Literary text mining using verb feature clustering

Marianna Dilai[1,*,†] and Iryna Dilai[2,†]

[1] *Lviv Polytechnic National University, Stepan Bandera Street, 12, Lviv, 79013, Ukraine*
[2] *Ivan Franko National University of Lviv, Universytetska Street, 1, Lviv, 79000, Ukraine*

Abstract

In this study, we explore text clustering techniques applied to a corpus of the works by the renowned Canadian postmodernist writer Margaret Atwood. Leveraging unsupervised machine learning methods, we investigate the thematic affinities within her literary legacy. Our approach involves employing document n-gram embeddings and bag-of-words clustering algorithms to analyze the structural similarities among the works. Additionally, we introduce a novel feature-based clustering model focusing on verbs, essential elements in English sentence structure and meaning formation. We assess the performance of verb-centered clustering through experimentation and evaluation, including the use of logistic regression classifiers and Rand index calculation. Ultimately, our findings shed light on the prevailing topics and thematic patterns permeating the author's diverse literary oeuvre, offering insights for computational text mining methodologies and literary analysis.

Keywords

text mining, machine learning, clustering, topic modeling, logistic regression, prediction, verb, distant reading, literary text

1. Introduction

Text is one of the most common and sophisticated types of data. Text mining as a type of large-scale data mining has become a promising direction in data science, combining NLP, machine learning, and information retrieval [1]. Transforming the unstructured text into a structured one and revealing hidden meaningful structural patterns enables the discovery of high-quality insights. Text mining is aimed at capturing key concepts, topics, trends, and latent relationships in vast collections of text material.

Recently, more and more attention has been paid to the computer-aided study of literary text. Computational methods are being applied to analyze large libraries of literary data. The most prominent of them is the method of 'distant reading' proposed by Franko Moretti [2] as a computer-assisted mode of reading that provides abstraction and the reader's/researcher's detachment to viewing, processing, and interpreting the text. Normally, distant reading encompasses topic modeling, stylometry, and network analysis.

MoDaST-2024: 6th International Workshop on Modern Data Science Technologies, May, 31 - June, 1, 2024, Lviv-Shatsk, Ukraine
* Corresponding author.
† These authors contributed equally.
marianna.p.dilai@lpnu.ua (M. Dilai); iryna.dilay@lnu.edu.ua (I. Dilai)
0000-0001-5182-9220 (M. Dilai); 0000-0001-9626-290X (I. Dilai)

Modern distant reading resorts to machine learning methods and techniques to produce robust computational findings and predictions.

This study aims to harness literary text mining techniques by testing a feature clustering model capable of eliciting the prevalent thematic patterns and measuring the affinity among the literary works exemplified by a postmodern Canadian writer Margaret Atwood.

The tasks set here are as follows:

- To create and process a corpus of works by M. Atwood.
- To elaborate an efficient text analysis model that can be used for text classification predictions.
- To test the performance of verb-based text clustering, assuming that verbs constitute meaningful and informative text structure features.
- To account for the thematic patterns and topics generated by the model.

Viewing fiction as information and applying text mining can shed light on sociocultural topics not immediately clear to readers/researchers. The new mode of reading, where the AI performs the function of an active participant, rather than a tool, opens new prospects for the literary revolution and requires boosting machine learning capabilities applied to literary text data.

2. Related works

By synthesizing existing research and scholarship on text mining with a focus on verb feature clustering, we seek to show the current state of the field and justify our methodology choice.

Literary text mining, a burgeoning field at the intersection of computational linguistics and literary studies, offers possibilities for uncovering hidden linguistic patterns within literary texts. As a result, it has been addressed by several modern researchers [3-5]. Authorship attribution methods were especially popular at the beginning of the 21st century [6]. Now, automatic text mining reveals genre-depending features, and is used for topic modeling [5-7] and cataloging documents [8-9], author's stylometry, and, most recently, distant reading [2]. F. Moretti initially applied the distant reading approach relying on computational techniques to represent the perspectives of the "great unread", which encompassed a massive collection of literary texts and strived to quantitively analyze the titles of the novels and classify them [2].

Clustering algorithms take center stage as they play a crucial role in organizing text documents into meaningful clusters. From the classic K-means to the more nuanced hierarchical clustering, each algorithm has shown its efficacy in revealing thematic structures. Agglomerative clustering is a type of hierarchical clustering that follows a bottom-up approach. Normally Euclidean distance and the Ward linkage method are used to measure the distance between the clusters.

In order to reduce dimensionality and visualize the data in the low dimensional space, the t-distributed stochastic neighbor embedding (t-SNE) method as a specific unsupervised machine learning technique is used [10].

If preprocessing is aimed at cleaning unstructured text, e.g., eliminating stop words, feature selection converts the text into structured data. One of the common ways of feature selection is a bag of words where words are represented as vectors. For instance, M. Short used this method to leverage machine learning and information extraction to assign subject headings to dime novels [5].

The first model to perform classification and predict labels for fiction test sets was tested in 2013 using Weka 3 [11]. Another popular machine learning tool used for text data mining is a Java application for classification called MALLET [12]. It uses Latent Dirichlet Allocation (LDA) for topic modeling and was, in particular, leveraged to analyze contemporary popular fiction, over 1,000 New York Times bestsellers, and genre fiction novels [4].

The vector space model relies on a bag of words or n-grams in the text. Feature selection is believed to reduce redundancy in the representation of text data and save computational time text. N-grams are often selected as features that successfully represent text for stylistic purposes [13]. Nonetheless, other meaningful features can be leveraged to facilitate automatic text mining and aptly classify meaningful information.

The major problems facing researchers nowadays pertain to the need for dimensionality reduction and boosting the performance of machine learning algorithms applied to text mining.

3. Methods and materials

Topic modeling is viewed as a type of 'soft' clustering of text documents [5]. It belongs to unsupervised machine learning techniques and is based on the structural similarity (or difference) in data. Statistical measures are applied to calculate the distance between the documents. In our study, we rely on the vectors obtained for each document by aggregation of n-gram embeddings. Documents close to one another have similar embeddings and can be considered semantically (thematically) related.

