Towards Computational Models to Theme Analysis in Literature

Abdulfattah Omar
Department of English
College of Sciences and Humanities
Prince Sattam Bin Abdulaziz University

***Abstract*—The recent years have witnessed the development of numerous computational methods that have been widely used in humanities and literary studies. In spite of their potentials of such methods in terms of providing workable solutions to different inherent problems within these domains including selectivity, objectivity, and replicability, very little has been done on thematic studies in literature. Almost all the work is done through traditional methods based on individual researchers' reading of texts and intuitive abstraction of generalizations from that reading. These approaches have negative implications to issues of objectivity and replicability. Furthermore, it is challenging for such traditional methods to deal effectively with the hundreds of thousands of new novels that are published every year. In the face of these problems, this study proposes an integrated computational model for the thematic classifications of literary texts based on lexical clustering methods. As an example, this study is based on a corpus including Thomas Hardy's novels and short stories. Computational semantic analysis based on the vector space model (VSM) representation of the lexical content of the texts is used. Results indicate that the selected texts were thematically grouped based on their semantic content. It can be claimed that text clustering approaches which have long been used in computational theory and data mining applications can be usefully used in literary studies.**

Keywords*—Computational models; computational semantics; lexical clustering; lexical content; philological methods; Thomas Hardy; Vector Space Model (VSM)

I. Introduction

An important development in literary studies over the past few decades is the increasing application of scientific methods in analysis of literary works [1-5]. It has been argued that the use of such scientific methods can assist in preventing the formation of false theories of criticism and the generation of unreliable thematic classifications [6, 7]. The present study is intended as a contribution to that development. The study seeks to propose a computational model that helps readers and critics of literary texts in an objective, replicable, therefore scientific way through exploring the thematic relationships of texts in a conceptually coherent way. The study is based on the novels and short stories of Thomas Hardy as an example. Thomas Hardy is one of the most important figures in the history of the English novel and he has ever sustained readers' interest in the themes and topics he tackled. Thomas Hardy was a Victorian poet and novelist and is considered by many critics as a main component of the English cultural heritage [8-10].

In spite of the proliferation of computational technology and the articulation of an explosive production of electronically encoded information of all kinds, computational methods have been very little used in humanities in general and literature in particular [11-13]. The wide cultural gap between the literary critic and computational research communities is the most obvious reason. The study is an attempt towards bridging the gap between traditional literary criticism and computational methods. The study employs experimentally replicable data representation and clustering methods. The greatest advantage of these methods is that they are completely objective in the sense that the results obtained are independent of the person applying the method.

The remainder of this article is organized as follows. Section 2 is a brief survey of the approaches to thematic studies of literary works. Section 3 outlines the methods and procedures of the study. It describes document clustering methods are used to classify the selected works in a thematically coherent way. Section 4 reports the results of the proposed methods and explores the thematic interrelationships between the texts. Section 5 is conclusion. It summarizes the main findings and suggests propositions that may be generalized to other literary texts and genres.

II. Literature Review

Theme analysis of literary texts is one of the oldest and most established disciplines in literary studies. Critics have been generally concerned with identifying the themes within literary texts. It was thought that part of the critic's job is to understand the deep meanings conveyed by authors, make observations about literary texts in order to construct the expression of themes in these works [14-16].

Although theme analysis is very old in literature studies, the issue of the way themes are defined is still controversial in literary criticism. There is no single agreed upon approach to theme analysis in literature. Over the years, there is no consensus among critics on the best ways of interpreting texts and deriving thematic concepts. It is true to claim that theme analysis is still controversial and problematic in literary criticism studies [17].

With the development of different literary theories including Marxism, Modernism, and feminism, theme analysis has been widely considered a reflection of these theories [18, 19]. Critics have been more concerned with identifying the relationship between author and work as reflected on themes of

race, class, and gender. In this way, thematic concepts of novelists and authors are usually confined and restricted to the critic's engagement with a given theory or selections from a text or some texts.

The issue of theme analysis with its complexities and controversies has its implications to the thematic studies of Thomas Hardy's literary texts. The thematic classification of Hardy's prose fiction ranges from a broad general classification of his novels and short stories to a discussion of a single thematic aspect in one, some, or all his writings [10, 20-25]. The main observation about almost the critical studies on the thematic structures of Hardy's work is that critics have been generally concerned with what Hardy himself classified as Major Works. Despite the rich thematic concepts exhibited in Hardy's prose fiction works exhibit rich thematic concepts, the majority of the thematic discussions of Hardy have been flawed in limiting their discussions to the series of novels and short stories he wrote between 1871 and 1895 [26].

