Feature Selection in Text Clustering Applications of Literary Texts: A Hybrid of Term Weighting Methods

Abdulfattah Omar
College of Science and Humanities
Prince Sattam Bin Abdulaziz University, Saudi Arabia
Department of English, Faculty of Arts, Port Said University

***Abstract*—The recent years have witnessed an increasing use of automated text clustering approaches and more particularly Vector Space Clustering (VSC) methods in the computational analysis of literary data including genre classification, theme analysis, stylometry, and authorship attribution. In spite of the effectiveness of VSC methods in resolving different problems in these disciplines and providing evidence-based research findings, the problem of feature selection remains a challenging one. For reliable text clustering applications, a clustering structure should be based on only and all the most distinctive features within a corpus. Although different term weighting approaches have been developed, the problem of identifying the most distinctive variables within a corpus remains challenging especially in the document clustering applications of literary texts. For this purpose, this study proposes a hybrid of statistical measures including variance analysis, term frequency-inverse document frequency, TF-IDF, and Principal Component Analysis (PCA) for selecting only and all the most distinctive features that can be usefully used for generating more reliable document clustering that can be usefully used in authorship attribution tasks. The study is based on a corpus of 74 novels written by 18 novelists representing different literary traditions. Results indicate that the proposed model proved effective in the successful extraction of the most distinctive features within the datasets and thus generating reliable clustering structures that can be usefully used in different computational applications of literary texts.**

***Keywords*—Feature selection; frequency; PCA; term weight; text clustering; TF-IDF; variance; VSC**

I. Introduction

With the increasing access to e-texts and the availability and power of computational tools, there has been an increasing amount of humanities computing literature on text analysis and interpretation. Studies of this kind are generally classified under the broad heading computer-assisted text analysis (CATA). CATA includes numerous applications including authorship attribution, stylometric analysis, theme analysis, the use of imagery, genre classification, characterization, and textual analysis [1-4]. In spite of the effectiveness of VSC methods in resolving different problems in these disciplines and providing evidence-based research findings, the problem of feature selection remains a challenging one. For reliable text clustering applications, a clustering structure should be based on only and all the most distinctive features within a corpus. For this purpose, this study proposes a hybrid of statistical measures including variance analysis, term frequency-inverse document frequency, TF-IDF, and Principal Component Analysis (PCA) successively for selecting only and all the most distinctive features that can be usefully used for generating more reliable document clustering that can be usefully used in authorship attribution tasks. The study is based on a corpus of 74 novels written by 18 novelists representing different literary traditions.

II. Literature Review

The literature suggests that text clustering (simply putting similar texts together) is central in almost all CATA applications [5, 6]. It is used as a starting point for many of the CATA applications including thematic analysis, genre classification, stylometry, and authorship attribution [5, 7-14]. It is known that studies in these disciplines have always been done using non-computational methods. With the development of computational approaches; however, critics and researchers have come to think about how effective computational approaches are in identifying meanings within texts. Now, it is often assumed that computational approaches prove effective in better understanding texts in question [15]. This is best described as a process of decoding meanings within texts [16]. Despite the relative success of studies of this kind, they are met with a strong wave of objections from a number of critics and scholars. They still think that their success in the interpretation of texts is still far from detecting what a text is exactly about [17, 18]. This can be attributed to the unfamiliarity of the world of computational theory and methodology to literary scholars. Ramsay [19] suggests that "the inability of computing humanists to break into the mainstream of literary critical scholarship may be attributed to the prevalence of scientific methodologies and metaphors in humanities computing research" [19, P. 167]. One might even suggest that the unfamiliarity with computational and mathematical approaches has generated in literary scholars the belief that all computational and statistical approaches are somehow antithetical to literary critical approaches. This would explain the gap we see between literary critical theory on the one hand and computer-based text analysis and quantitative approaches on the other: the majority of critical theory researchers have never argued the need for using computational mathematical approaches to supplement widely

Paper Submission Date: January 19, 2020
Acceptance Notification Date: February 12, 2020

used critical approaches [20-22]. Critics of the involvement of computational methods in literary criticisms always argue that human reasoning is crucial and can never be replaced in understanding and interpreting texts. They argue that so far there is no computer-assisted system that is capable of accounting only for all the linguistic and meta-linguistic features of texts.

Defenders of computational text analysis, on the other hand, argue that the use of a computational framework in literary studies is objective, quantifiable, and methodologically consistent [23-27]. Hockey asserts that computational tools are useful adjuncts to literary criticism. She contends that without computational tools, critics have only human reading, intuition, and serendipity to use in literary criticism. Many of the defenders even go beyond that, arguing "without the computer, the interpreter is nothing more than some Romantic Aeolian harpist drowning in the phenomenological abyss of their own impressions" [19, P. 168]. This can be reflected in the significant increase in the application of computational methods in literary studies over the recent years. In numerous thematic reviews of different literary texts, text clustering is central in thematic analysis applications. This is the arrangement of texts by topic with the purpose of investigating thematic interrelationships within texts [7, 9, 14, 28, 29]. The main assumption is that text clustering methods are effective in identifying what a text is about. Consequently, thematic hypotheses can be based on clustering results. It is even argued that computational techniques are effective in generating new insights and interpretative ideas about thematic reviews of different literary texts [14, 28]. Likewise, Ramsay [13] indicates that genre classification which remained distant from computational and mathematical applications for a long time, is now making use of computation technologies and more specifically text clustering approaches to adjudicate some genre classification problems and objectively assign literary texts to appropriate genres. With the high development of text clustering algorithms and methods, genre classification studies draw more heavily on computational methods for more accurate results and better performance [13, 30-35]. Interestingly, the works of Shakespeare have been the subject of many computer-based genre classifications [13, 34, 36]. Using cluster analysis methods, Jockers classified 37 Shakespearean plays into three main clusters, comedy, history, and tragedy as shown in Fig. 1.