Though n-gram embeddings are generally accepted as a reliable text clustering technique, we test and compare here also other clustering techniques. In particular, we presume that text clustering can be done based on the most informative features. As far as content structuring is concerned, such features can be all verbs (verb forms) in the text. Stemming from the pivotal role of the verb in the sentence as the predicate, its low possibility of being omitted, a high semantic load of lexical verbs, and their central role in the narrative as plot building blocks, we assume that verbs contribute significantly to uncovering the content and theme of the text. As a result, they can serve as meaningful features for document clustering.

The automated data processing is performed here by applying Orange, a machine learning and data mining suite for data analysis through Python scripting [14]. It was developed by the Laboratory of Bioinformatics, Faculty of Computer Science, University of Ljubljana. We use Orange 3.36.2, which contains several text mining widgets [15].

The material of the research is a corpus of 19 books (documents) by a prominent Canadian postmodernist writer Margaret Atwood. It includes the novels "The Edible Woman" (1969), "Surfacing" (1972), "Lady Oracle"(1976), "The Handmaid's Tale" (1985), "Cat's Eye" (1988), "Alias Grace" (1996), "The Blind Assassin" (2000), "Oryx and Crake"

(2003), “The Year of Flood” (2009), “MaddAddam” (2013), “The Testaments” (2019), “The Heart Goes Last” (2015), a novella “The Penelopiad” (2005), collections of short fiction “Murder in the Dark” (1983), “Bluebeard’s Egg and Other Stories” (1983), “Wilderness Tips” (1991), “Good Bones” (1992), “The Tent” (2007), and a non-fiction “The Payback” (2007). The total size of the corpus is 294,606 tokens.

In order to check the thematic affinity between the novels and pinpoint the principal topics in the literary legacy of M. Atwood by utilizing computational procedures, some assumptions can be made. First, we predict that the works written consecutively at a certain period of the literary activity have more similarities. Besides, parts of the trilogy “MaddAddam” (“Oryx and Crake” (2003), “The Year of Flood” (2009), “MaddAddam” (2013) presumably share the same thematic scope and characters, as well as “The Testaments”, the sequel to the novel “The Handmaid’s Tale”. On a broader scale, the topics raised by the author are related to feminism, ecology, ecofeminism, self-identity, dystopian societies, power and oppression. However, their distribution across the works of the author and salience differs.

The applied text mining model splits the corpus of texts into two datasets: a training dataset – a corpus of 16 works by M. Atwood and a testing dataset (245,084 tokens; 23,863 types) – consisting of three works (49,522 tokens; 12,211 types). The training text dataset with unlabeled data is used to conduct unsupervised document clustering. The testing text data set includes the later works of the author which might combine a variety of the previous topics and/or can be hard to classify: “The Penelopiad” (2005), the last book of the trilogy “The MaddAdam” (2013) and “The Blind Assassin” (2000), which stands out in the literary legacy due to its complex narrative structure (novel-within-a-novel), historical context and intricate plot construction.

The experiment consists of the following interrelated stages:

1. Text preprocessing (performed by Orange to eliminate the noise).
2. Clustering of the training dataset (16 documents) is implemented by applying document n-gram-based embedding.
3. The results are compared to the clustering of the training dataset by applying a bag of words.
4. Visualizations are provided in the form of dendrograms and visualization maps.
5. The word clouds representing the clusters visualize the prominence of the most frequent lexical items in the texts.
6. The clusters are analyzed in terms of the topic affinity based on the common semantic content.
7. Logistic regression classifier is applied.
8. Predictions for the testing dataset are made based on the trained model.

This text mining model is enhanced by selecting the most informative features and thus saving computational time by reducing the size of data. Though our dataset is not of a particularly large size, testing a feature-based document clustering model can be beneficial for larger datasets.

The feature tested in this study is the verb as a pivotal element of the English sentence structure and meaning construction. Thus, we have repeated the same stages of the analysis with a reduced, verb-filtered dataset, compared the results with the results obtained from the bag of words-based clustering, and assessed the performance of the verb-centered clustering model. For part of speech tagging a Penn Treebank POS Tagger has been used.

The performance of the verb-centered model is measured by calculating the Rand index. Finally, we draw conclusions about the prevailing topics generated by the model and interpret the findings.

4. Experiment

The experimental part consists in elaborating a text mining model that relies on the stepwise combination of unsupervised machine learning methods, such as clustering, and supervised machine learning methods, such as logistic regression, as well as training and testing a model for predictions.

4.1. Text analysis model

The corpus of texts (16 novels) was loaded into the model using the Import Documents widget. The Corpus Viewer widget allows us to see all the texts in the corpus. An important initial stage of the analysis is text preprocessing, which is done with the help of the Preprocess Text widget. It runs transformation (lowercase), tokenization (Regexp), normalization (lemmatization by Lemmagen Lemmatizer), and filtering (stopwords, numbers).

Clustering implementation by Document Embedding (1) and Bag of Words (2) and the overall workflow is depicted in Figure 1.