It can be claimed thus that the work on Thomas Hardy's prose fiction is widely selective. Some critics focus on what is referred to Wessex novels. They think of Hardy's works as a cry for the lost beauty of the English countryside. Evidently, many commentators have characterized Hardy as a regional novelist, attribute this regionalist focus to his fascination with Wessex, an old English kingdom covering an area that provided the fictionalised setting for Hardy [27, 28]. Balanced against this argument, others insist that Hardy was a Victorian social critic since his writings depict the sufferings of England's working class and society's responsibility for their tragic fates. Through this process, Hardy is seen to have been preoccupied with improving conditions in society. These concerns mark Hardy out as a realistic writer who took on the role of expressing the joys and woes of the victims of the merciless harshness of their lives [25, 29].

One major problem with studies in this tradition is that they ignore much of the thematic richness in Hardy's works. In the face of this limitation, this study suggests the use of empirical approaches and new technologies. These should have the impact of developing a comprehensive and more detailed structuring of Hardy's thematic concepts.

III. Methods and Procedures

For developing a computational model for deriving taxonomies of thematic concepts in literary texts, document clustering theory is adopted. Document clustering theory has been widely used in data mining and information retrieval (IR) applications [30, 31]. Document clustering methods are generally used for grouping similar texts together [32, 33]. The hypothesis is that texts grouped together are more likely to have the same theme [34-36]. Document clustering methods have been proved effective in grouping and categorizing unstructured text data and exploring. In such processes, similar texts are separated together in distinct groups or clusters. Accordingly, document clustering methods can be usefully used in the domains of theme analysis in literary studies.

There are numerous document clustering methods. For the purposes of the study, vector space clustering is used. VSC is one of the earliest computer-based clustering methods [33, 37]. It is thought however that it is appropriate for the study. The rationale is that the study is concerned with build thematic structures of the texts based on their lexical semantics. VSC is thus appropriate for the purposes of the study. In VSC, documents can be grouped into distinct classes based on their lexical content [30, 38, 39]. In this regard, it is assumed that VSC is appropriate for the purposes of the study. VSC is used for organizing the novels and short stories of Thomas Hardy into distinct classes based on the lexical content of these texts. Herein, lexical clustering (one of VSC methods) is used.

Conventional lexical-clustering algorithms treat text fragments as a mixed collection of words, with a semantic similarity between them calculated based on the term of how many the particular word occurs within the compared fragments. Whereas this technique is appropriate for clustering large-sized textual collections, it operates poorly when clustering small-sized texts such as sentences [40].

In so doing, a corpus including all the selected texts is designed. The tradition of building a corpus for text clustering applications has always been based on the assumption that the corpus is both large and representative of the research domain. An important question in the context of this study is what size the corpus should be in order to support objective and reliable generalizations about Thomas Hardy's prose fiction. The corpus on which this analysis is based consists of all the known (published and unpublished) prose fiction texts of Hardy.

As a first step for data representation, the corpus was confined to what is referred to a bag of words. It was also decided that the corpus to be built of only the content words. All function words were thus removed. The hypothesis is that they do not usually carry semantic meaning; thus, they cannot be considered as distinctive features. The corpus should include only and all the distinctive features [41]. Content words can act as strong predictors of the topic(s) or content of a document [42]. Moreover, the experimental results of document classification indicate that content-word representation gives good results in identifying the content of a document and its latent structure [43-45]. Equally important, most studies seem to agree that content-word representation has been proved to give much better results than any other approach to clustering. This study considers content words to be indicators of semantic content. In other words, the analysis identifies all the morphological variants of a given stem as just one lexical type. It can be observed that variant word forms with similar semantic conceptions can be treated as equivalent. To take an example, the words 'marry', 'marries', 'married', and 'marriage' deal with a single semantic concept, which is necessarily different from, for example, *dogs* and *cats*. The analysis thus reduces all these variant forms to just one form, i.e. *marry*.

In order for the texts to be amenable for computational analysis, texts were mathematically represented using vector space model (VSM). The reason is that it is conceptually simple as well as it is convenient for computing semantic similarity within documents.

A data matrix was created including all the 62 selected texts and lexical types (45,298 variables) included in this study.

An initial observation about the corpus is that the 62 texts vary substantially in size, ranging from 002 Kb to 389 Kb. One major problem with a corpus of the kind is that documents will be clustered based on size rather than lexical content and semantic similarity. The row vectors of were thus normalized to compensate for variations in length. Mean document length method was used for the purpose. This had the effect that the documents were equally represented in the matrix and their lexical frequency profiles could be meaningfully clustered.

In order to extract the most distinctive lexical variables, term frequency inverse document frequency (TFIDF) was used. Based on the TFIDF of the Hardy matrix, the highest 200 variables were decided to be the most distinctive lexical features within the corpus. It was also clear that the texts are best categorized into four distinct classes as seen in Fig. 1.