The literature also suggests that text clustering methods are now used in stylometry- the investigation of the quantitative properties of an author's style, and authorship attribution [33, 37-45]. The claim is that results based on computer-based methods are accepted by many as more accurate than those based on conventional non-computational methods. In spite of the potentials of computational approaches and text clustering methods especially the capacities for analyzing large quantities of data and generating results that are objective and replicable, there are still many problems and challenges with these approaches that may affect the reliability and acceptability of such methods [46-49]. One main problem is the effectiveness of text classifiers to identify and extract only and all the most distinctive features or variables within a corpus for generating clustering structures that can be usefully used in different applications. Although the issue has been extensively investigated in different disciplines including data mining and information retrieval, very little has been done in relation to the problem of feature selection in text clustering applications on literary texts. This study addresses this gap in the literature by proposing a model that combines together three statistical methods, namely variance, TF-IDF, and PCA.

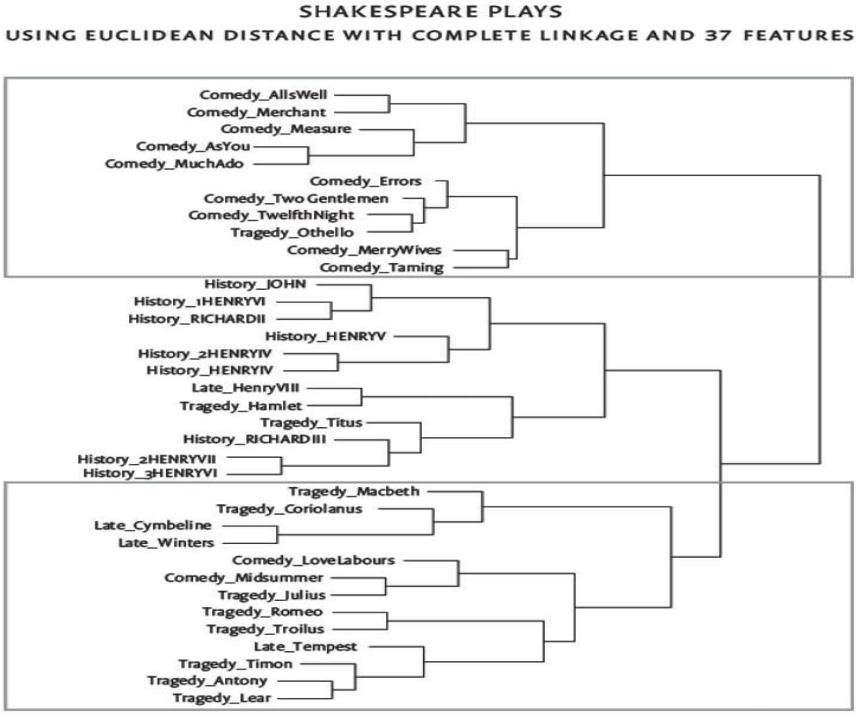


Fig. 1. Jockers' Genre Classification of 37 Shakespearean Plays.

III. METHODOLOGY

A. Methods

For the purposes of the study, an experimental study is used where different term-weighting methods are tried to develop a model that best identifies and extracts only and all the most distinctive variables within datasets. Term weighting is a pre-processing step in text clustering applications where each term is assigned its appropriate weight in all documents within a corpus with the purpose of enhancing the text clustering performance [50-52]. Term frequency is still one of the most widely used term weighting approaches in text clustering applications [53-57]. However, term frequency approaches alone are unsuitable for the text clustering of literary texts. This study experiments a combination of different term weighting methods including variance, TF-IDF, and PCA.

1). Variance: Document clustering depends on there being variation in the characteristics of interest to the research question; if there is no variation, the documents are identical and cannot be classified relative to one another [57-60]. The assumption is that variables describing the characteristics of interest are thus only useful for clustering if there is significant variation in the values they take. The intuition for variance is that if a word is used in all or most of the documents in a document collection then that word is more likely to be more important than words that do not vary considerably [53]. Accordingly, documents can be clustered according to the basis of variance. The implication is that variables of significant variation can be retained and variables with little or no variation can be removed. Although variance is an important factor in the assessment of variable importance, retaining the variables that have significant significance is not a guarantee that the data matrix is built up of the most distinctive vectors. Consequently, it should be used along with different term-weighting methods.

2). TF-IDF: TF-IDF is currently the most common method of calculating term frequency. It is widely used in information retrieval and text mining for identifying the most important variables within datasets. Numerous studies have concluded that TF-IDF works well but they do not explain why this happens [51, 59, 61-64]. The development of IDF came at the hands of Karen Spärck Jones in 1972 with the publication of her article "A statistical interpretation of term specificity and its application in retrieval". Spärck Jones [65] was the first to propose the measure of term specificity and the term came to be known as Inverse Document Frequency IDF later. The underlying principle of specificity is the selection of particular terms, or rather the adoption of a certain set of effective vocabulary that collectively characterizes the set of documents. In statistical terms, specificity is a statistical property of index terms. Statistical specificity is explained in relation to term frequency. This is based on counting the number of documents in the collection being searched which contain the query [61, 65]. Given that the term frequency of a document is the number of terms it contains, specificity of a term is the number of documents to which it pertains [65]. Logically, if descriptions are longer, terms will be used more often. This may lead to the assumption that if a query is frequently repeated in a document, this document is related to the query. This assumption can be, however, falsified. Spärck Jones [65] argues that a query term that occurs in many documents is not necessarily a good discriminator, and should be given less weight than one which occurs in a few documents. Spärck Jones' specificity or inverse document frequency IDF was later coupled with term frequency where it has been extensively used in many term weighting schemes [61, 66, 67]. In TF-IDF, the most discriminant terms are the highest TF-IDF variables. This is computed by summing the TF-IDF for each query term and a high weight in TF-IDF is reached by a high term frequency in the given document and a low document frequency of the term in the whole collection of documents [51, 59, 66, 67]. The implication to document clustering is that if the highest TF-IDF variables, which are taken to be the most discriminant terms, are identified, then unimportant variables can be deleted and data dimensionality is reduced.