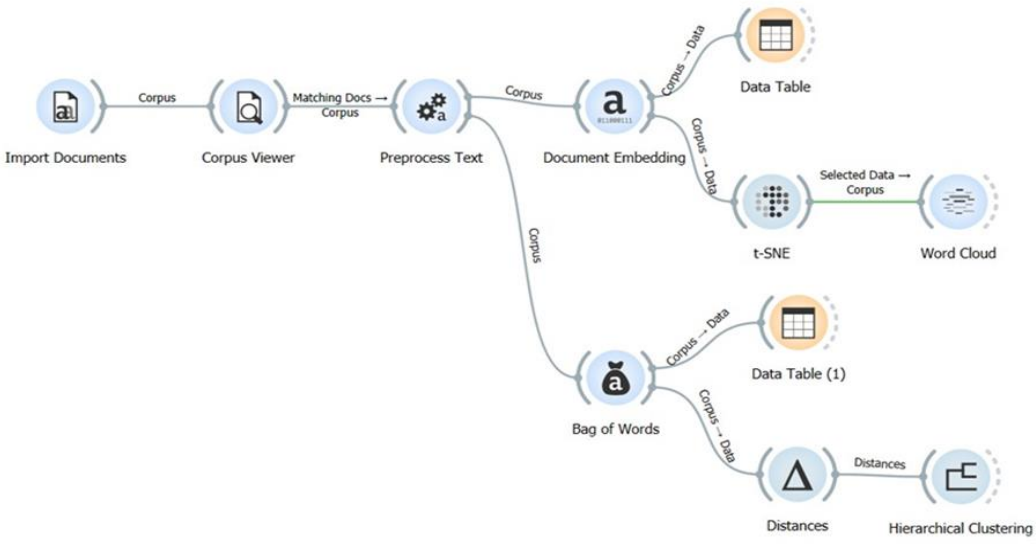

Figure 1: Clustering in Orange

4.1.1. Document embedding

The Document Embedding widget embeds documents from the corpus into vector space by using pre-trained fastText models [16]. The Document Embedding parses n-grams of each document in the corpus, obtains embedding for each n-gram using the pre-trained model, and obtains one vector for each document by aggregation of n-gram embeddings using Mean aggregator. In the Data Table, we can see 300 features for 16 instances. These features (vectors) are compared to find similar documents. We visualize documents on the map using two-dimensional data projection with t-SNE. The t-SNE widget plots the data with a t-distributed stochastic neighbor embedding method. t-SNE is a dimensionality reduction technique, similar to MDS, where points are mapped to 2-D space by their probability distribution Exaggeration 3, PCA components 20. The fastText embedding is shown in Figure 2. Documents close to each other on the t-SNE map have similar embeddings and are viewed as semantically related. The content of the clusters on the map can be explored in the Word Cloud. The clusters and their word clouds are presented in Figures 3-5.

