accuracy, confirming the significant prediction volatility of the SVM model. On the other hand, the prediction intervals of other models were relatively close. In summary, the descriptive statistics and box plots of prediction accuracy provide a rough representation of the model's predictive performance. However, they are not detailed enough to provide a comprehensive understanding. Therefore, we proceeded with further performance comparisons to obtain a more refined evaluation of the models.

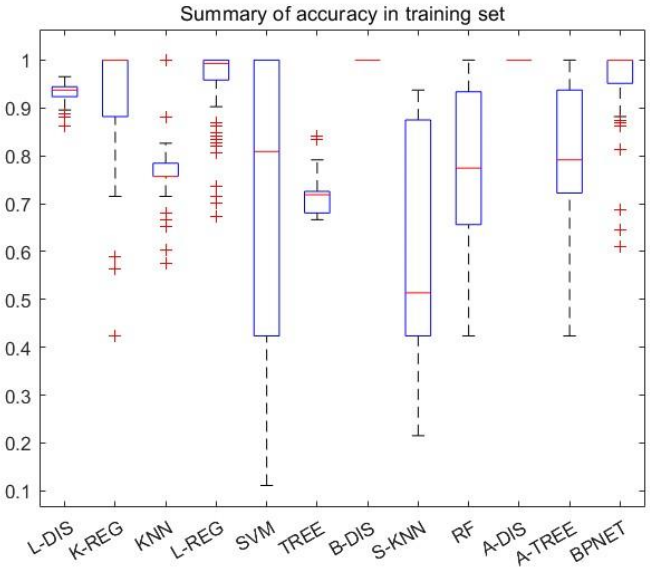

Figure 2: Prediction results of the 100 repeated experiments on the training set

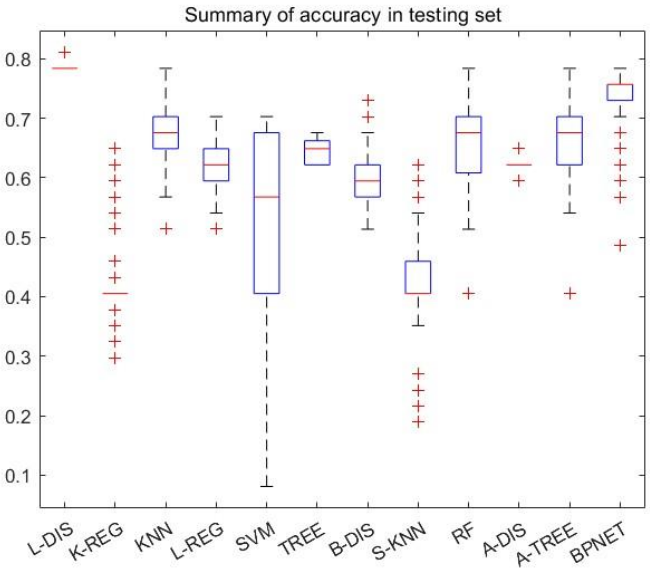

Figure 3: Prediction results of the 100 repeated experiments on the testing set

3.3 Comparison of classifier performance

Due to the relatively close descriptive statistics of model prediction accuracy, we conducted one-sided t-tests on the repeated experiments to validate whether different classification methods obtained significantly superior performance from the 100 independent runs. Firstly, we examined whether the model performance was superior to random guessing. As we categorized the drama genre into five major classes, we conducted one-sided t-tests comparing the model prediction accuracy to 0.2. The null hypothesis was that the model prediction accuracy is less than 0.2, and the alternative hypothesis was that the model prediction accuracy is greater than or equal to 0.2. The results of the one-sided t-tests for the training and testing sets are presented in Table 5. It is evident that for both the training and testing sets, each model's one-sided t-test rejected the null hypothesis at a significance level of 0.01. This signifies that the prediction accuracy of each model is significantly greater than 0.2. This

experiment provides preliminary validation of the predictability of text genres based on word frequencies, laying the foundation for further exploration in subsequent analyses.