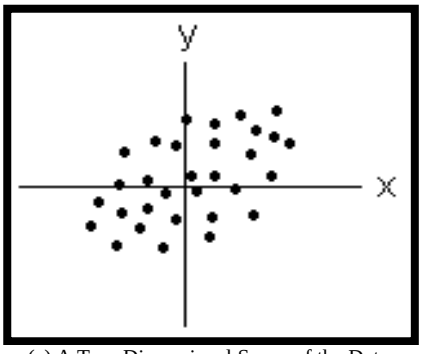
3). PCA: PCA is one of the basic geometric tools that are used to produce a lower-dimensional description of the rows and columns of a multivariate data matrix [50, 68-70]. The main function of PCA is to find the most informative vectors within a data matrix. Jolliffe [71] explains "The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data sets [71]. It can be thus described as a technique for data quality [69]. To put it simply, PCA performs two complementary tasks: (1) organizing sets of data and (2) reducing the number of variables without much loss of information. In many text clustering applications, PCA is used along with cluster analysis so that clustering is based on the most distinctive vectors within data sets. The literature suggests that PCA is used a great deal in text clustering applications prior to performing cluster analysis. The link between both cluster analysis and PCA is that both are concerned with finding patterns in data. It is sometimes advised that cluster analysis is based on PCA results so that the clustering structure is built on uncorrelated vectors. In spite of the computational mathematical nature of PCA, this section is only concerned with the idea of data reduction.

The main assumption behind PCA is that a matrix with huge data sets can be reduced so that the most distinctive vectors are identified with the purpose of best expressing the data and revealing hidden structures. Although some of the discarded or deleted variables can be important for clustering, PCA works to perform a 'good' dimensionality reduction with no great loss of information. The underlying principle of PCA is that it removes correlated variables within datasets so that it describes the covariance relationships among these variables. Fielding [72] explains that PCA "transforms an original set of variables (strictly continuous variables) into derived variables that are orthogonal (uncorrelated) and account for decreasing

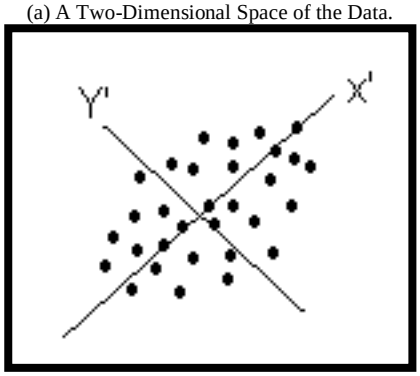
amounts of the total variance in the original set" [72, P. 16]. The process is done by means of computing the principal components scores by measuring all the variables in the data set. In so doing the variables that have the highest loading or weight are identified as principal components and other variables are discarded. The resulting principal components can then be used in subsequent analyses. Given a two-dimensional vector space with dimensions *x* and *y* shown in Fig. 2A, it is possible to transform the distribution of the data as an orthogonal linear representation as shown in Fig. 2B.

The data vector coordinates are then recalculated relative to the new basis. This has the effect of generating a highly correlated 2-dimensional vector space, as shown in Fig. 3.

Finally, the data vector coordinates are then computed on a given principal component. The variables are weighted in such a way that the resulting components account for a maximal amount of variance in the dataset. This is shown in Fig. 4.



(a) A Two-Dimensional Space of the Data.



(b) An Alternative Orthogonal Basis for Data.

Fig. 2. A Representation of a 2-Dimensional Space in Two different Ways.

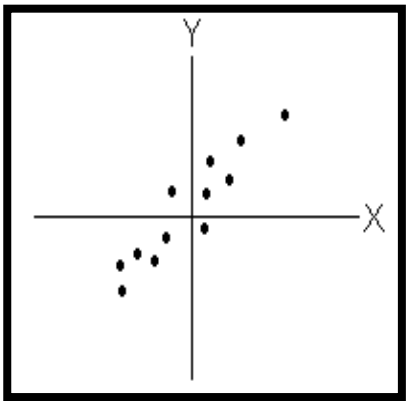


Fig. 3. A Highly Correlated 2-Dimensional Vector Space.

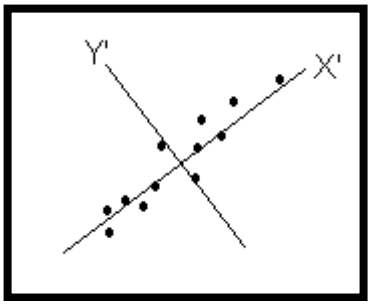


Fig. 4. Testing Variance in Data using TF-IDF.

As seen in the above figure, X' captures almost all the variation in the data, and Y' only a small amount. If Y' is simply disregarded, then the data can be restated in just one rather than the original two dimensions with minimal loss of information, and the data dimensionality has been reduced. The idea is extended to any data dimensionality. So given a data matrix of 100 rows and 1000 columns, the data matrix can be re-described in a lower number of dimensions given that there is redundancy among the variables; that is, they overlap with one another in terms of the information they present. One of the main issues in PCA, however, is determining the number of meaningful principal components (PCs).

B. Data

This is based on a corpus of 74 novels written by 18 novelists representing different literary traditions. These were alphabetically ordered and coded as shown in Table I.

C. Procedures

For text clustering purposes, a data matrix M was built. The matrix included all the 74 novels. Three pre-processing steps were carried out. First, all non-alphabetical ad punctuation marks were removed. The texts were converted into what is called bag of words (BOW). Second, stemming was carried out where only lexical types were retained. Third, texts were normalized in terms of length so that variation in text length has no negative impacts on the reliability of text clustering results. A matrix M was thus generated consisting of 74 rows (the number of texts) and 37435 vectors (all the lexical types in the texts). One major problem with this matrix is data dimensionality. That is, the matrix is composed of so many variables which makes it impossible for any text clustering system to generate reliable clustering structures. In the face of this problem, a model of three term weighting methods was proposed.

First, a variance analysis test using ANOVA was carried out for the $M_{74,\ 37435}$. It was found out that the only 1000 variables are the highest density ones. So it was decided that variables 1-1000 to be retained and variables 1001-37435 to be removed. This can be shown in Fig. 5.

Second, a TF-IDF analysis was carried out. Based on the TF-IDF test shown in Fig. 6, only the variables with the highest TF-IDF values are retained. It was decided that the highest 200 TF-IDF frequencies to be retained. So far, the Matrix is composed of only 200 variables (M74, 200).