Cluster analysis was then used to find meaningful clusters in the data. Cluster analysis was used to generate a centroid-based lexical clustering structure that captures and describes the semantic similarities of the selected texts in the data matrix based on the lexical resource. The hypothesis is that texts grouped together should have common thematic features and based on the lexical semantic properties of the variables of these texts; it is easy to assume the recurrent themes in these texts. For visualization of the clustering structure, hierarchical cluster analysis is used. Hierarchical cluster analysis is one of the main statistical approaches that is used for finding distinct classes or groups based on the shared and common features. Despite the development of different clustering algorithms, hierarchical cluster analysis or hierarchical clustering remains one of the most widely unsupervised clustering algorithms in clustering applications to find discrete groups with varying degrees of (dis)similarity in a data set represented by a (dis)similarity matrix [46]. The selected texts fall into four main clusters as shown in Fig. 2.

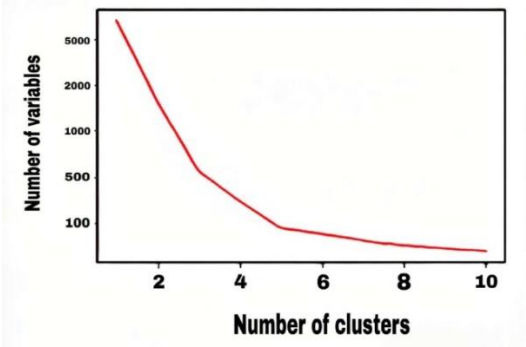

Fig. 1. A TFIDF Analysis of the Matrix H62, 200.

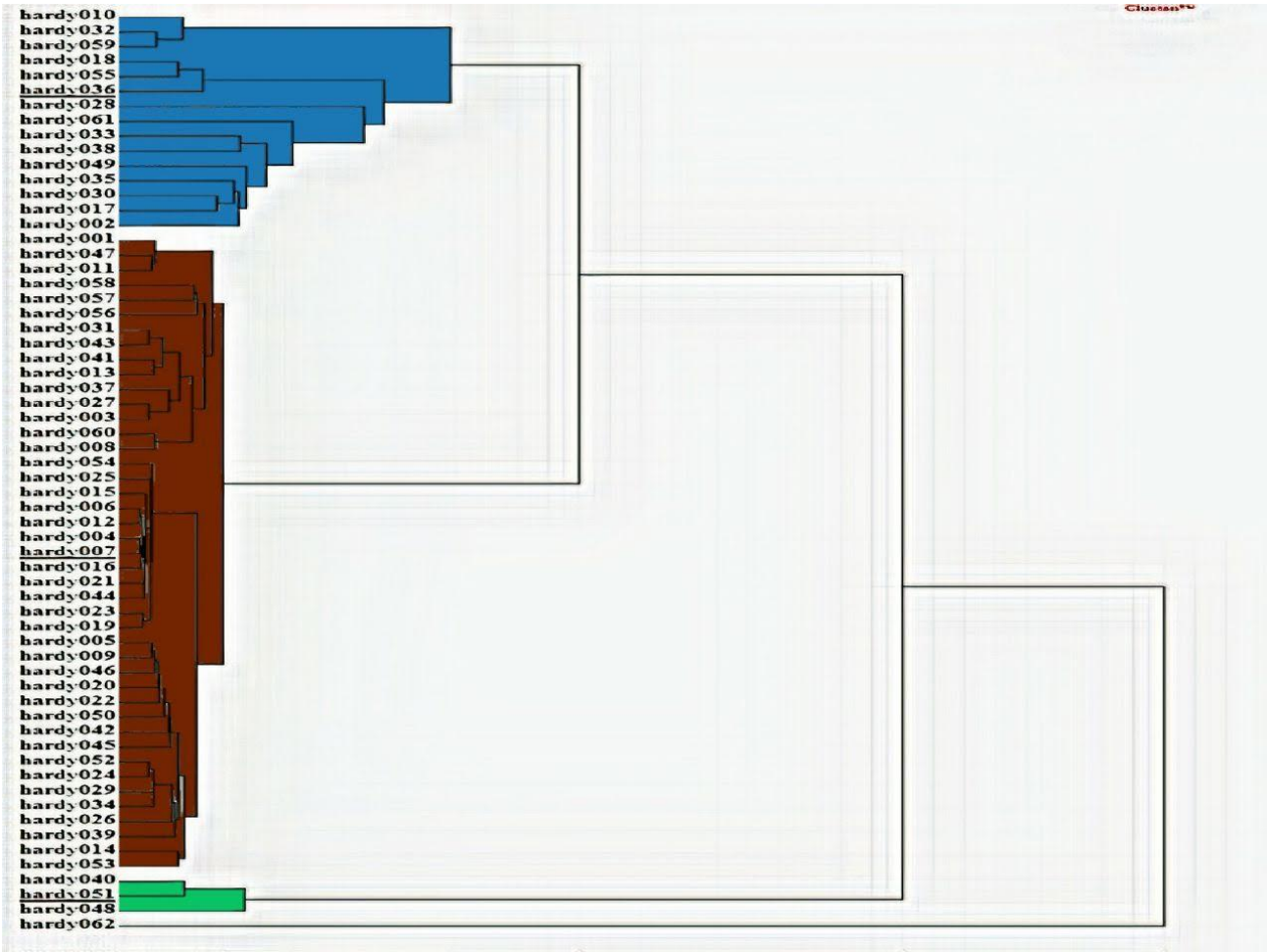

Fig. 2. A Clustering Structure of Hardy's Matrix using Euclidean Distance and Ward Linkage Clustering.

