Available online at www.jlls.org

JOURNAL OF LANGUAGE AND LINGUISTIC STUDIES

ISSN: 1305-578X

Journal of Language and Linguistic Studies,17(Special Issue 1),580-594; 2021

Identifying themes in fiction: A centroid-based lexical clustering approach

Abdulfattah Omar [a] [1]

[a] *Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia*

APA Citation:

Omar, A. (2021). Identifying themes in fiction: A centroid-based lexical clustering approach. *Journal of Language and Linguistic Studies, 17*(Special Issue 1), 580-594.

Submission Date:01/09/2020

Acceptance Date:11/10/2020

Abstract

In recent years, numerous computational methods have been developed that have been widely used in humanities and literary studies. In spite of the potential of such methods in providing workable solutions to various inherent problems in research within these domains, including selectivity, objectivity, and replicability, very little empirical work has been done on thematic studies in literature. Such studies are almost entirely undertaken through traditional methods based on individual researchers' reading of texts and intuitive abstraction of generalizations from their reading. This has negative implications in terms of issues of objectivity and replicability. Furthermore, there are challenges in dealing effectively with the hundreds of thousands of new novels that are published every year using traditional methods. In the face of these problems, this study proposes an integrated computational model for the thematic classification of literary texts based on lexical clustering methods. This study is based on a corpus comprising Thomas Hardy's novels and short stories. The study employs computational semantic analysis based on a vector space model (VSM) representation of the lexical content of the texts. The results indicate that the selected texts could be grouped thematically based on their semantic content. Thus, there is now evidence that text clustering approaches, which have long been used in computational theory and data mining applications, can be usefully applied in literary studies.

Keywords: computational models; computational semantics; lexical clustering; lexical content; Thomas Hardy; vector space model (VSM)

1. Introduction

There has been an important development in literary studies over the past few decades with the increasing application of scientific methods in the analysis of literary works (Balossi, 2014; Mani, 2013; Mullings, Kenna, Deegan, & Ross, 2019; Shanahan, Qu, & Wiebe, 2005; Siemens & Schreibman, 2013; Zyngier, 2008). It has been argued that the use of such scientific methods can inhibit the development of spurious theories of criticism and the generation of unreliable thematic classifications (Jockers & Thalken, 2020). This study is intended as a contribution to this methodological development. The study

[1] Corresponding author.
E-mail address: a.a.omar2010@gmail.com

seeks to propose a computational model that can help readers and critics address literary texts in an objective, replicable, and therefore scientific way by exploring the thematic relationships in texts in a conceptually coherent manner. To exemplify the use of this computational model, it is applied to the novels and short stories of Thomas Hardy, one of the most important figures in the history of the English novel whose works have continued to sustain readers' interest in the themes and topics he tackled over the years. Thomas Hardy was a Victorian poet and novelist and is considered by many critics to be a major part of the English cultural heritage (Bevis, 2013; Mallett & Maier, 2013; Page, 2000).

In spite of the proliferation of computational technology, and indeed an explosion in the production of electronically encoded information of all kinds, computational methods have been used very little in the humanities in general and in literature in particular (Gold & Klein, 2016; Hoover, Culpeper, & O'Halloran, 2014). The sizeable cultural gap between the literary critic and computational research communities is the most obvious reason. This study is an attempt to bridge 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 that the results obtained are independent of the person applying them.

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

2. 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, and make observations about literary texts in order to construct the expression of themes in these works (Headrick, 2013; Pugh & Johnson, 2013).

Although thematic analysis has long been used in literature studies, the issue of how themes are defined is still controversial in literary criticism. Over the years, there has been no consensus among critics on the best ways of interpreting texts and deriving thematic concepts, and there is no single agreed approach to thematic analysis in literature. Therefore, it is still the case that thematic analysis is controversial and problematic in literary criticism studies (Mulhern, 2014).

Thematic analysis has been widely considered to be a reflection of the development of different literary theories, including Marxism, modernism, and feminism (Kachuck, 1995; Wellek & Warren, 1963). Critics have primarily been concerned with identifying the relationship between the author and work as reflected in themes of race, class, and gender. Thus, exploration of the thematic concepts developed by novelists and authors is usually confined and restricted to the critic's engagement with a given theory or selections from a text or some texts.