TABLE. I. A LIST OF THE SELECTED NOVELS AND SHORT STORIES

Code	Title of the novel/short story	Author
M01	A Daughter of Isis	Nawal El-Saadawi
M02	A Portrait of the Artist as a Young Man	James Joyce
M03	A Shabby Genteel Story	Thackeray
M04	Adventures of Huckleberry Finn	Mark Twain
M05	Aisha	Ahdaf Soueif
M06	Arabian Jazz	Diana Abu Jaber
M07	Basil	Wilkie Collins
M08	Beloved	Toni Morrison
M09	Bird Summons	Leila Aboulela
M10	Birds of Paradise	Diana Abu Jaber
M11	Catherine	Thackeray
M12	Colored Lights	Leila Aboulela
M13	Daisy Miller	Henry James
M14	David Copperfield	Charles Dickens
M15	Dubliners	James Joyce
M16	Elsewhere, Home	Leila Aboulela
M17	Emma	Jane Austen
M18	Far From the Madding Crowd	Thomas Hardy
M19	God Help the Child	Toni Morrison
M20	Hard Times	Charles Dickens
M21	Home	Toni Morrison
M22	I Think of You	Ahdaf Soueif
M23	In Love and Trouble: Stories of Black Women	Alice Walker
M24	In the Eye of the Sun	Ahdaf Soueif
M25	Jude the Obscure	Thomas Hardy
M26	Lady Chatterley's Lover	D. H. Lawrence
M27	Memoirs of a Woman Doctor	Nawal El-Saadawi
M28	Meridian	Alice Walker
M29	Minaret	Leila Aboulela
M30	Mrs. Dalloway	Virginia Woolf
M31	My Name is Salma	Fadia Faqir
M32	Nisanit	Fadia Faqir
M33	Northern Abbey	Jane Austen
M34	Oliver Twist	Charles Dickens
M35	Origin	Diana Abu Jaber
M36	Orlando: A Biography	Virginia Woolf
M37	Paradise	Toni Morrison
M38	Persuasion	Jane Austen
M39	Pillars of Salt	Fadia Faqir
M40	Pride and Prejudice	Jane Austen
M41	Sandpiper	Ahdaf Soueif
M42	Sense and Sensibility	Jane Austen
M43	Song of Solomon	Toni Morrison
M44	Sons and Lovers	D. H. Lawrence
M45	Sula	Toni Morrison
M46	Tar Baby	Toni Morrison
M47	Tess of the D'Urberville	Thomas Hardy
M48	The Bluest Eye	Toni Morrison
M49	The Captain's Doll	D. H. Lawrence
M50	The Cask of Amortillado	Edgar Allan Poe
M51	The Celebrated Jumping Frog of Calaveras County	Mark Twain
M52	The Color Purple	Alice Walker
M53	The Fox	D. H. Lawrence
M54	The Glided Age	Mark Twain
M55	The Luck of Barry Lyndon	Thackeray
M56	The Map of Love	Ahdaf Soueif
M57	The Mayor of Casterbridge	Thomas Hardy
M58	The Moon Stone	Wilkie Collins
M59	The Portrait of a Lady	Henry James
M60	The Rainbow	D. H. Lawrence
M61	The Raven	Edgar Allan Poe
M62	The Tell Tale Heart	Edgar Allan Poe
M63	The Translator	Leila Aboulela
M64	The Voyage Out	Virginia Woolf
M65	The Waves	Virginia Woolf
M66	The Woman in White	Wilkie Collins
M67	To the Lighthouse	Virginia Woolf
M68	Ulysses	James Joyce
M69	Under the Greenwood Tree	Thomas Hardy
M70	Vanity Fair	Thackeray
M71	Washington Square	Henry James
M72	Willow Trees Don't Weep	Fadia Faqir
M73	Women in Love	D. H. Lawrence
M74	Zeina	Nawal El-Saadawi

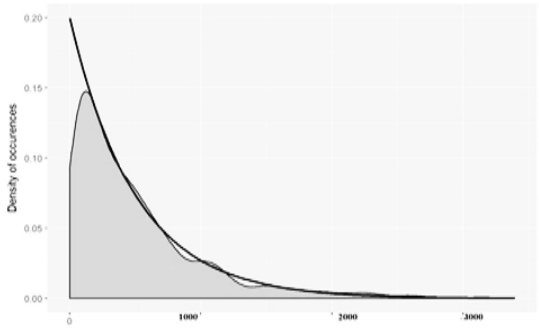


Fig. 5. Variance Analysis Test of the Matrix $M_{74, 37435}$ using ANOVA.

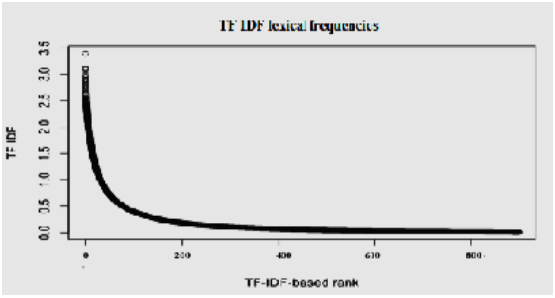


Fig. 6. TF-IDF Test of the Data Matrix $M_{74, 1000}$.

As a final step, PCA was carried out in order to extract only the most distinctive variables within the matrix $M_{74, 200}$. Based on the PCA test shown in Fig. 7, only the first 50 variables were retained. The matrix thus is reduced to only 50 variables which are thought to be the most distinctive features within the corpus.

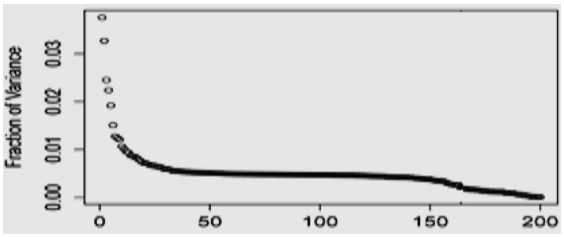


Fig. 7. A PCA of the Data Matrix $M_{74, 200}$.