The result is a centroid-based lexical-clustering structure that can be used in any application in which the relationship between patterns is expressed in terms of pairwise semantic similarities [40]. In our case, this clustering structure is used for building hypotheses and making generalizations about the thematic relations of the texts in each group or cluster.

For validity purposes, PCA was used. The validity of cluster analysis results is an important requirement since different cluster structures may lead to completely different interpretations of the same data and thus generating contradictory hypotheses about the data. The purpose of clustering validation here thus is to see whether the same analytical methods applied to an alternative representation of the data gives identical or at least similar results. The alternative data representation was generated by principal component analysis, a dimensionality reduction method whereby H62, 200 was reduced to a dimensionality of 61, yielding the matrix H62, 50 [47, 48]. The result is that there is a full agreement between the results of the clustering structures.

In order to identify the most distinctive lexical features of each group, the columns of the matrix were rearranged in order of descending variance. Centroid vectors for the clusters A, B, C, and D were constructed by taking the means of the vectors in the matrix that constitute Groups A, B C, and D in accordance with the function

$$V_i = \frac{\sum_{i=1\ldots m} Hij}{m})$$

Where

V*j* is the *j*th element of the centroid vector (for *j* = 1 …the number of columns in H),

H is the data matrix, and

m is the number of row vectors in the cluster in question

The resulting vectors Group A $_{centroid}$, Group B $_{centroid}$, Group C $_{centroid}$ and Group D $_{centroid}$ were compared to show how, on average, the three groups differ on each of the extracted lexical variables, the aim being to identify the variables on which they differ most and thereby the thematic characteristics of each group can be inferred.

IV. Analysis and Discussions

The aim in this section is to see whether the clustering structures thus far validated are meaningful. Given that the texts were clustered on the basis of lexical frequency vectors, this implies that each cluster has a characteristic lexical frequency profile which distinguishes it from the others [49]. By doing so, it should be possible to identify the most important variables for each group, and, on the basis of the lexical semantics of these items, to infer thematic characteristics of the respective groups [50, 51].

According to the computation of the quantitative findings and an intuitive understanding of the texts, each of these groups displays the distinctive lexical variables that make them thematically distinct. The frequent use of words like ‘duke’, ‘baron’, ‘duchess’, ‘knight’, ‘estate’, and ‘squire’ in Group A is a good indication that this group is particularly concerned with aristocratic life and class differences. It can be suggested that this group touches on many aspects of class difference, adventure, romance, matrimony and mismatched unions and the conflicts they bring. Parallel to these themes, the Napoleonic era appears as a recurring theme in many of the texts of this group, as reflected by the frequent use of words like ‘Napoleon’, ‘France’, ‘French’, and ‘*war*’. This quantitative finding is supplemented by an intuitive reading of the texts and is also supported by critical assessments of the texts included here. Gilmartin and Mengham [52] argue that The Poor Man and the Lady and A Group of Noble Dames feature one of the most recurrent themes in Hardy’s books: that of cross-class relationships or marriages. The texts included here discuss issues of elopement, failure in marriage, and illegitimate children. This finding also agrees in principle with Hardy’s classification of his own works since the majority of the texts included here he classified under the category of Romance and Fantasies.

The hierarchical clustering structure, which is based on pure mathematical methods, supports Hardy’s tendency to group similar short stories together. Six texts of this group are included in his volume of short stories, A Group of Noble Dames. It also includes The Doctor’s Legend, which was first collected in Noble Dames when it was published in serial form in Harper’s Weekly and the Graphic in late 1890 [53]. Purdy [54] comments that the text appeared later in the collection A Group of Noble Dames under the title Barbara, which is thematically similar to The Legend. As such, the results of this analysis agree with the thematic structure that Hardy defined for his books.

Although the texts included here can be placed under the heading of Romance and Fantasies, as Hardy classified them, the element of social criticism persists through almost all of the texts. In The Poor Man and the Lady and the stories of A Group of Noble Dames, discussion of social problems is clear. Hardy is concerned with the problem of mismatched unions in a very class-conscious society. This argument is supported by Brady [55], who writes: In its subject matter, however, A Group of Noble Dames has interesting links with Hardy’s earlier work. The book is one of his many attempts, beginning with The Poor Man and the Lady, to portray the fascination and the difficulty of sexual alliances that cross class boundaries [55]. The texts involved in this group highlight the historical development of Hardy as a novelist and it is clear that Hardy was preoccupied with social issues throughout his career as a novelist and prose writer. This is supported by Dalziel [56], who stresses the essential continuity of Hardy’s thinking on social issues from the beginning to the end of his career as a writer.