The issue of thematic analysis, with its complexities and controversies, has implications for the thematic study 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 discussion of a single thematic aspect in one, some, or all his writings (Bownas, 2012; Cox, 1970; Dillion, 2016; Mallett & Maier, 2013; Nemesvari, 2011; Vigar, 2014; Wilson, 2010). The main observation concerning almost all the critical studies on the thematic structures of Hardy's work is that critics have generally been concerned with what Hardy himself classified as "major works." Although Hardy's prose fiction exhibits 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 (Ireland, 2014).

It can thus be claimed that works exploring Thomas Hardy's prose fiction have been highly selective. Some critics have focused on what are referred to as the 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, attributing this focus to his fascination with Wessex, an old English kingdom covering an area that provided the fictionalized setting his most renowned works (Brantlinger & Thesing, 2008; Hodson, 2017). 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 as a realistic writer who took on the role of expressing the joys and woes of the working class as victims of the merciless harshness of their lives (King, 1978; Wilson, 2010). 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 make it possible to develop a comprehensive and more detailed structuring of Hardy's thematic concepts.

3. Methods

The study employed document clustering theory to develop a computational model that could lead to the derivation of taxonomies of thematic concepts in literary texts. Document clustering theory has been widely used in data mining and information retrieval (IR) applications (Srivastava & Sahami, 2009; Wu, Xiong, & Shekhar, 2013). Document clustering methods are generally used for grouping similar texts (Aggarwal & Reddy, 2016; Wu, Xiong, & Shekhar, 2013), based on the hypothesis that texts grouped together are more likely to share themes (Chakraborty, Pagolu, & Garla, 2014; França & de Souza, 2008; Somani, Shekhawat, Mundra, Srivastava, & Verma, 2019). Such methods have proved effective in exploring, grouping, and categorizing unstructured text data. Using such processes, similar texts are separated out into distinct groups or clusters. Thus, document clustering methods can usefully be employed in the domain of thematic analysis in literary studies.

As a response to the growing overflow of information which has made it difficult for many search engines to fill people's needs, various computer-based clustering methods have been developed. The standard approach and the most widely used in statistical text clustering applications, however, is clustering by content. This is the clustering of documents by the words they contain. Content clustering is carried out by means of computing semantic similarity/ distance or what can be called measuring proximity within documents. It is thus a lexical semantic function. It has always been argued that semantic information within documents is key to understanding and determining the content of such documents.

3.1. Instrument

There are numerous document clustering methods. For the purposes of the study, vector space clustering (VSC) was used. VSC is one of the earliest computer-based clustering methods (Riesen & Bunke, 2010; Wu et al., 2003) but was considered appropriate for the study as it was concerned with building the thematic structures of the texts based on their lexical semantics. In VSC, documents can be grouped into distinct classes based on their lexical content (Kogan, 2007; Moisl, 2015; Wu et al., 2013), as was done in this study using a corpus of the novels and short stories of Thomas Hardy, i.e., following a method of lexical clustering.

Conventional lexical-clustering algorithms treat text fragments as a mixed collection of words, with the semantic similarity between them calculated based on how many times the particular word occurs within the compared fragments. Although this technique is appropriate for clustering large textual collections, it operates poorly when clustering small amounts of texts, such as sentences (Abdalgader, 2018, p. 378). 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 was what size the corpus should be to support objective and reliable generalizations about Thomas Hardy's prose fiction. The corpus on which the analysis is based consists of all the known (published and unpublished) prose fiction texts of Hardy.

As a first step in data representation, the corpus was confined to what is referred to a bag of words. It was also decided that the corpus would be built of only the content words. All function words were thus removed. The hypothesis is that they do not usually carry semantic meaning and thus they cannot be considered distinctive features, and a corpus should include only and all the distinctive features (Glynn & Robinson, 2014). Content words can act as strong predictors of the topic(s) or content of a document. Moreover, the experimental results of document clustering indicate that content word representation gives good results in identifying the content of a document and its latent structure (Eaton, 2007; Gani, Siddiqa, Shamshirband, & Hanum, 2016; Hofmann, 2017). Equally important, most studies seem to agree that content word representation has been shown to give much better results than any other approach to clustering. This study considered content words to be indicators of semantic content. In other words, the analysis identified 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*.

3.2. Data collection procedures

For the texts to be amenable to computational analysis, they were mathematically represented using a vector space model (VSM) as this is conceptually simple, as well as being 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 this kind is that documents will be clustered based on size rather than lexical content and semantic similarity. The row vectors were thus normalized to compensate for variations in length using the mean document length method. This had the effect that the documents were equally represented in the matrix and their lexical frequency profiles could meaningfully be clustered.