Figure 2: The fastText embedding

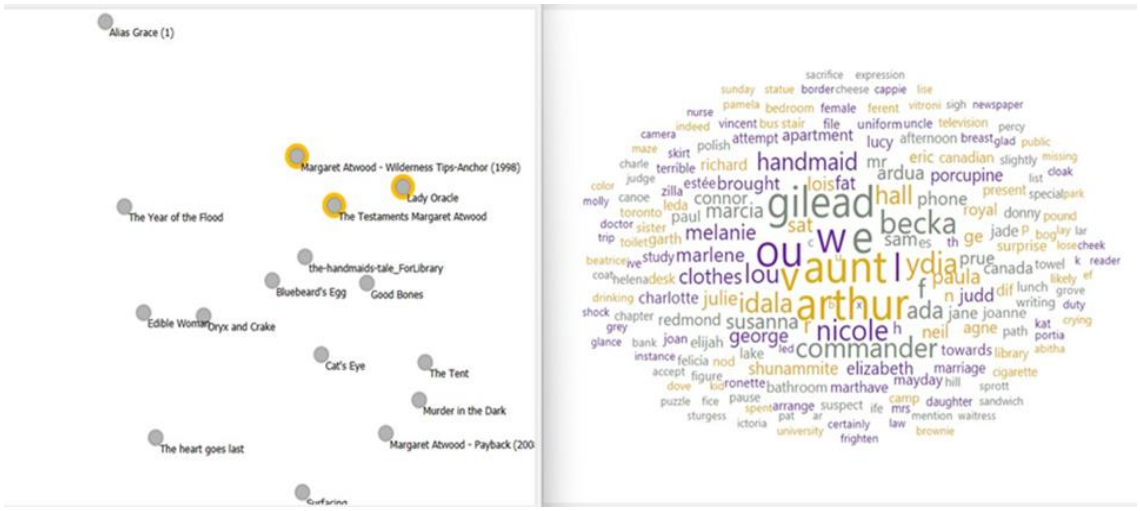

Figure 3: Cluster 1 and its Word Cloud

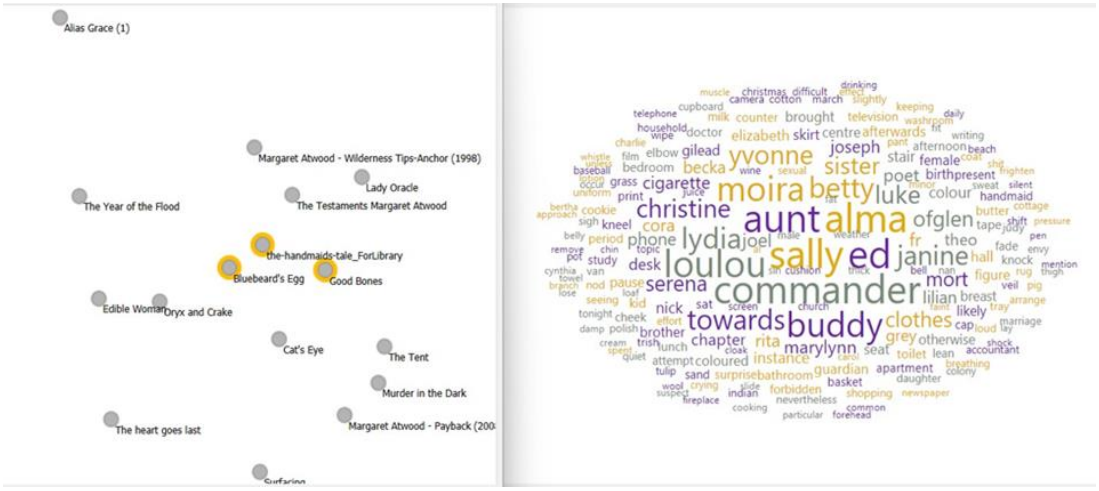

Figure 4: Cluster 2 and its Word Cloud

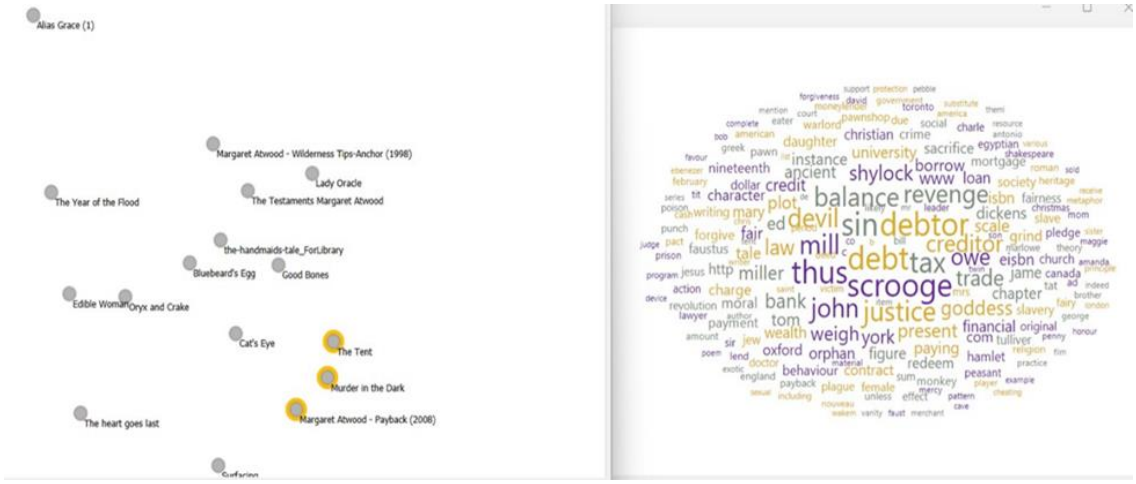

Figure 5: Cluster 3 and its Word Cloud

4.1.2. Clustering by using Bag of Words

We turn the text into a numerical representation, counting how many times each word occurs in the text. This approach is called Bag of Words. Figure 6 shows Bag of Words outputs in Data Table (23,862 features for 16 texts, term frequencies for each document).

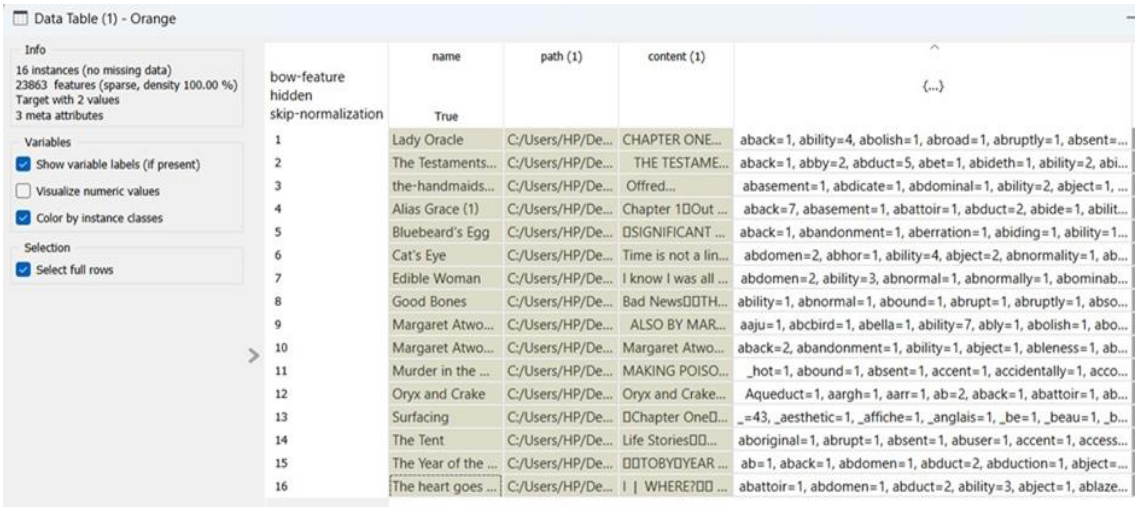

Figure 6: Bag of Words

Then we connect the Bag of Words to the Distances widget (cosine distance, which is 1-similarity) and Hierarchical Clustering. The results of clustering are given in Figure 7.

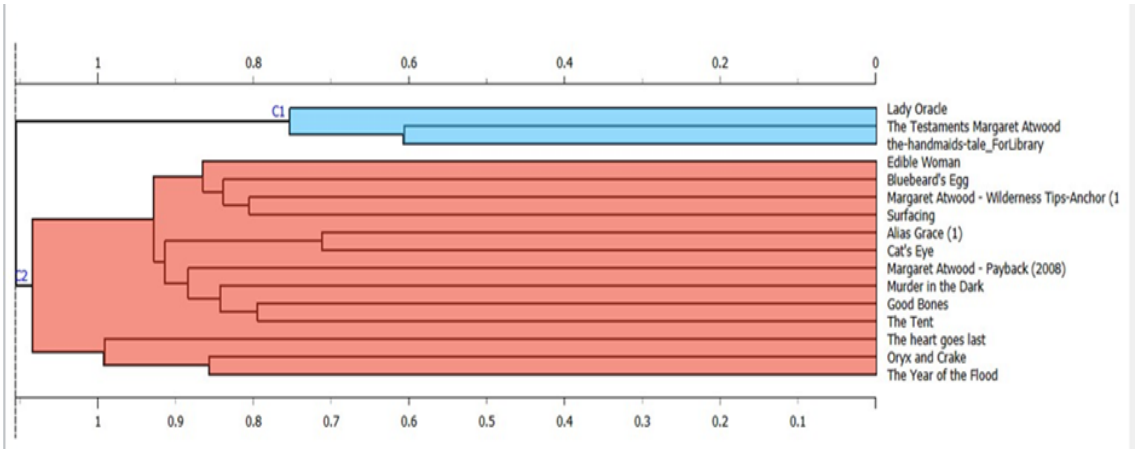

Figure 7: Clusters of texts (C1 and C2)

The next stage lies in extracting all verbs from the classified texts and eliciting the most common key verbs. The results are provided in Figures 8-9.