Table 5: Significance tests of machine learning models against random prediction accuracy

Model	P value in training set	P value in testing set
L-DIS	0.00^{***}	0.00^{***}
K-REG	0.00^{***}	0.00^{***}
KNN	0.00^{***}	0.00^{***}
L-REG	0.00^{***}	0.00^{***}
SVM	0.00^{***}	0.00^{***}
TREE	0.00^{***}	0.00^{***}
B-DIS	0.00^{***}	0.00^{***}
S-KNN	0.00^{***}	0.00^{***}
RF	0.00^{***}	0.00^{***}
A-DIS	0.00^{***}	0.00^{***}
A-TREE	0.00^{***}	0.00^{***}
BPNET	0.00^{***}	0.00^{***}

Furthermore, we aimed to determine the relative superiority of the models in terms of their generalization performance. To assist in evaluating the experimental models based on accuracy obtained from the testing set, we conducted one-sided t-tests, and the results are presented in Table 6. Based on the test results, we generated a heatmap in Figure 4. The t-test matrix element (i,j) corresponds to the hypothesis that the model's accuracy in the i-th row is less than or equal to the model's accuracy in the j-th column, with the alternative hypothesis being greater. From the heatmap, it is evident that L-DIS stands out as the best-performing model, significantly outperforming all other models at the 0.01 significance level. This further supports our previous conclusion that L-DIS is the optimal choice for handling high-dimensional, small-sample datasets. On the other hand, K-REG performs as the worst model, exhibiting significant inferiority to all other experimental models. This is primarily attributed to the dataset's excessive features in the Shakespeare dataset, causing multicollinearity issues even after feature extraction, thus affecting the high-dimensional projection effect of the kernel function and resulting in the worst predictive performance.

BPNET emerges as the second-best model in predictive performance, significantly outperforming all models except L-DIS. This can be attributed to the strong fitting ability of BPNN [29], which effectively learns the underlying relationship between word frequencies and genres within high-dimensional features. However, due to the small sample size and a large number of parameters to be estimated in BPNN, its performance is weaker than that of L-DIS. Additionally, we observed that ensemble learning methods affected the models' generalization performance, as the predictive performance of models using ensemble learning was weaker than that of the original models. This indicates that ensemble learning methods are not suitable for small-sample datasets. The aforementioned analysis aids in determining the models' relative advantages and disadvantages, providing valuable insights into their performance on the given dataset.