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 taken as the most distinctive lexical features within the corpus, as seen in Figure 1.

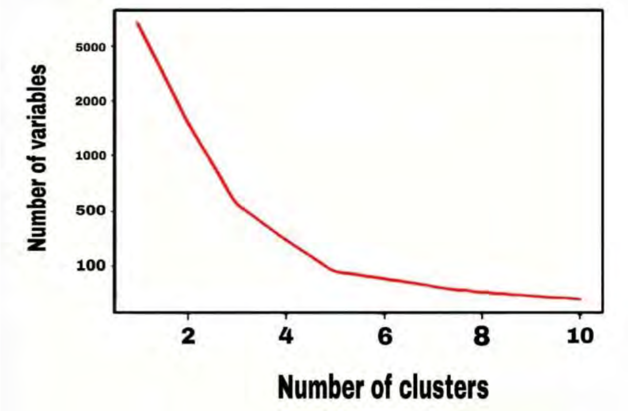

Figure 1. TFIDF analysis of the Matrix H62, 200

3.3. Data analysis

For the purposes of the study, cluster analysis was used. The rationale is that through cluster analysis, it is possible for vectors in an $\mathcal{N}$-dimensional space where variables are distributed in a non-random way to be assigned into distinct clusters or groups. Consider, for example, the plot of 100 three-dimensional randomly-generated vectors shown in Figure 2.

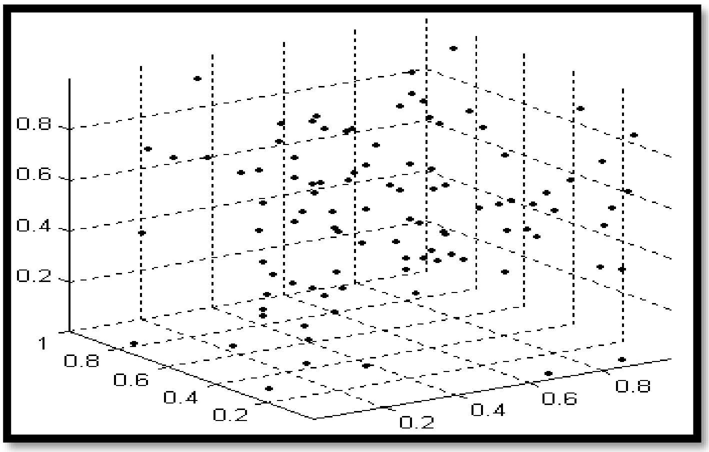

Figure 2. A plot of 100 three-dimensional randomly-generated vectors

As seen in Figure 2, the vectors are not uniformly distributed in the vector space; they are randomly distributed. The way the data was generated indicates that such structure is an accidental by-product of randomness. Visual inspection suggests some weak regularities, which makes it hard to identify any relationship between these vectors. If vectors, on the other hand, are organized into a non-random dimensional space, as shown in Figure 3, it is possible to define relationships among these vectors.