The majority of the texts in this group as a whole are thus thematically related around romance and adventure. This does not contradict, however, the inclusion of texts like A Tradition of Eighteen Hundred and Four, Anna, or A Committee Man of “The Terror”, which are all about the political upheavals that took place in England and France as a result of the French Revolution and English Civil War—the main thematic frame in the first story is adventure while in the other two stories it is romance. Gilmartin and Mengham [52] argue that in spite of

the fact that the story is concerned with the theme of English-French conflicts: "it exhibits many of the expected features of a Christmas story (being written for the annual Harper's Christmas); it is meant to give a frisson of fear to those within the story who are sheltering from the rain and cold by the inn's fireside, and also to the readers of the periodical sitting by the Christmas hearth" [52].

The largest group, Group B, includes 43 texts out of the matrix's 62 rows, and is concerned with the English countryside; domestic life (as reflected in words such as 'river', 'cabbage', 'village', 'horse', 'mare', 'farmer', 'mill', 'tub', 'heath', 'cloth', 'sky', 'vicar', 'cover', 'passage', 'stream', 'hut', 'lane', and 'rain'); and struggle, outrage, and the frustrations of the poor ('public', 'money', 'children', 'work', 'fact', 'trade', 'bureau', and 'penny'). A common theme of contemporary social life can be suggested. Nevertheless, each subclass displays characteristic thematic features. One subclass which can be defined as Group B1, for instance, is tragedy, which correlates with ideas of social promotion/hostility and struggle. This subclass includes texts referred to by many critics as Hardy's major works. These include: Far From the Madding Crowd*;* The Return of the Native*;* The Woodlanders*;* The Mayor of Casterbridge*;* Tess of the D'Urbervilles*;* Jude the Obscure; and The Trumpet Major. The texts included here reflect Hardy's sense of disdain for the fashionable world and mock the social mores of the age. The texts talk generally about heroes and heroines who aspire for a better life and their attempts to achieve social promotion; as well as how they discover the falsity of their lives. They cannot escape the miserable conditions in which they live and are destined to suffer. Fate is an important factor in their suffering. The combination of social elements with these tragedies suggests the theme of social tragedy. Love is a recurrent theme in the other subcluster. The realistic representation, however, is always there. This is represented in texts such as The Romantic Adventures of a Milkmaid and The Trumpet-Major. Given that the texts represent different historical stages of Hardy's career, it may be claimed that the social element is heavily emphasized from the beginning of his career as a novelist up until he gave up writing novels. Unlike Hardy's classification of his own works, hierarchical cluster analysis along with qualitative analysis results point to social indicators influencing his career as a novelist. The social dimension is never absent in his writing.

There is also a correlation between the texts included here and Hardy's vision of Wessex and the English countryside. Many of the texts are set in that imaginary world of Wessex, the name of an Anglo-Saxon kingdom that covered a large area of south and southwest England prior to the Norman Conquest. It may also be claimed that woman and feminist issues are central themes in the texts of this group. Thomas Hardy was keen on describing Victorian hypocrisy in relation to women's issues. *Tess* highlights the rampant sexual assault and exploitation of the age. The novel also reflects Hardy's disapproval of the Victorians' obsession with female virginity. Fanny Robin in Far from the Madding Crowd is another example. When Troy refuses to marry her and abandons her, she tries to pick up her life as best she can. Finally, she becomes unable to work and is left without any money. As a result, she and her child die of need and starvation. This offers another typical example of the suffering of women in the Victorian age.

Texts in Group C seem to form a distinct thematic relationship. The three short stories in this group, What the Shepherd Saw (Hardy048), The Duke's Reappearance (Hardy051), and The Duchess of Hamptonshire (Hardy040), are concerned with the idea of hidden or unrevealed death. This idea is repeated in the three texts where problems of jealousy and suspicion in marriage lead to death. The main idea of each of these three texts is that that there is a beautiful married woman who belongs to the elite. Her husband, as a man of high position, feels jealous and decides to take revenge against the person who he thinks to be her lover, because of the disgrace such an illicit relationship causes him. Finally, Group D includes just one text which is The Unconquerable (Hardy062). The most important variables of this group are 'book', 'linger', 'occupation', 'measure', 'copying', 'bold', 'quaint', 'style', 'architecture', 'graveyard', 'figure', 'draughtsman', 'antique', 'masonry', and 'rose'. Correlating this cluster with the bibliographical data, it emerged that the text was written by Hardy in collaboration with his wife Florence Dugdale-Hardy. This can be an indication that it has unique lexical features that makes it distinct from other texts.