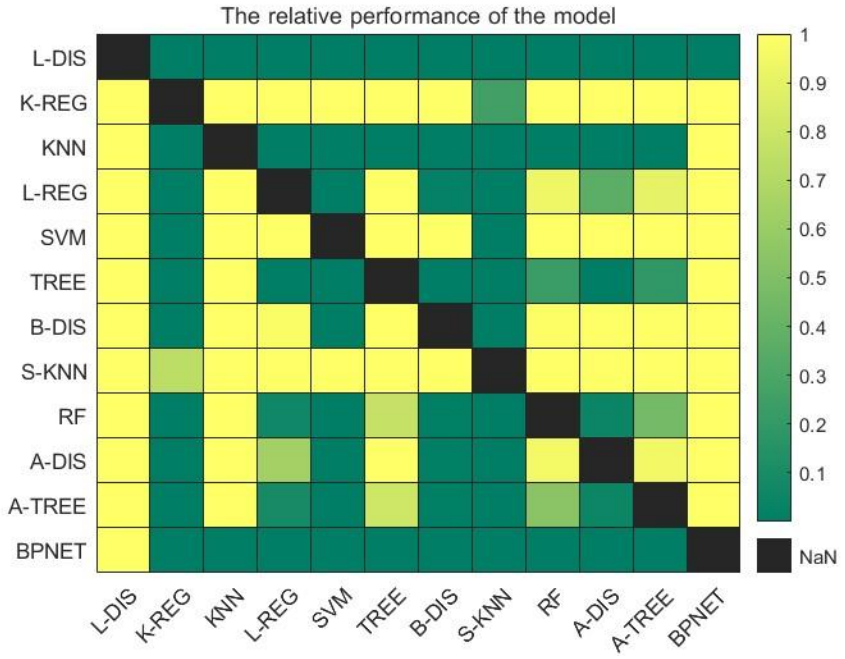

Figure 4: Heatmap of Model Predictive Performance

Table 6: Significance Testing of Relative Performance of Machine Learning Models

Model	L-DIS	K-REG	KNN	L-REG	SVM	TREE	B-DIS	S-KNN	RF	A-DIS	A-TREE	BPNET
L-DIS	NAN	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}
K-REG	1.0	NAN	1.0	1.0	1.0	1.0	1.0	0.3	1.0	1.0	1.0	1.0
KNN	1.0	0.0^{***}	NAN	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	1.0
L-REG	1.0	0.0^{***}	1.0	NAN	0.0^{***}	1.0	0.0^{***}	0.0^{***}	0.9	0.4	0.9	1.0
SVM	1.0	0.0^{***}	1.0	1.0	NAN	1.0	1.0	0.0^{***}	1.0	1.0	1.0	1.0
TREE	1.0	0.0^{***}	1.0	0.0^{***}	0.0^{***}	NAN	0.0^{***}	0.0^{***}	0.2	0.0^{***}	0.2	1.0
B-DIS	1.0	0.0^{***}	1.0	1.0	0.0^{***}	1.0	NAN	0.0^{***}	1.0	1.0	1.0	1.0
S-KNN	1.0	0.7	1.0	1.0	1.0	1.0	1.0	NAN	1.0	1.0	1.0	1.0
RF	1.0	0.0^{***}	1.0	0.1^{*}	0.0^{***}	0.8	0.0^{***}	0.0^{***}	NAN	0.0^{***}	0.5	1.0
A-DIS	1.0	0.0^{***}	1.0	0.6	0.0^{***}	1.0	0.0^{***}	0.0^{***}	1.0	NAN	1.0	1.0
A-TREE	1.0	0.0^{***}	1.0	0.1^{*}	0.0^{***}	0.8	0.0^{***}	0.0^{***}	0.5	0.1^{*}	NAN	1.0
BPNET	1.0	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	0.0^{***}	NAN

(***p<0.01 **p<0.05 *p<0.1)

4. Discussion of findings

We have determined that the optimal model for Shakespearean literary genre classification is the L-DIS model, which effectively establishes the mapping relationship between Shakespearean text features and genre categories. However, literary research requires further exploration and refinement of the connections between the vocabulary features constituting this mapping and the dramatic imagery and themes. To address this, we utilized the popular Shapley value method to assess which words have the greatest importance in the model's predictions, thereby revealing their intrinsic literary associations with Shakespearean genres. In Section 4.1, we computed the important words for all

models to assess their abilities to identify significant words. Furthermore, in Section 4.2, we computed the important words corresponding to different genres in the L-DIS model, uncovering significant literary features associated with Shakespearean themes.

4.1 The 20 most important words in predictions for different models

Table 7: The 20 most important words for each model

Model	Words
L-DIS	blood death london flowers lives roots bed moved made lord england we_royalplural_ earth power majesty marry heart kingdom lies therefore
K-REG	and i my a the is it to_infinitive_ for_preposition_ but as to_preposition_ be that_relative_ thee lord sir shall that_demonstrative_ there
KNN	a it is my sir to_infinitive_ lord to_preposition_ be but that_relative_ as for_preposition_ thee shall death that_demonstrative_
L-REG	sir lord death is a but we_royalplural_ blood that_relative_ to_preposition_ as and for_preposition_ to_infinitive_ shall am sweet well heart england
SVM	a sir is lord death it but and to_infinitive_ blood we_royalplural_ as to_preposition_ that_relative_ i the for_preposition_ be that_demonstrative_ shall
TREE	death england moved power it blood we_royalplural_ lord for_preposition_ bed but majesty thee live i lives fit the and a
B-DIS	a but i lord death am is to_preposition_ england that_demonstrative_ and we_royalplural_ blood it made my TRUE marry sweet the
S-KNN	it shall well to_preposition_ blood must day lives we_royalplural_ is and fit i my be heart sir that_relative_ the as
RF	death blood england lives marry majesty we_royalplural_ a ass sword power english earth sir fool richard lord london sweet am
A-DIS	a but earth moved made TRUE to_preposition_ lord power sweet marry and my being the majesty is flowers as lood
A-TREE	england death blood lives majesty moved english am fool lord kingdom wits ass therefore roots bed marry made we_royalplural_ sword
BPNET	the and i a my is to_infinitive_ it to_preposition_ be but for_preposition_ as that_relative_ shall by_preposition_ that_demonstrative_ sir thee they