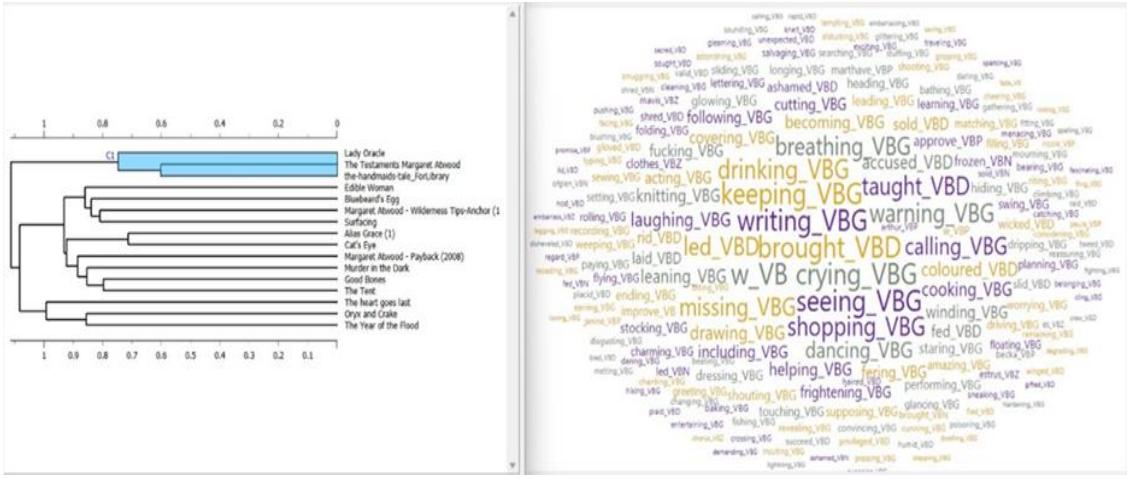

Figure 8: Verbs in C1 of the classified texts

Figure 9: Verbs in C2 of the classified texts

4.2. Predicting a text type

In order to predict the type of unclassified texts by the same author, we build a new analysis model by importing the corpus of classified texts (C1 and C2) (Figure 10 shows the analysis flow on the canvas). Using Logistic Regression, we construct the model to predict the class of texts (classification algorithm with ridge (L2) regularization). The Nomogram visualizes the Logistic Regression classifier. It offers an insight into the structure of the training data and the effects of the attributes on the class probabilities. Apart from visualization of the classifier, the widget offers interactive support for predicting class probabilities. Continuous attributes can be plotted in 2D for the selected target class C1 (Figure 11). The Nomogram displays the top words important for the classifier, which most contribute to the prediction.

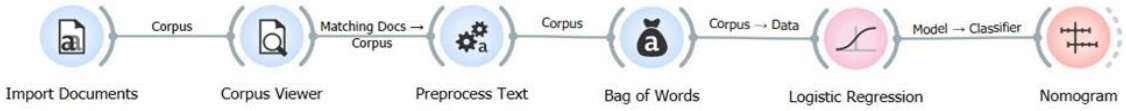

Figure 10: Analysis of classified texts

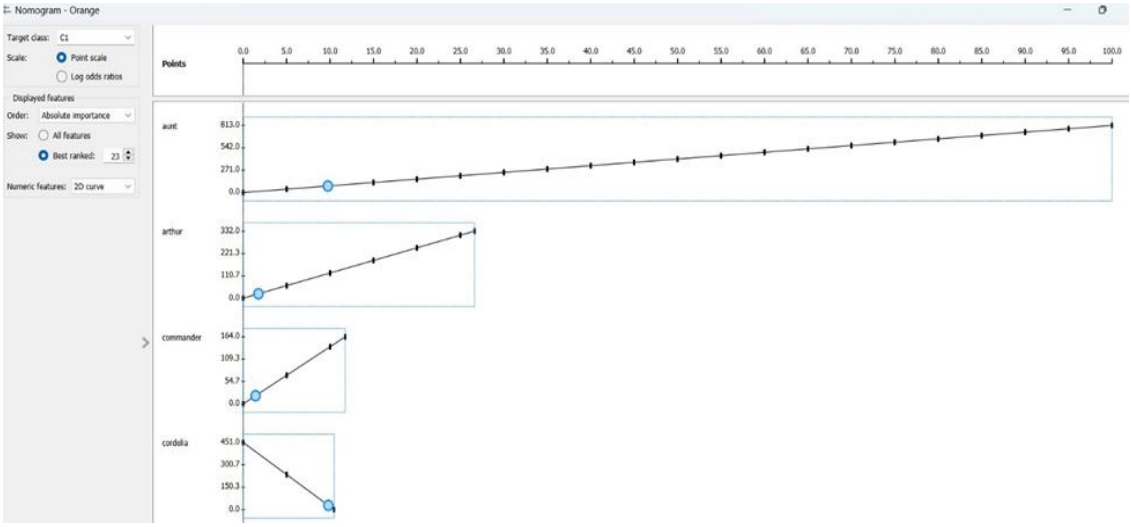

Figure 11: Nomograms for the visualization of Logistic Regression classifier (C1)

After that, we import unclassified texts ("The Penelopiad", "MaddAddam", "The Blind Assassin") using Import Documents (1) widget on the canvas (Figure 12). We connect it to the Predictions widget and Logistic Regression. Predictions widget receives a dataset and predictors (predictive models). It shows the probabilities and final decisions of predictive models. The output of the widget is another dataset, where predictions are appended as new metaattributes. The result can be observed in a Data Table (Figure 13).