IV. Analysis

In order to test the effectiveness of the proposed model, cluster analysis is used. This is a technique whereby similar texts are grouped together. The assumption is that there is a strong association between members of the same group or cluster as sharing the same characteristics. The closer texts to each other, the more similar they are and vice versa. These should be texts that can be classified under a given genre and/or written by the same author. $\mathcal{K}$-means clustering, one of the simplest and most popular cluster analysis methods, is used for the task [73-75]. In this process, every data point (the novels in our case) is assigned to the closest center or nearest mean based on their Euclidean distance. Then, new centers are calculated and the data points are updated. This process continues until there is no further iterations and changes within the clusters as seen in Fig. 8.

Using K-means clustering, the texts or data points of the matrix $M_{72, 50}$ were assigned to three groups as seen in Fig. 9. This is based on the number of centroids within the clustering structure. It should be noted, however, that the identification of the number of classes can be different from one researcher to another.

In order to validate the results of the clustering performance, hierarchical cluster analysis is used. Hierarchical clustering is as simple as $\mathcal{K}$-means clustering and it results in a clustering structure consisting of nested partitions. The results can be seen in Fig. 10.

In testing the clustering performance based on our proposed model, results of the K-means clustering are compared to those of hierarchical cluster analysis. Results indicate that there is complete agreement between the members of each cluster/group in the two clustering structures. In the two clustering structures, there are three main distinct classes. These are shown as follows.

Group 1 includes 20 texts. These are 20 novels and short stories. The most distinctive lexical features of this group are words like Islam, veil, marriage, obedience, exile, young, woman, and virginity. Texts included in this cluster are Ahdaf Soueif's Aisha, I Think of You, In the Eye of the Sun, Sandpiper, and The Map of Love; Diana Abu Jaber's Arabian Jazz, Birds of Paradise, and Origin; Fadia Faqir's My Name is Salma, Nisanit, Pillars of Salt, and Willow Trees Don't Weep; Leila Aboulela's Bird Summons, Colored Lights, Elsewhere, Home, Minartet, and The Translator; and Nawal EL-Saadawi's A Daughter of Isis, Memoirs of a Woman Doctor, and Zeina. These texts can be suggested to be belonging to a class of literature known as Anglophone Arabic literature [76-78].

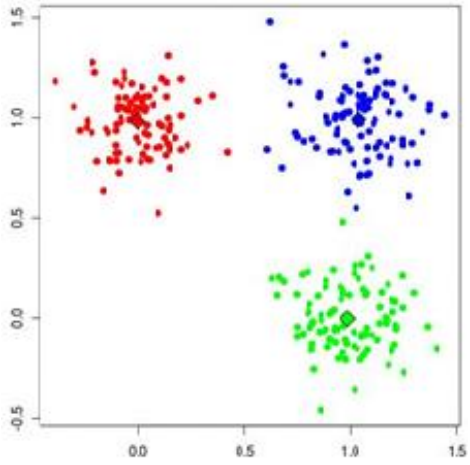


Fig. 8. K-Means Clustering.

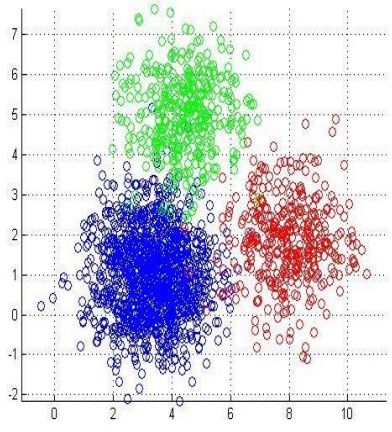


Fig. 9. K-Means Clustering of the Data Matrix $M_{74, 50}$.