We computed the Shapley value of each word in the feature subset for each model during predictions and listed the 20 most important words. It is important to note that since our research involves a multi-classification task, each word in each sample may have multiple Shapley values. To aid in comprehending word importance, we selected only the Shapley values corresponding to the

predicted true class for each sample and calculated the average over all samples. Table 7 presents the 20 most important words for each model during predictions.

For each predictive model, we selected the top 20 words in terms of importance and compared them with the 20 most frequently occurring words in the best-performing L-DIS model in terms of prediction accuracy. Firstly, it is evident that the L-DIS model exhibits the strongest ability to distinguish between function words and grammar-connecting words, followed by the RF model and A-TREE model. In the list of words from the L-DIS model, only one conjunction, "therefore," is present, while the rest are meaningful nouns or verbs. Moreover, the L-DIS model effectively filters out meaningful content words, as only one word, "lies," in its top 20 words does not appear in the top 20 words provided by other models. Finally, there is strong consistency in the selection of grammar-connecting words among different models. We can see that,the consistent occurrence of conjunctions such as "but", "and" and "and" that ", auxiliary verbs "is", "am", "be", prepositions "the", "for", personal pronouns "thee", "it", "my", quantifier "a" and infinitive "to" in multiple patterns meets the basic features of English grammar.

4.2 The most important words corresponding to different genres

To further investigate the inherent relationship between word frequencies and genres when using the L-DIS model for classification, we calculated the Shapley value of each word when predicting different genres. Table 8 presents the top twenty important words for each category based on their corresponding Shapley values.

Table 8: The 20 most important words for different genres in the L-DIS model

Genre	Words
Comedy	'a' 'and' 'death' 'blood' 'TRUE' 'sweet' 'but' 'moved' 'london' 'we_royalplural_' 'thee' 'being' 'lord' 'earth' 'lives' 'power' 'lies' 'am' 'as' 'i'
History	'lord' 'but' 'we_royalplural_' 'lives' 'england' 'that_demonstrative_' 'to_preposition_' 'i' 'is' 'a' 'to_infinitive_' 'sweet' 'teach' 'fit' 'therefore' 'marry' 'TRUE' 'fool' 'made' 'am'
Tragedy	'a' 'death' 'is' 'to_preposition_' 'as' 'moved' 'sir' 'but' 'earth' 'i' 'lord' 'done' 'made' 'and' 'flowers' 'majesty' 'richard' 'sacred' 'bed' 'TRUE'
Tragicomedy	'same' 'to_preposition_' 'faith' 'roots' 'that_demonstrative_' 'sir' 'shall' 'day' 'sung' 'being' 'but' 'am' 'power' 'flowers' 'is' 'befallen' 'majesty' 'made' 'nearest' 'it'