On the basis of the foregrounding discussions, it can be claimed that clustering structures are meaningful. Each cluster or Group has its distinctive lexical profile that distinguishes it from other groups or clusters. It may be claimed that cluster analysis points to significant facts regarding the novels and short narratives of Hardy. This cluster analysis relates some works to each other in ways not found in the established criticisms of Hardy. In Hardy's classification of his works, The Return of the Native is classified under the category of Novels of Environment and Character, while The Hand of Ethelberta is classified under the category of Novels of Ingenuity. However, here the two texts are clustered together in Group B. The dominant realistic approach of the works of Hardy may be one reason that many critics, who have attempted the thematic classification of Hardy's work, have not thought about connections and similarities between the two texts. In our case, we suggest that these two texts are related to each other in terms of their dealing with class consciousness. In his introduction to the New Wessex Edition of The Hand of Ethelberta, for example, Gittings [57] underestimates the novel, classifying it as 'the joker in the pack' of Hardy's novels. Widdowson [58], on the other hand, insists that The Hand of Ethelberta is not merely a romance, as Hardy classified it. He argues that the novel demonstrates Hardy's concern with the issue of class consciousness. He gives evidence that the text reflects bibliographical elements of Hardy's own life and draws parallels between the narrator of the story, who takes novel writing as a means for social promotion, and Thomas Hardy himself. Widdowson [58] comments that in all his novels, especially in The Hand of Ethelberta, Hardy appears concerned with the idea of class consciousness.

Equally important, the clustering structures provide ways of classifying the novels and short stories of Hardy according to genre. The idea is that thematic classification has pointed to

tragic, historic, and fantasy elements in the texts. Consequently, as far as thematic interrelationships are concerned, it is clear that the texts exhibit obvious features for genre classification. This can be a starting point for a comprehensive genre classification of the novels and short stories of Hardy and texts can be classified under the main categories of tragedy, comedy, romance, epic, fantasy, history, and pastoral histories, etc. One advantage of such a classification is that it can narrow down the ways in which we think about them. Here, I give some examples. Those texts that critics have usually considered tragedies are included in Group B—The Woodlanders, Tess, and Jude, for instance are all included in just one group. These are modern social tragedies and in these novels, Hardy deals with the social factors that determine the tragic end of his protagonists.

V. Conclusion

This study addressed the question whether thematic concepts can be identified in literary texts using computational models. In this regard, document clustering methods were used for grouping the selected texts into distinct classes based on their semantic similarity. Results indicate clearly that document clustering methods can be usefully used for generating distinct and meaningful classes that express some thematic concepts such as class status, sex, marriage, love, romance, and the English countryside. The analytical results are objective in the sense that they are generated by mathematically based computational methods working on empirically derived data, and as such are not open to influences from any theoretical presuppositions that the researcher might have. Unlike results from the philological method, the computational results are replicable and therefore testable and scientifically respectable. This is not to overlook the subtle elaborations of literary criticism of Thomas Hardy over the past years. In point of fact, these elaborations generate a number of hypotheses which have not been empirically confirmed. Although the results of the present study largely agree with non-computational philological classifications of Hardy's texts, the contribution, however, is that the results obtained here are objective.

Acknowledgments

This project was supported by the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University under the research project No. 2020/02/11847.

References

[1] C. Mullings, S. Kenna, M. Deegan, and S. Ross, New Technologies for the Humanities. De Gruyter, 2019.

[2] G. Balossi, A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's The Waves. John Benjamins Publishing Company, 2014.

[3] I. Mani, Morgan, and C. Publishers, Computational Modeling of Narrative. Morgan & Claypool, 2013.

[4] R. Siemens and S. Schreibman, A Companion to Digital Literary Studies. Wiley, 2013.

[5] J. G. Shanahan, Y. Qu, and J. Wiebe, Computing Attitude and Affect in Text: Theory and Applications. Springer Netherlands, 2005.

[6] M. L. Jockers and R. Thalken, Text Analysis with R: For Students of Literature. Springer International Publishing, 2020.

[7] W. van Peer and S. Zyngier, Directions in Empirical Literary Studies: In Honor of Willie Van Peer. John Benjamins Publishing Company, 2008.

[8] N. Page, Oxford Reader's Companion to Hardy. Oxford University Press, 2000.

[9] M. Bevis, The Oxford Handbook of Victorian Poetry. OUP Oxford, 2013.

[10] P. Mallett and S. E. Maier, Thomas Hardy in Context. Cambridge Cambridge University Press, 2013.

[11] J. Burrows, "Textual Analysis," in A Companion to Digital Humanities, S. Schreibman, R. Siemens, and J. Unsworth, Eds. Oxford: Blackwell, 2004, pp. 88-97.

[12] M. K. Gold and L. F. Klein, Debates in the Digital Humanities. University of Minnesota Press, 2016.