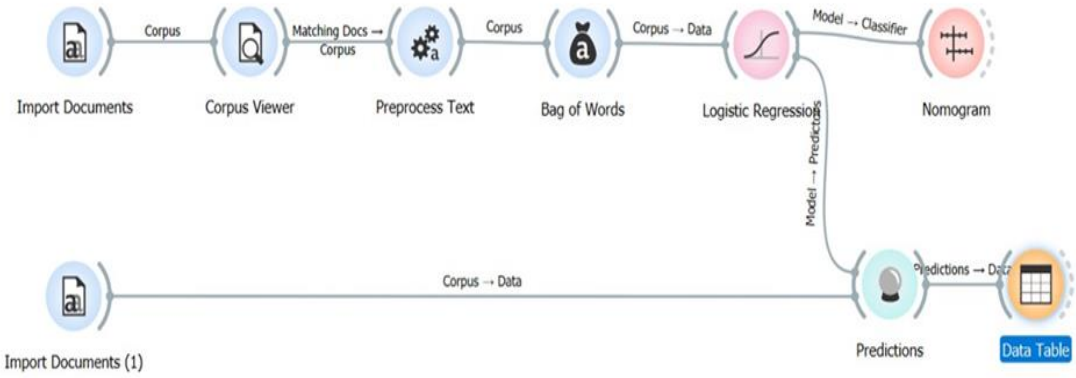

Figure 12: Prediction workflow

Data Table - Orange

	name True	path	Logistic Regression	Logistic Regression (C2)
2	Atwood Margaret. The Penelopiad_ ...	C:/Users/HP/De...	C2	0.99998
1	Atwood Margaret. The Blind Assassin...	C:/Users/HP/De...	C2	0.99989
3	Atwood Margaret. MaddAddam - ro...	C:/Users/HP/De...	C2	1

Figure 13: Prediction results for unclassified texts

4.3. Verb feature clustering model

Finally, we apply filter-based feature clustering of texts. Feature selection as a preprocessing stage is aimed at reducing the size of data and saving computational time. Choosing informative features, especially without relying on specific machine learning algorithms, can be problematic. We intend to test how the model will perform if such features are all verbs in the texts.

Verbs are viewed as essential features of content construction and, presumably, topic modeling. The question is whether it is enough to use the list of verbs from the texts to obtain the same thematic affinity clusters as with the complete texts; in other words, how informative are verbs in the text, and whether the same classification accuracy can be achieved as with the methods described above. Filter-based selection is applied to extract all the verbs from the corpus of texts and create verb datasets.

Figure 14 shows the model for finding similar texts filtered by verbs (Treebank POS tagger). With the help of the Select Columns widget, we selected all the verbs (_V) and ignored words of other POS in the Bag of Words (Figure 15). Then we iterate the procedure for hierarchical clustering using cosine distance.

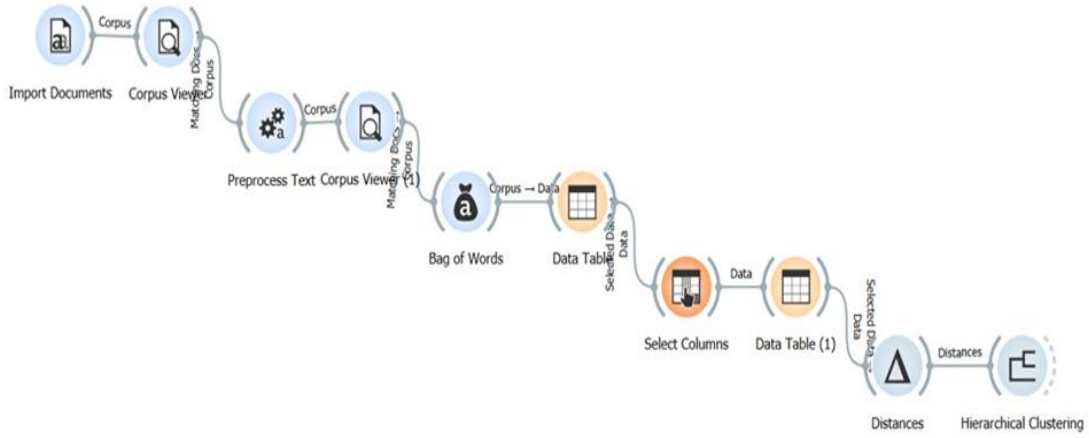

Figure 14: Verb-filtered workflow

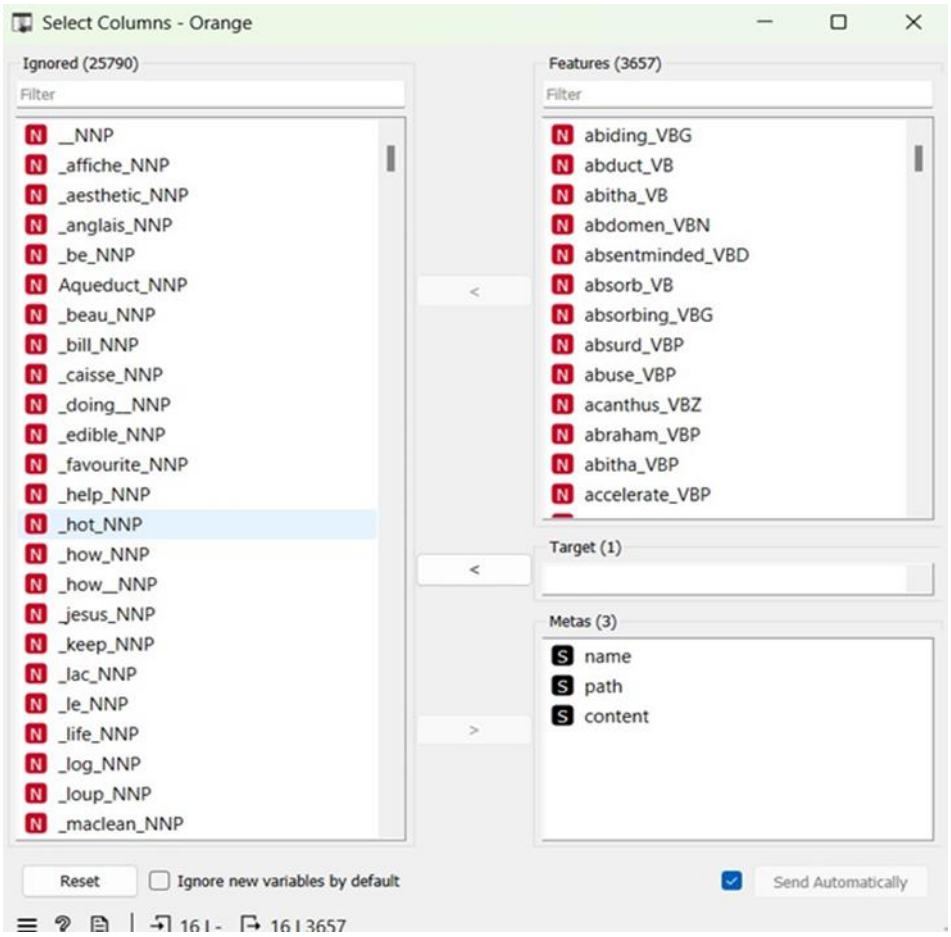

Figure 15: Verb selection

5. Results

The results of the verb feature clustering described above do not fully coincide with the previously completed text clustering (Figures 16-17).