Shakespeare's comedies and historical plays were written in the early stages of his dramatic career, characterized by an optimistic and cheerful tone, filled with confidence in resolving social conflicts through humanistic ideals. His comedies, in particular, are renowned for their complex plots, well-developed characters, and humorous dialogues. From Table 8, it can be observed that comedies often employ light and cheerful imagery. The word 'sweet' is frequently associated with love and romantic emotions. In Shakespeare's comedies, sweetness can represent the allure, joy, and happiness of love, often appearing in songs and soliloquies depicting romantic relationships. Nevertheless, despite the prevalent themes of merriment and celebration in Shakespeare's comedies, they also feature sharp twists and conflicts. The word 'death' often emerges as a potential threat or turning point, emphasizing the contrast and uncertainty of life within the comedy. 'Blood' carries multiple symbolic meanings in Shakespeare's works. It can symbolize family, lineage, and hereditary relationships, as well as

violence, passion, and the force of destiny. In comedies, blood can be used to create dramatic conflicts or imply close connections between characters. Additionally, 'TRUE' represents another essential theme in Shakespeare's plays. In his comedies, truth often contrasts with misunderstandings, disguises, and illusions. The true identities and genuine emotions of characters are often concealed or distorted, creating comedic effects.

Through his historical plays, Shakespeare showcases the development of the English nation, the struggle for royal power, and the interplay between national glory and individual destinies. These plays delve deep into important themes such as power, politics, family, love, and national identity. We can see that terms such as "Sir", "Lord", "we_royalplural_" and "England" for noble titles and politics have become more prominent. In the historical plays, England's royalty and nobility often take center stage, and the plot revolves around the power, status, and political struggles within the aristocracy, depicting the rise and fall and glory of the nation. The word 'marry' suggests that marriage is frequently used as a means of political alliances, inheritance rights, or national interests. Additionally, the term 'fool' frequently appears as a comedic character in historical plays. For instance, in "Henry IV, Part 1," Falstaff is portrayed as a 'fool,' characterized by his libertine lifestyle and revelry, often seen as a humorous figure. Through his comical actions and dialogues, the play satirically reveals issues concerning political power and social order.

Shakespeare's tragedies were written during the mid-period of his dramatic career. Death is one of the most significant themes in Shakespearean tragedies. In his tragedies, characters often undergo a philosophical exploration of the meaning of death and existence by depicting their fear, acceptance, or resistance towards death. The term 'bed' may refer to a literal bed, but it also symbolizes a place of sleep, rest, and death. In tragedies, 'bed' and 'sleep' imply the arrival of death, the end of characters' destinies, or tragic conclusions. For example, in Hamlet's soliloquy, sleep is used as a metaphor for death: "To sleep, perchance to dream—ay, there's the rub. For in that sleep of death what dreams may come. When we have shuffled off this mortal coil, Must give us pause." [30]. The words 'made' and 'done' represent the result of formation, completion, or actions. In tragedies, they may be related to characters' fate, personality, or decisions, hinting at the consequences of their molded inner selves or behaviors. Shakespeare's tragedies vividly depict the conflict and contradiction between humanistic ideals and harsh reality. The protagonists awaken from medieval ages, eager to rectify the time and enlighten the era. As Hamlet realizes after the death of his father and the usurpation of the throne by his uncle, "the time is out of joint: O cursed spite, that ever I was born to set it right!" [31]. However, they are unable to overcome the limitations of their era and themselves, ultimately facing inevitable failure and sacrifice in their unequal struggle against hostile forces around them. In the end, the conflicts between ambition and failure lead to the crisis of humanism.

Shakespeare's tragicomedies were written in the later stages of his dramatic career, during which he remained committed to humanistic ideals. These plays exhibit strong elements of legend and romance, often culminating in resolutions of forgiveness and reconciliation. In Table 8, words like 'faith,' 'flowers,' and 'sung' are frequently used in the plays as symbols and metaphors, reflecting the celebration of beautiful moments, love, and happiness.

5. Conclusion

In our study, we utilized a Shakespeare dataset and employed 12 different machine learning classification models with word frequencies as features. To determine feature importance and achieve dimensionality reduction, we employed decision trees. Subsequently, we utilized the reduced feature set for predicting Shakespearean drama genres. Using the Shapley value method, we further determined the important words corresponding to different genres in Shakespearean plays and revealed their underlying imagery and themes. This digital approach has contributed to the sustainable

development of Shakespearean research. Specifically, we trained each machine learning model using randomly selected 80% of the dataset as training data, and evaluated their predictive performance on the remaining 20% as the testing set. Among the models, the linear discriminant model based on Bayesian optimization demonstrated the best predictive ability and effectively prevented overfitting.