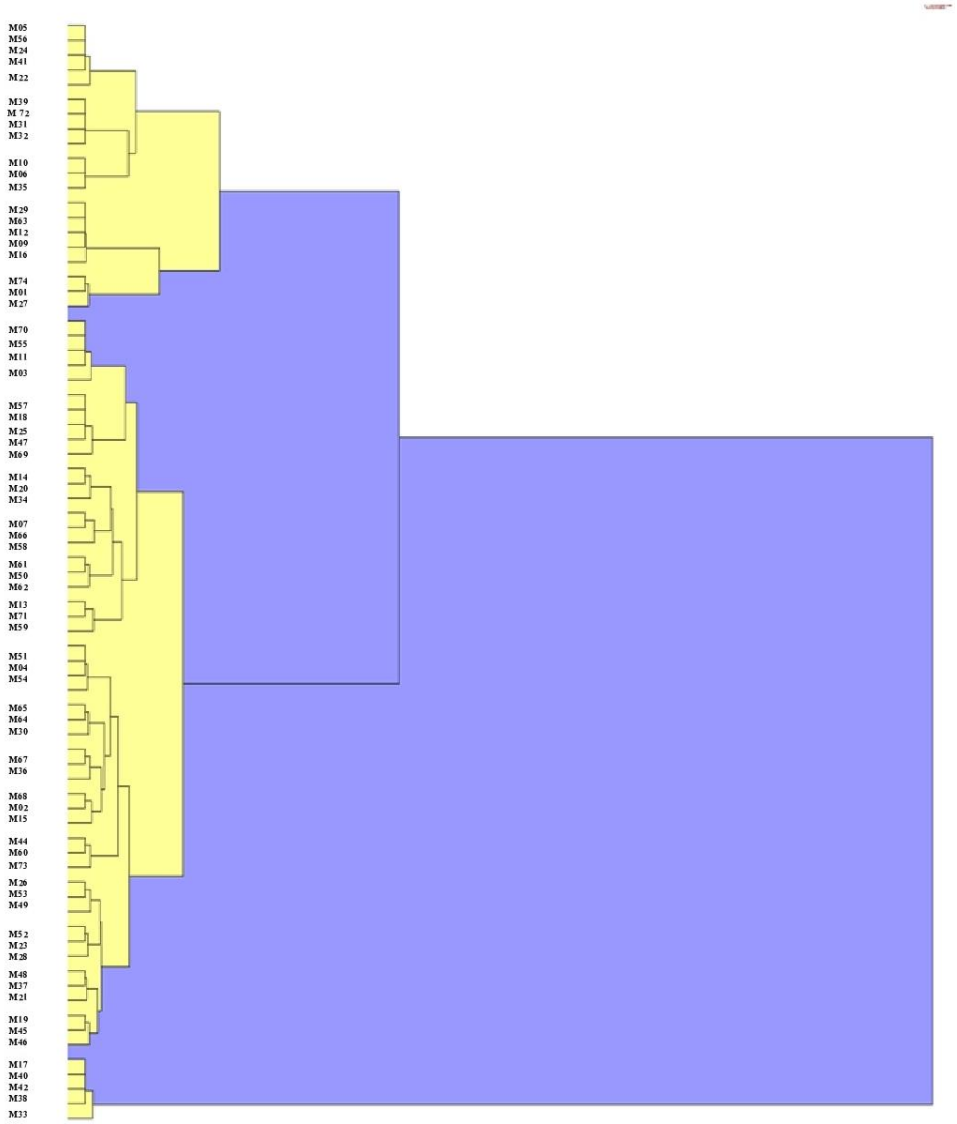


Fig. 10. A Cluster Analysis of the Data Matrix M 74, 50.

Group 2 is the biggest one as it includes 49 novels and short stories. These include Charles Dickens' Bleak House, David Copperfield, Great Expectations, Hard Times, and Oliver Twist; Thomas Hardy's Jude the Obscure, Far From the Madding Crowd, Tess of the D'Urbervilles, The Mayor of Casterbridge, and Under the Greenwood Tree; Henry James' Washington Square, D. H. Lawrence's Sons and Lovers, and Virginia Woolf's Mrs. Dalloway and The Wave. It can be seen that these texts share some features such as the portrayal of the world as we know it and the discussion of realistic problems. This cluster includes the novels that can be described as realistic novels.

Within Cluster 2, however, we can identify 4 sub-clusters or subclasses. The first subclass includes the texts written by Charles Dickens, Thomas Hardy, William Thackeray, and Wilkie Collins. These are described as social realistic novels [79, 80]. The second subclass includes the texts written by American Victorian writers Henry James, Mark Twain, and Edgar Allan Poe. Poe's texts are, however, distant from those of James and Twain as Poe is adopting a different style, the Gothic tradition, in addressing some realistic problems. The third subclass includes the novels and short stories that best described as modernist novels. These are the books written by James Joyce, D. H. Lawrence, and Virginia Woolf. These represent the modernist novels. The fourth subclass includes 11 novels. These are Toni Morrison's novels Beloved, God Help the Child, Home, Paradise, Song of Solomon, Sula, Tar Baby, and The Bluest Eye; and Alice Walker's In Love and Trouble: Stories of Black Women, Meridian and The Color Purple. These texts are similar to other members of the same group (Cluster 2) in the sense that they all address realistic problems. However, they form a distinct class by themselves as focusing more on the problems of the Black communities.

Group 3 includes only 5 novels. These are Emma, Northanger Abbey, Persuasion, Pride and Prejudice, and Sense and Sensibility. These are all written by Jane Austen and belong to the same literary tradition of what is referred to as the Romanticism [81-83]. It is also clear that the four texts

Emma, Persuasion, Pride and Prejudice, and Sense and Sensibility are very close to each other forming a subclass while Northanger Abbey represents a separate subclass. This hints that the first four texts are thematically similar to each other while Northanger Abbey has a different theme.

It is obvious that the intra-cluster similarity is high. That is, members of each group are similar to each other as the data inside each cluster is similar to one another. It is also clear that each cluster holds information that isn't similar to the other clusters. It can be claimed then that the clustering performance based on our proposed model generated a distinct structure even though different interpretations can be suggested.

V. Conclusion

This study addressed the problem of feature selection in the text clustering applications of literary texts. It proposed an integrated model for extracting the most distinctive features within datasets. The proposed model combines together three different term weighting methods: variance, TF-IDF, and PCA. In order to test the proposed model, a corpus of 74 novels and short stories was designed. Using VSC methods, the selected texts were classified into three distinct classes. It can be concluded that the proposed model is successful in extracting the most distinctive features within datasets. The findings of this study support the claim that traditional or conventional term weighting methods based solely on frequency methods are not sufficient or effective in extracting the most distinctive features within datasets. The proposed model is suggested to be usefully used in CATA applications for its high accuracy in grouping similar texts together.

Acknowledgment

I take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Scientific Deanship, for all technical support it has unstintingly provided towards the fulfillment of the current research project.

References

[1] R. Popping, Computer-assisted Text Analysis, London: SAGE, 2000.

[2] G. Wiedemann, Text Mining for Qualitative Data Analysis in the Social Sciences: A Study on Democratic Discourse in Germany. Springer Fachmedien Wiesbaden, 2016.

[3] D. N. Bengston and U. S. F. S. N. C. R. Station, Applications of Computer-aided Text Analysis in Natural Resources. U.S. Department of Agriculture, Forest Service, North Central Research Station, 2000.

[4] B. D. Hirsch, Digital Humanities Pedagogy: Practices, Principles and Politics. Open Book Publishers, 2012.

[5] B. Yu, "An Evaluation of Text Classification Methods for Literary Study," Lit Linguist Computing, vol. 23, no. 3, pp. 327-343, September 1, 2008 2008.

[6] C. Crompton, R. J. Lane, and R. Siemens, Doing Digital Humanities: Practice, Training, Research. Taylor & Francis, 2016.

[7] B. Yu and J. Unsworth, "Toward Discovering Potential Data Mining Applications in Literary Criticism," presented at the Digital Humanities, 5-9 July 2006, Paris-Sorbo, 2006.

[8] J. Unsworth, "Scholarly Primitives: What Methods do Humanities Researchers Have in Common, and How Might Our Tools Reflect This? ," Symposium on Humanities Computing: Formal Methods, Experimental Practice, King's College, London, 13 May 2000 2000.

[9] S. Argamon and M. Olsen, "Toward Meaningful Computing," Communications of ACM, vol. 49, no. 4, pp. 33-35, 2006.