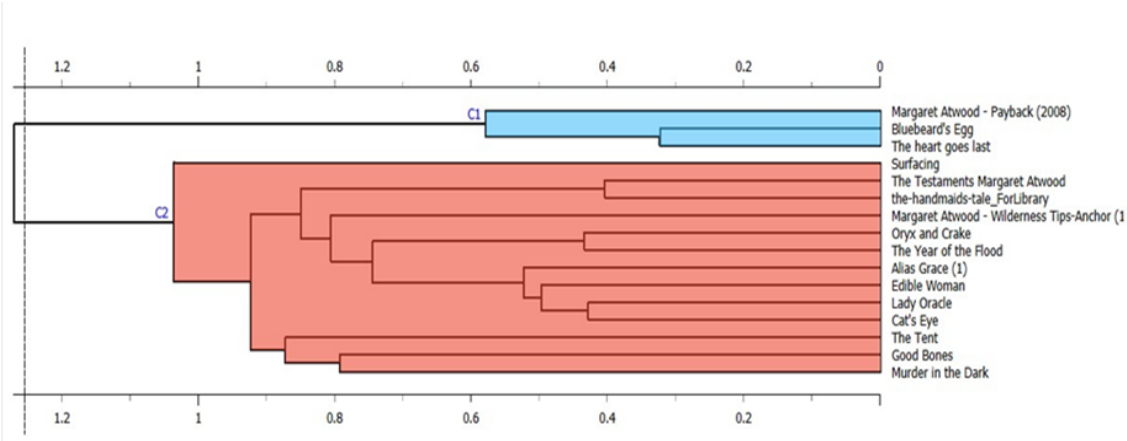

Figure 16: Clusters of texts filtered by verbs

Figure 17: Verb clusters after feature clustering and their congruence with text clusters in Figure 7

The performance of the verb feature-based document clustering model has been calculated by measuring the Rand index. We compare its performance against the performance achieved by the bag of words clustering. The Rand index is a common measure of similarity between clusterings. In our case, the rand index of 0.688 is achieved, which indicates a reasonable level of agreement between the two clusterings.

To evaluate the performance of the model it is also worthwhile to test the prediction of the classification of the new texts. The workflow for the predictions for unclassified texts filtered by verbs is shown in Figure 18. Predictions for unclassified texts filtered by verbs ("The Penelopiad", "MaddAddam", "The Blind Assassin") are provided in Figure 19.

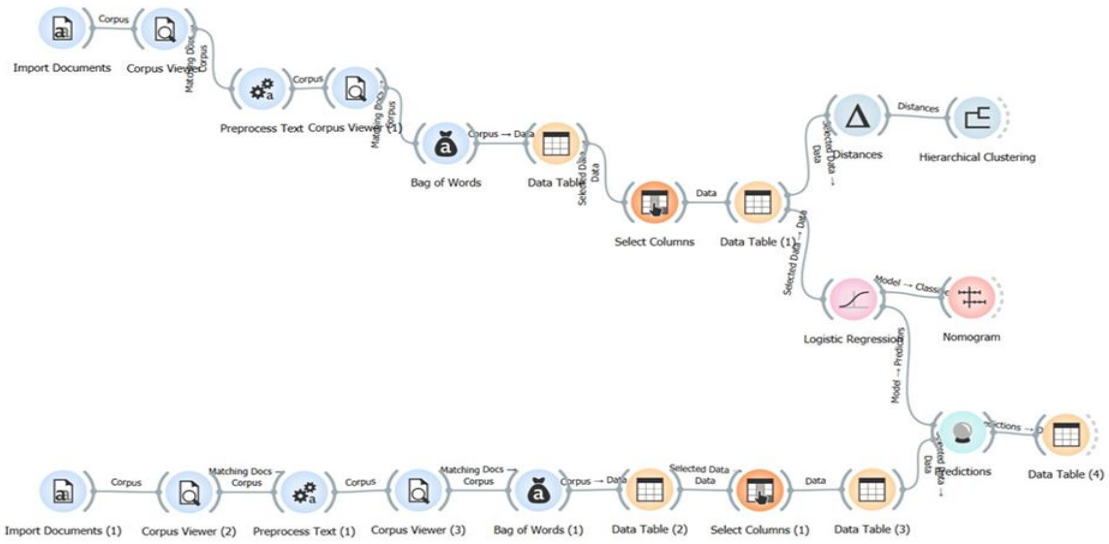

Figure 18: Prediction workflow for unclassified texts filtered by verbs

bow-feature hidden skip-normalization	name True	Logistic Regression	Logistic Regression (C2)	Logistic Regression (C1)
1	Atwood Margar...	C2	0.999748	0.000252183
2	Atwood Margar...	C2	0.99989	0.000110268
3	Atwood Margar...	C2	0.998865	0.00113502

Figure 19: Prediction results for unclassified texts filtered by verbs

The results of the classification are the same as the results obtained for unfiltered texts. The model shows high performance when predicting the class of the document, referring all three works from the testing dataset to the second cluster, which is consistent with the bag of words clustering.

6. Discussions

The applied text analysis model relies on document embedding and bag of words clustering, further used for class predictions. An attempt has been made to test the feature clustering based on verb filtering.