[13] D. L. Hoover, J. Culpeper, and K. O'Halloran, Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama. Taylor & Francis, 2014.

[14] T. Pugh and M. E. Johnson, Literary Studies: A Practical Guide. Taylor & Francis, 2013.

[15] P. P. Headrick, The Wiley Guide to Writing Essays About Literature. Wiley, 2013.

[16] M. L. Kamil, P. B. Mosenthal, P. D. Pearson, and R. Barr, Handbook of Reading Research. Taylor & Francis, 2014.

[17] F. Mulhern, Contemporary Marxist Literary Criticism. Taylor & Francis, 2014.

[18] R. Wellek and A. Warren, Theory of Literature. Dalkey Archive Press, 2020.

[19] B. Kachuck, "Feminist Social Theories: Theme and Variations," Sociological Bulletin, vol. 44, no. 2, pp. 169-193, 1995.

[20] J. L. Bownas, Thomas Hardy and Empire: The Representation of Imperial Themes in the Work of Thomas Hardy. Taylor & Francis, 2016.

[21] R. G. Cox, Thomas Hardy; the critical heritage (Critical heritage series). New York: Barnes & Noble, 1970, pp. xlvii, 473 p.

[22] J. Dillion, Thomas Hardy: Folklore and Resistance. Palgrave Macmillan UK, 2016.

[23] R. Nemesvari, Thomas Hardy, Sensationalism, and the Melodramatic Mode. New York: Palgrave Macmillan US, 2011.

[24] P. Vigar, The Novels of Thomas Hardy: Illusion and Reality. Bloomsbury Publishing, 2014.

[25] K. Wilson, A Companion to Thomas Hardy. Wiley, 2010.

[26] K. Ireland, Thomas Hardy, Time and Narrative: A Narratological Approach to his Novels. Palgrave Macmillan UK, 2014.

[27] P. Brantlinger and W. Thesing, A Companion to the Victorian Novel. Wiley, 2008.

[28] J. Hodson, Dialect and Literature in the Long Nineteenth Century. Taylor & Francis, 2017.

[29] J. King, "Thomas Hardy: Tragedy Ancient and Modern," in Tragedy in the Victorian NovelCambridge: Cambridge University Press, 1978, pp. 97–126.

[30] W. Wu, H. Xiong, and S. Shekhar, Clustering and Information Retrieval. Springer 2013.

[31] A. N. Srivastava and M. Sahami, "Text Mining Classification, Clustering, and Applications," (Data Mining and Knowledge Discovery Series. Chapman and Hall, 2009, p.^pp. Pages.

[32] C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms and Applications. CRC Press, 2016.

[33] W. Wu, H. Xiong, and S. Shekhar, Clustering and Information Retrieval. Springer US, 2003.

[34] F. M. G. França and A. F. de Souza, Intelligent Text Categorization and Clustering. Springer Berlin Heidelberg, 2008.

[35] A. K. Somani, R. S. Shekhawat, A. Mundra, S. Srivastava, and V. K. Verma, Smart Systems and IoT: Innovations in Computing: Proceeding of SSIC 2019. Springer Singapore, 2019.

[36] G. Chakraborty, M. Pagolu, and S. Garla, Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS. SAS Institute, 2014.

[37] K. Riesen and H. Bunke, Graph Classification And Clustering Based On Vector Space Embedding. World Scientific Publishing Company, 2010.

[38] J. Kogan, Introduction to Clustering Large and High-Dimensional Data. Cambridge: Cambridge University Press, 2007.

[39] H. Moisl, Cluster Analysis for Corpus Linguistics. De Gruyter, 2015.

[40] K. Abdalgader, "Centroid-Based Lexical Clustering," in Recent Applications in Data Clustering, H. Pirim, Ed.: IntechOpen, 2018, pp. 378-403.

[41] D. Glynn and J. A. Robinson, Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy. John Benjamins Publishing Company, 2014.

[42] K. Luyckx, Scalability Issues in Authorship Attribution. UPA, 2011.

[43] M. L. Eaton, Multivariate Statistics: A Vector Space Approach (Institute of Mathematical Statistics. Lecture notes-monograph series). Beachwood, Ohio: Institute of Mathematical Statistics, 2007, pp. viii, 512 p.

[44] A. Gani, A. Siddiqa, S. Shamshirband, and F. Hanum, "A survey on indexing techniques for big data: taxonomy and performance evaluation," Knowledge and information systems, vol. 46, no. 2, pp. 241-284, 2016.

[45] T. Hofmann, "Probabilistic latent semantic indexing," in ACM SIGIR Forum, 2017, vol. 51, no. 2, pp. 211-218: ACM.

[46] TomTullis and B. Albert, Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics, Second Edition ed. Elsevier 2013.