[10] G. Tambouratzis and M. Vassiliou, "Employing Thematic Variables for Enhancing Classification Accuracy Within Author Discrimination Experiments," Lit Linguist Computing, vol. 22, no. 2, pp. 207-224, June 1, 2007 2007.

[11] C. Labbe and D. Labbe, "A Tool for Literary Studies: Intertextual Distance and Tree Classification," Lit Linguist Computing, vol. 21, no. 3, pp. 311-326, September 1, 2006 2006.

[12] J. Nakamura and J. Sinclair, "The World of Woman in the Bank of English: Internal Criteria for the Classification of Corpora," Lit Linguist Computing, vol. 10, no. 2, pp. 99-110, January 1, 1995 1995.

[13] S. Ramsay, "In Praise of Pattern," TEXT Technology: the Journal of Computer Text Processing, vol. 14, no. 2, pp. 177-190, 2005.

[14] T. Horton, C. Taylor, B. Yu, and X. Xiang, "'Quite Right, Dear and Interesting': Seeking the Sentimental in Nineteenth Century American Fiction," presented at the Digital Humanities, Paris-Sorbonne, France, 5-9 July 2006, 2006.

[15] G. Rockwell, "What is Text Analysis, Really?," Lit Linguist Computing, vol. 18, no. 2, pp. 209-219, June 1, 2003 2003.

[16] P. Boot, "Decoding Emblem Semantics," Lit Linguist Computing, vol. 21, no. suppl_1, pp. 15-27, January 1, 2006 2006.

[17] T. Rommel, "Literary Studies," in ACompanion to Digital Humanities, S. Schreibman, R. Siemens, and J. Unsworth, Eds. Oxford: Blackwell, 2004, pp. 88-97.

[18] T. N. Corns, "Computers in the Humanities: Methods and Applications in the Study of English Literature," Lit Linguist Computing, vol. 2, no. 2, pp. 127-130, January 1, 1987 1987.

[19] S. Ramsay, "Special Section: Reconceiving Text Analysis: Toward an Algorithmic Criticism," Lit Linguist Computing, vol. 18, no. 2, pp. 167-174, June 1, 2003 2003.

[20] T. W. Machan, "Late Middle English Texts and the Higher and Lower Criticisms," in Medieval Literature: Texts and Interpretation. Medieval and Renaissance Texts and Studies, T. W. Machan, Ed. New York: Binghamton, 1991, pp. 3–16.

[21] R. Siemens, "A New Computer-assisted Literary Criticism?," Computers and the Humanities, vol. 36, no. 3, pp. 259-267, 2002.

[22] R. Cohen, The Future of literary theory. New York: Routledge, 1989, pp. xx, 445 p.

[23] S. M. Hockey, Electronic Texts in the Humanities: Principles and Practice. Oxford: Oxford University Press, 2000, pp. xii, 216 p.

[24] M. Terras, J. Nyhan, and E. Vanhoutte, Defining Digital Humanities: A Reader. Taylor & Francis, 2016.

[25] T. H. Howard-Hill, Literary Concordances: A Complete Handbook for the Preparation of Manual and Computer Concordances. Elsevier Science, 2014.

[26] M. L. Jockers, Macroanalysis: Digital Methods and Literary History. University of Illinois Press, 2013.

[27] N. Dershowitz and E. Nissan, Language, Culture, Computation: Computing - Theory and Technology: Essays Dedicated to Yaacov Choueka on the Occasion of His 75 Birthday (no. pt. 1). Springer Berlin Heidelberg, 2014.

[28] C. Plaisant, J. Rose, and B. Yu, "Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces," presented at the Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06), Chapel Hill, North Carolina, 11-15 June 2006, 2006.

[29] R. Horton, M. Olsen, G. Roe, and R. Voyer, "Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclope´die," presented at the Digital Humanities,. Urbana-Champaign, Illinois, 2-8 June 2007, 2007.

[30] Z. Xiao and A. McEnery, "Two Approaches to Genre Analysis: Three Genres in Modern American English," Journal of English Linguistics, vol. 33, no. 1, pp. 62-82, March 1, 2005 2005.

[31] M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically Categorizing Written Texts by Author Gender," Lit Linguist Computing, vol. 17, no. 4, pp. 401-412, November 1, 2002 2002.

[32] M. Wolters and M. Kirsten, "Exploring the Use of Linguistic Features in Domain and Genre Classification," presented at the Proceedings of the

ninth conference on European chapter of the Association for Computational Linguistics, Bergen, Norway, 1999.

[33] D. I. Holmes, "The Evolution of Stylometry in Humanities Scholarship," Lit Linguist Computing, vol. 13, no. 3, pp. 111-117, September 1, 1998 1998.

[34] M. L. Jockers. (2009, 16 March 2010). Machine-Classifying Novels and Plays by Genre. Available: https://www.stanford.edu/~mjockers/cgi-bin/drupal/node/27.

[35] B. Kessler, G. Numberg, and H. Schtze, "Automatic Detection of Text Genre," presented at the Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Madrid, Spain, 1997.

[36] S. Ramsay, "Algorithmic Criticism," in A companion to digital literary studies, vol. A companion to digital literary studies, R. G. Siemens and S. Schreibman, Eds. no. Blackwell companions to literature and culture) Malden, MA: Blackwell Publishers, 2007, pp. xx, 620 p.

[37] J. F. Burrows, "Modal Verbs and Moral Principles: An Aspect of Jane Austen's Style," Lit Linguist Computing, vol. 1, no. 1, pp. 9-23, January 1, 1986 1986.

[38] J. F. Burrows, "Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style," Literary and Linguistic Computing, vol. 2, pp. 60-71, 1987.

[39] J. F. Burrows, Computation into criticism : a study of Jane Austen's novels and an experiment in method. Oxford: Clarendon, 1987, pp. xii,255p.

[40] J. F. Burrows, "'An ocean where each kind. . .': Statistical analysis and some major determinants of literary style," Computers and the Humanities, vol. 23 (4), no. 4, pp. 309-321, 1989.