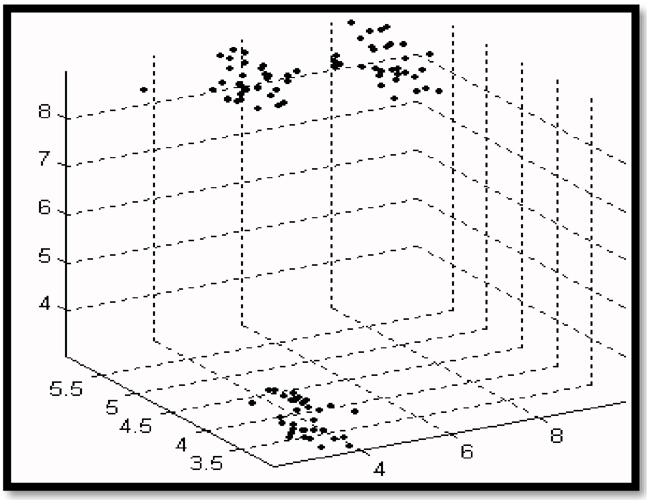

Figure 3. A plot of 100 three-dimensional nonrandomly-generated vectors

Unlike the random distribution of the data set shown in Figure 2, the distribution of points in this example (Figure 3) is non-random: the vectors are assigned into three clearly defined groups. The reason for the selection of cluster analysis methods can be discussed under the heading of appropriateness. Cluster analysis can effectively deal with data that is large in terms of the number of objects being studied for dimensionality which makes it useful in generating a clustering of the novels and short stories of Thomas Hardy. If we look at the prose works, it is difficult to discover any interrelationships that may exist among texts simply because the volume of writing is large and takes a lot of time to assimilate by reading. The overview of 62 texts, the selected data, must be foggy. Similarly, data derived from such a volume of texts will be difficult if not impossible to interpret. It is generally cognitively difficult for people to deal with unorganized sets of data, take the texts of Thomas Hardy as an example. Cluster analysis makes it cognitively easier for researchers to discover interrelationships of objects based on group membership. For this reason, it was considered that cluster analysis would usefully supplement the objectives of the thesis in generating an objective, replicable classification of Hardy's prose fiction.

Within cluster analysis, large sets of data are subdivided into smaller sets that can describe the original observations without sacrificing critical information. Cluster analysis aims to discover a system of organizing observations into distinct groups where members of each group share properties in common. This is foreshadowed by the objective of the thesis in constructing a sensible and informative categorization the 62 texts of Thomas Hardy where the members of each class share certain thematic properties. Cluster analysis results stimulate the making of generalizations about the data. This is again one of the objectives of the thesis. The study seeks to propose some hypotheses, based on the results of the analyses- that may serve in making generalizations about the data.

Cluster analysis was thus used to find meaningful clusters in the data, generating a centroid-based lexical clustering structure that could capture and describe 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 identify the recurrent themes. Hierarchical cluster analysis, one of the main statistical approaches used to find distinct classes or groups based on shared and common features, was used to visualize the clustering structure. Despite the development of different clustering algorithms, hierarchical cluster analysis, or hierarchical clustering, remains one of the most widely used algorithms in clustering applications to find discrete groups with varying degrees of (dis)similarity in a data set represented by a (dis)similarity matrix (Tullis & Albert, 2008). The selected texts fall into four main clusters, as shown in Figure 4.