[47] A. KASSAMBARA, Practical Guide To Principal Component Methods in R: PCA, M(CA), FAMD, MFA, HCPC, factoextra. CreateSpace Independent Publishing Platform, 2017.

[48] H. Bozdogan and A. K. Gupta, Multivariate Statistical Modeling and Data Analysis: Proceedings of the Advanced Symposium on Multivariate Modeling and Data Analysis May 15–16, 1986. Springer Netherlands, 2012.

[49] A. A. Omar, "Addressing Subjectivity in Thematic Classification of Literary Texts: Using Cluster Analysis to Derive Taxonomies of Thematic Concepts in the Thomas Hardy's Prose Fiction," Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science, vol. 1, no. 2, 2010.

[50] A. A. Omar, "Feature Selection in Text Clustering Applications of Literary Texts: A Hybrid of Term Weighting Methods," International Journal of Advanced Computer Science and Applications, vol. 11, no. 2, pp. 99-107, 2020.

[51] A. A. Omar, "On the Digital Applications in the Thematic Literature Studies of Emily Dickinson's Poetry," International Journal of Advanced Computer Science and Applications, vol. 11, no. 6, pp. 361-365, 2020.

[52] S. Gilmartin and R. Mengham, Thomas Hardy's Shorter Fiction: A Critical Study. Edinburgh: Edinburgh University Press, 2007, pp. x, 144 p.

[53] P. Dalziel, "Thomas Hardy: The Excluded and Collaborative Stories." Oxford: Clarendon Press, 1992, p.^pp. Pages.

[54] R. L. Purdy, Thomas Hardy : A Bibliographical Study. [S.l.]: Oxford University Press, 1979.

[55] K. Brady, The Short Stories of Thomas Hardy. New York: St. Martin's Press, 1982, pp. xii, 235.

[56] P. Dalziel, "Hardy's Unforgotton 'Indiscretion': The Centrality of an Uncontrolled Work," Review of English Studies, vol. XLIII, no. 171, pp. 347-366, August 1, 1992 1992.

[57] R. Gittings, "An Introduction to The Hand of Ethelberta." New York: St. Martin's Press, 1978, p.^pp. Pages.

[58] P. Widdowson, On Thomas Hardy : late essays and earlier. Basingstoke: Macmillan, 1998, pp. x, 218.

APPENDIX NO. 1: TEXTS AND NAME CODES

Title	Code	Title	Code
A Laodicean	hardy001	The First Countess of Wessex	hardy032
A Pair of Blue Eyes	hardy002	Barbara of the House of Grebe	hardy033
An Indiscretion in the Life of an Heiress	hardy003	The Marchioness of Stonehenge	hardy034
Desperate Remedies	hardy004	Lady Mottisfont	hardy035
Far from the Madding Crowd	hardy005	The Lady Icenway	hardy036
Jude the Obscure	hardy006	Squire Petrick's Lady	hardy037
Tess of the D'Urbervilles	hardy007	Anna, Lady Baxby	hardy038
The Hand of Ethelberta	hardy008	The Lady Penelope	hardy039
The Mayor of Casterbridge	hardy009	The Duchess of Hamptonshire	hardy040
The Poor Man and the Lady	hardy010	The Honourable Laura	hardy041
The Well-Beloved	hardy011	A Changed Man	hardy042
The Return of the Native	hardy012	The Waiting Supper	hardy043
The Trumpet-Major	hardy013	Alicia's Diary	hardy044
The Woodlanders	hardy014	The Grave by the Handpost	hardy045
Two on a Tower	hardy015	Enter a Dragoon	hardy046
Under the Greenwood Tree	hardy016	A Tryst At An Ancient Earthwork	hardy047
The Three Strangers	hardy017	What The Shepherd Saw	hardy048
The Three Strangers	hardy018	A Committee-Man of The Terror	hardy049
A Tradition of Eighteen Hundred and Four	hardy019	Master John Horseleigh, Knight	hardy050
The Melancholy Hussar of The German Legion	hardy020	The Duke's Reappearance	hardy051
The Withered Arm	hardy021	A Mere Interlude	hardy052
Fellow-Townsmen	hardy022	The Romantic Adventures of a Milkmaid	hardy053
Interlopers At The Knap	hardy023	How I Built Myself a House	hardy054
The Distracted Preacher	hardy024	Destiny and a Blue Cloak	hardy055
An Imaginative Woman	hardy025	The Thieves Who Couldn't Help	hardy056
The Son's Veto	hardy026	Our Exploits at West Poley	hardy057
For Conscience' Sake	hardy027	Old Mrs. Chundle	hardy058
A Tragedy of Two Ambitions	hardy028	The Doctor's Legend	hardy059
On the Western Circuit	hardy029	The Spectre of the Real	hardy060
To Please His Wife	hardy030	Blue Jimmy: The Horse Stealer	hardy061
The Fiddler of the Reels	hardy031	The Unconquerable	hardy062