Literary text mining using verb feature clustering appears to be a plausible technique for reducing data dimensionality and saving computational time. The procedure applied above can be extrapolated to other types of text as well. The efficiency of verbs as meaningful features in the studied literary texts can be explained by the genre characteristics of the texts. In their majority, these texts are narratives and have a linear structure of alternating

processes signaled by verbs. The variation of verbs and their saturation in the text is also higher in fiction.

The patterns identified with verb feature clustering are quite meaningful. For example, in Cluster 1 of the resultant verb clustering (Figure 17), the lexical verbs *to knit, to cry, to fuck, to shop, to dress, to cook,* etc. are allusive of female or gender-related themes and motifs, the verbs *to pay, to keep, to save, to drink, to lead, to push, to gather* indicate the social issues in the texts from this cluster. Cluster 2 gives more prominence to the creation motifs: *to write, to draw, to color*, etc. While in Cluster 1 prevail the verbs that denote inner perception and mental states, such as *to consider, to suppose, to see*, in Cluster 2, we come across the verbs denoting the external expression of the feelings and emotions: *to laugh, to cry, to call, to shout,* etc.

The prevailing verb forms in the clusters were _VBG and _VBD.

The Rand index has been calculated to measure the agreement between the two main clusterings. A score of 0.688 has been achieved. Though the model still requires testing and improvements, the result is quite satisfactory, taking into account that the lower-level clusters within the hierarchical clustering are better classified. It is worth mentioning that both the bag of words model and the verb feature model yield the same results in correctly classifying the novels which constitute the trilogy "MaddAddam" and refer to the same (mini)cluster "The Handmaid's Tale" and its sequel "The Testaments".

The model can be augmented by introducing verb arguments, but this already leads us to the level of a clause, which can be a promising direction in the future.

The most significant result of the conducted text mining is that the trained model has been successfully tested for the prediction of text classification. Both the bag of words model and the verb feature model referred "The Penelopiad", "MaddAddam", "The Blind Assassin" to the second cluster.

7. Conclusions

The applied feature engineering lies in defining, selecting and testing verbs as informative clustering features. It encompasses text preprocessing, feature generation, verb feature selection, pattern discovery, and evaluation. The clusters have been identified with hierarchical clustering algorithms.

Verb clusters uncover insights about the analyzed texts. Patterns emerging from clusters testify to the recurring themes. The applied visualization techniques, such as scatter plots, dendrograms, and word clouds, show the intricate relationships between the clusters and, respectively, between the texts analyzed.

The conducted literary text mining with a focus on verb feature clustering shows a good performance (Rand index = 0.688), though it still needs training on a bigger dataset and other texts. It sheds light on the linguostylistic characteristics of literary works and can be viewed as a plausible distant reading technique.

As a prospect of further research, we can see the application of the multilingual Sentence-BERT model embeddings to establish semantic textual similarity, and since verbs are the core elements of any sentence structure, it can provide further insights into the informativeness of verbs as computationally efficient topic modeling features.

References

[1] J. Atkinson-Abutridy, Text Analytics. An Introduction to the Science and Applications of Unstructured Information Analysis, Chapman & Hall, 2022.

[2] F. Moretti, Distant reading, Verso, London, 2015.

[3] M. L. Jockers, D. Mimno, Significant themes in 19th-century literature, Poetics, 41(6) (2013) 750–769. doi:10.1016/j.poetic.2013.08.0052013.

[4] M. Lundy, Text Mining Contemporary Popular Fiction: Natural Language Processing-Derived Themes Across Over 1,000 New York Times Bestsellers and Genre Fiction Novels, Master's thesis, University of South Carolina, Columbia, SC, US, 2020.

[5] M. Short, Text Mining and Subject Analysis for Fiction; or, Using Machine Learning and Information Extraction to Assign Subject Headings to Dime Novels, Cataloging & Classification Quarterly, 57(5) (2019) 315–336. doi:10.1080/01639374.2019.1653413.

[6] K. van Dalen-Oskam, The Riddle of Literary Quality: A Computational Approach, Amsterdam University Press, Amsterdam, the Netherlands, 2023. doi:10.1515/9789048558155.

[7] R. S. Purves, O. Koblet, B. Adams, Analysing Environmental Narratives Computationally, in: R. S. Purves, O. Koblet, B. Adams (Eds.), Unlocking Environmental Narratives: Towards Understanding Human Environment Interactions through Computational Text Analysis, Ubiquity Press, London, UK, 2022. pp. 43–84. doi:10.5334/bcs.c.

[8] J. O. Cain, Using topic modeling to enhance access to library digital collections, Journal of Web Librarianship, 10(3) (2016) 210–225. doi:10.1080/19322909.2016.1193455.

[9] A. L. Neatrour, E. Callaway, R. Cummings, Kindles, card catalogs, and the future of libraries: a collaborative digital humanities project, Digital Library Perspectives, 34(3) (2018) 162–187. doi:10.1108/dlp-02-2018-0004.

[10] C. Wang, X. Ma, Text mining, in: Encyclopedia of Mathematical Geosciences, Springer International Publishing, Cham, 2023, pp. 1535-1537.

[11] Weka 3: Machine Learning Software in Java, 2013. URL: https://www.cs.waikato.ac.nz/ml/weka/

[12] A. K. McCallum, MALLET: a machine learning for language toolkit, 2002. URL: http://mallet.cs.umass.edu/

[13] T. Georgieva-Trifonova, M. Duraku, Research on N-grams feature selection methods for text classification. IOP Conference Series: Materials Science and Engineering, volume 1031, 2021. 012048. doi:10.1088/1757-899X/1031/1/012048.

[14] J. Demsar, T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, M. Mozina, M. Polajnar, M. Toplak, A. Staric, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M. Zitnik, B. Zupan, Orange: Data Mining Toolbox in Python, Journal of Machine Learning Research, 14 (2013) 2349−2353.

[15] Orange Data Mining, 2015. URL: https://orangedatamining.com/docs/

[16] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning Word Vectors for 157 Languages, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 3483–3487