[41] R. A. J. Matthews and T. V. N. Merriam, "Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher," Lit Linguist Computing, vol. 8, no. 4, pp. 203-209, January 1, 1993 1993.

[42] M. Q. Patton, Qualitative Research & Evaluation Methods, 3rd ed ed. London: Sage, 2002, pp. xxiv, 598, [65].

[43] R. S. Forsyth and D. I. Holmes, "Feature-finding for test classification," Lit Linguist Computing, vol. 11 (4), no. 4, pp. 163-174, December 1, 1996 1996.

[44] D. I. Holmes, "Authorship Attribution," Computers and the Humanities, vol. 28, pp. 87-106, 1994.

[45] D. I. Holmes and R. S. Forsyth, "The Federalist Revisited: New Directions in Authorship Attribution," Lit Linguist Computing, vol. 10, no. 2, pp. 111-127, January 1, 1995 1995.

[46] M. W. A. Smith, "Shakespeare, Stylometry and "Sir Thomas More"," Studies in Philology, vol. 89, no. 4, pp. 434-444, 1992.

[47] M. W. A. Smith, "An investigation of Morton's method to distinguish Elizabethan playwrights," Comput. Hum., vol. 19, no. 1, pp. 3-21, 1985.

[48] C. Delcourt, "About the statistical analysis of co-occurrence," Computers and the Humanities, vol. 26, no. 1, pp. 21-29, 1992.

[49] T. Sing, S. Siraj, R. Raguraman, P. Marimuthu, and K. Nithiyananthan, "Cosine similarity cluster analysis model based effective power systems fault identification," International Journal of Advanced and Applied Sciences, vol. 4, no. 1, pp. 123-130, 2017.

[50] M. W. Berry, Survey of Text Mining: Clustering, Classification, and Retrieval. Springer New York, 2013.

[51] I. Zelinka, P. Vasant, V. H. Duy, and T. T. Dao, Innovative Computing, Optimization and Its Applications: Modelling and Simulations. Springer International Publishing, 2017.

[52] S. Sirmakessis, Text Mining and its Applications: Results of the NEMIS Launch Conference. Springer Berlin Heidelberg, 2012.

[53] T. Jo, Text Mining: Concepts, Implementation, and Big Data Challenge. Springer International Publishing, 2018.

[54] C. C. Aggarwal and C. X. Zhai, Mining Text Data. Springer New York, 2012.

[55] K. L. Du and M. N. S. Swamy, Neural Networks and Statistical Learning. Springer London, 2019.

[56] C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms and Applications. CRC Press, 2018.

[57] R. Nisbet, G. Miner, and K. Yale, Handbook of Statistical Analysis and Data Mining Applications. Elsevier Science, 2017.

[58] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer New York, 2010.

[59] S. M. Weiss, N. Indurkhya, and T. Zhang, Fundamentals of Predictive Text Mining. Springer London, 2015.

[60] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery, Model-Based Clustering and Classification for Data Science: With Applications in R. Cambridge University Press, 2019.

[61] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," Journal of Documentation, vol. 60, no. 5, pp. 503-520, 2004.

[62] D. H. Kraft, E. Colvin, and G. Marchionini, Fuzzy Information Retrieval. Morgan & Claypool Publishers, 2017.

[63] B. Mitra and N. Craswell, An Introduction to Neural Information Retrieval. Now Publishers, 2018.

[64] M. Gopal, Applied Machine Learning. McGraw-Hill Education, 2019.

[65] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval " Journal of Documentation, vol. 28, pp. 11-21, 1972.

[66] G. Salton and C. Buckley, "Term-weighing approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513-523, 1988.

[67] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval," Cornell University1987.

[68] W. Härdle and L. Simar, Applied multivariate statistical analysis. Berlin ; New York: Springer, 2003, p. 486 p.

[69] J. E. Jackson, A user's guide to principal components (Wiley series in probability and mathematical statistics. Applied probability and statistics). New York: Wiley, 1991, pp. xvii, 569.

[70] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R. Springer New York, 2013.

[71] I. T. Jolliffe, Principal component analysis, 2nd ed. ed. (Springer series in statistics). Berlin ; London: Springer, 2002, p. 500 p.

[72] A. Fielding, Cluster and Classification Techniques for the Biosciences. Cambridge, UK ; New York: Cambridge University Press, 2007, pp. xii, 246 p.

[73] A. Khan, S. Baseer, and S. Javed, "Perception of students on usage of mobile data by K-mean clustering algorithm," International Journal of Advanced and Applied Sciences, vol. 4, no. 2, pp. 17-21, 2017.

[74] P. Kaur, S. Singla, and S. Singh, "Detection and classification of leaf diseases using integrated approach of support vector machine and particle swarm optimization," International Journal of Advanced and Applied Sciences, vol. 4, no. 8, pp. 79-83, 2017.

[75] Z. Ullah, S. Lee, and M. Fayaz, "Enhanced feature extraction technique for brain MRI classification based on Haar wavelet and statistical moments," International Journal of Advanced and Applied Sciences, vol. 6, no. 7, pp. 89-98, 2019.

[76] L. Maleh and L. A. Maleh, Arab Voices in Diaspora: Critical Perspectives on Anglophone Arab Literature. Rodopi, 2009.

[77] G. Nash, The Anglo-Arab Encounter: Fiction and Autobiography by Arab Writers in English. Peter Lang, 2007.

[78] Z. Halabi, Unmaking of the Arab Intellectual: Prophecy, Exile and the Nation. Edinburgh University Press, 2017.

[79] E. Freedgood, Worlds Enough: The Invention of Realism in the Victorian Novel. Princeton University Press, 2019.

[80] D. David, D. Deirdre, P. E. E. D. David, and C. U. Press, The Cambridge Companion to the Victorian Novel. Cambridge University Press, 2001.

[81] C. Lamont and M. Rossington, Romanticism's Debatable Lands. Palgrave Macmillan UK, 2007.

[82] S. Ailwood, Jane Austen's Men: Rewriting Masculinity in the Romantic Era. Taylor & Francis, 2019.

[83] M. Ferber, Romanticism: A Very Short Introduction. OUP Oxford, 2010.