We further analyzed the 20 most important words in the LDA model and conducted a detailed analysis of the literary connotations expressed by these words in different genres. In comedies, positive imagery such as 'sweet' and 'TRUE' is frequently used, reflecting the light-hearted and joyful tone prevalent in this genre. However, comedies also feature intense plot twists and dramatic conflicts symbolized by words like "blood" and "death," which may not be readily apparent using traditional qualitative literary research approaches. Historical plays often explore the struggles and political conflicts within the royal nobility, leading to the frequent use of terms like 'sir,' 'lord,' and 'we_royalplural_' in character dialogues. The plotlines of political alliances in historical plays also make 'marry' an important vocabulary. In tragedies, Shakespeare often employs characters' dialogue to delve into the philosophical meaning of death and existence. The important words 'death,' 'bed,' and 'sleep' directly or metaphorically support this exploration, heightening the conflict between humanistic ideals and harsh realities. Lastly, for Shakespeare's tragicomedies, key words include 'faith,' 'flowers,' and 'sung,' which often represent yearning for happiness, aligning well with the strong romantic and legendary elements in Shakespeare's tragicomedies.

Given the effectiveness of LDA and the Shapley value method in Shakespearean genre analysis, they hold the potential for further application in the future. We believe that the methods presented above will offer a novel digital perspective for future literary research, contributing to the popularization of literary studies and promoting its sustainable development.

Acknowledgement

Funding: This research received no external funding.

Data Availability Statement

You can find the data in ‘http://archive.ics.uci.edu/dataset/747/181 +early+modern+english+ plays+transcriptions+of+early+editions+in+tei+encoding’

Conflicts of Interest

The author declare no conflict of interest.

References

[1] Bloom. Dramatists and Dramas, 1st ed., Chelsea House: London, England, 2005, p 7.
[2] Colyvas, K., Egan, G., Craig, H. Changes in the length of speeches in the plays of William Shakespeare and his contemporaries: A mixed models approach. Plos One 2023, 18.
[3] Pivetti, K. Publishing the History Play in the Time of Shakespeare: Stationers Shaping a Genre. Shakespeare Quart 2022, 73, 150-152.
[4] Murphy, S., Archer, D., Demmen, J. Mapping the links between gender, status and genre in Shakespeare's plays. Lang Lit 2020, 29, 223-245.
[5] Vickers, B. Shakespeare and Authorship Studies in the Twenty-First Century. Shakespeare Quart 2011, 62, 106-142.
[6] Aristotle. Poetics., Penguin: London, UK, 1996, pp. 18-53.
[7] Whissell, C. Quantifying genre: An operational definition of tragedy and comedy based on Shakespeare's plays. Psychol Rep 2007, 101, 177-192.
[8] Culpeper, J. Keyness Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. Int J Corpus Linguis 2009, 14, 29-59.
[9] Vickers, B. Identifying Shakespeare's Additions to The Spanish Tragedy (1602): A New(er) Approach. Shakespeare

2012, 8, 13-43.
[10] Papp-Zipernovszky, O., Mangen, A., Jacobs, A., Ludtke, J. Shakespeare sonnet reading: An empirical study of emotional responses. Lang Lit 2022, 31, 296-324.
[11] LeCun, Y., Bengio, Y., Hinton, G. Deep learning. Nature 2015, 521, 436-444.
[12] Xiang, Z., Du, Q. Z., Ma, Y. F., Fan, W. G. A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Manage 2017, 58, 51-65.
[13] Hirschberg, J., Manning, C. D. Advances in natural language processing. Science 2015, 349, 261-266.
[14] Alghazzawi, D. M., Alquraishee, A., Badri, S. K., Hasan, S. H. ERF-XGB: Ensemble Random Forest-Based XG Boost for Accurate Prediction and Classification of E-Commerce Product Review. Sustainability-Basel 2023, 15.
[15] Fedotova, A., Romanov, A., Kurtukova, A., Shelupanov, A. Authorship Attribution of Social Media and Literary Russian-Language Texts Using Machine Learning Methods and Feature Selection. Future Internet 2022, 14.
[16] Jacobs, A. M., Kinder, A. "The Brain Is the Prisoner of Thought": A Machine-Learning Assisted Quantitative Narrative Analysis of Literary Metaphors for Use in Neurocognitive Poetics. Metaphor Symbol 2017, 32, 139-160.
[17] Ustaszewski, M. Towards a machine learning approach to the analysis of indirect translation. Transl Stud 2021, 14, 313-331.
[18] Xue, S. W., Ludtke, J., Sylvester, T., Jacobs, A. M. Reading Shakespeare Sonnets: Combining Quantitative Narrative Analysis and Predictive Modeling-an Eye Tracking Study. J Eye Movement Res 2019, 12.
[19] Plechac, P. Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns. Digit Scholarsh Hum 2021, 36, 430-438.
[20] Liu, X. T., Xu, A. B., Liu, Z., Guo, Y. F., Akkiraju, R., Assoc, C. M. Cognitive Learning: How to Become William Shakespeare. In CHI Conference on Human Factors in Computing Systems (CHI), 2019.
[21] Moscato, P., Craig, H., Egan, G., Haque, M. N., Huang, K. V., Sloan, J., de Oliveira, J. C. Multiple regression techniques for modelling dates of first performances of Shakespeare-era plays? Expert Syst Appl 2022, 200.
[22] Subhan, F., Saleem, S., Bari, H., Khan, W. Z., Hakak, S., Ahmad, S., El-Sherbeeny, A. M. Linear Discriminant Analysis-Based Dynamic Indoor Localization Using Bluetooth Low Energy (BLE). Sustainability-Basel 2020, 12.
[23] Moradzadeh, A., Sadeghian, O., Pourhossein, K., Mohammadi-Ivatloo, B., Anvari-Moghaddam, A. Improving Residential Load Disaggregation for Sustainable Development of Energy via Principal Component Analysis. Sustainability-Basel 2020, 12.
[24] Ding, C., Wang, D. G., Ma, X. L., Li, H. Y. Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees. Sustainability-Basel 2016, 8.
[25] Mayakuntla, P. K., Ghosh, D., Ganguli, A. Classification of Corrosion Severity in Concrete Structures Using Ultrasonic Imaging and Linear Discriminant Analysis. Sustainability-Basel 2022, 14.
[26] Hoyle, D. C. Accuracy of Pseudo-Inverse Covariance Learning-A Random Matrix Theory Analysis. Ieee T Pattern Anal 2011, 33, 1470-1481.
[27] Palm, N., Landerer, M., Palm, H. Gaussian Process Regression Based Multi-Objective Bayesian Optimization for Power System Design. Sustainability-Basel 2022, 14.
[28] Gao, S. Y., Zhang, F. R., Ning, W., Wu, D. Y. Optimization of Cargo Shipping Adaptability Modeling Evaluation Based on Bayesian Network Algorithm. Sustainability-Basel 2022, 14.
[29] Tang, R. X., Yan, E. C., Wen, T., Yin, X. M., Tang, W. Comparison of Logistic Regression, Information Value, and Comprehensive Evaluating Model for Landslide Susceptibility Mapping. Sustainability-Basel 2021, 13.
[30] Shakespeare., Raffel., Bloom. Hamlet(The Annotated Shakespeare), 1st ed., Yale University Press: New Haven, America, 2003, pp. 97.
[31] Shakespeare., Raffel., Bloom. Hamlet(The Annotated Shakespeare), 1st ed., Yale University Press: New Haven, America, 2003, pp. 52.