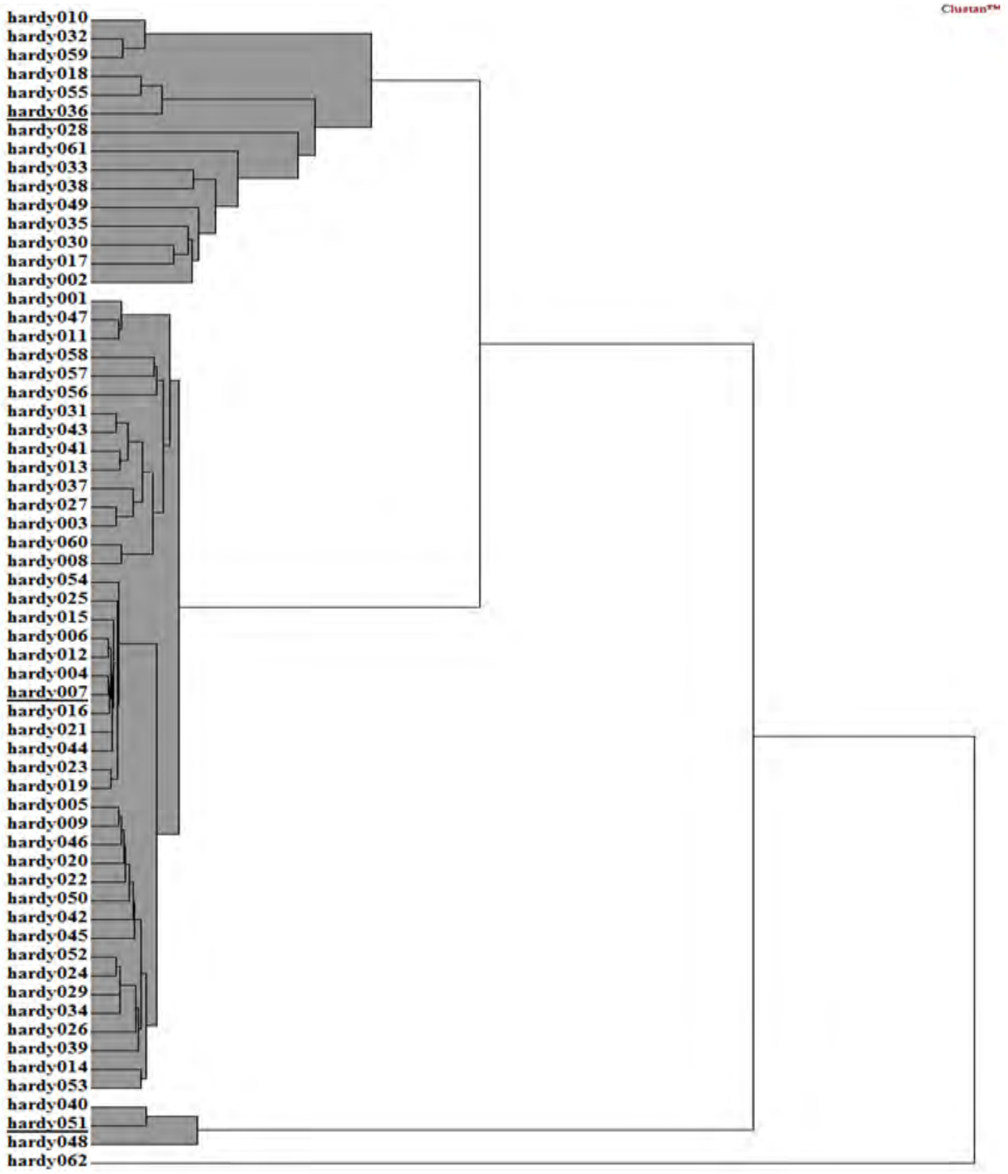

Figure 4. Clustering structure of Hardy's Matrix using Euclidean distance & Ward linkage clustering

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 (Abdalgader, 2018). In this case, the clustering structure was used for building hypotheses and making generalizations about the thematic relations of the texts in each group or cluster.

For validity purposes, principal component analysis (PCA) was used. The validity of the results of cluster analysis is an important requirement since different cluster structures may lead to completely different interpretations of the same data, thus generating contradictory hypotheses about the data. The purpose of clustering validation is thus to see whether the same analytical methods applied to an alternative representation of the data will give identical or at least similar results. The alternative data representation was generated by PCA, a dimensionality reduction method whereby H62, 200 was reduced to a dimensionality of 61, yielding the matrix H62, 50 (Bozdogan & Gupta, 2012; Kassambara, 2017). The analysis showed full agreement between the results of the clustering structures.

To identify the most distinctive lexical features of each group, the columns of the matrix were rearranged in the 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 constituting Groups A, B C, and D in accordance with the function

$$\mathbf{V}_{\boldsymbol{i}} = \frac{\sum_{\boldsymbol{i}=\mathbf{1}\ldots\boldsymbol{m}} \mathbf{H}\boldsymbol{ij}}{\boldsymbol{m}})$$

where

V_j is the jth 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$_{\text{centroid}}$, Group B$_{\text{centroid}}$, Group C$_{\text{centroid}}$, and Group D$_{\text{centroid}}$, were compared to show how the three groups differed on average on each of the extracted lexical variables, the aim being to identify the variables on which they differed most, and thereby infer the thematic characteristics of each group.

4. Results

This section considers 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 that distinguishes it from the others (Omar, 2010). Thus, 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 the thematic characteristics of the respective groups (Omar, 2020a, 2020b).

Based on 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 (2007) 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 he classified the majority of the texts included here under the category of Romance and Fantasy.

The hierarchical clustering structure, which is based on pure mathematical methods, supports Hardy’s tendency to group similar short stories together. Six texts in 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 collated in *Noble Dames* when it was published in serial form in *Harper’s Weekly* and the *Graphic* in late 1890 (Dalziel, 1992b). Purdy (1979) 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 Fantasy, 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 (1982), 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. (p. 52)

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 (1992a), 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. However, this is not in contradiction with the inclusion of texts such as *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 (2007) 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" (p. 24).

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, correlating with ideas of social promotion/hostility and struggle. This subclass includes texts referred to by many critics as Hardy's major works: *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 they mock the social mores of the age. The texts talk generally about heroes and heroines who aspire to 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 together with the results of qualitative analysis point to social indicators influencing his career as a novelist. The social dimension is never absent from 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 women's 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.

The 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 in which 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, suspects her of infidelity, is jealous, and decides to take revenge against the person who he thinks is her lover because of the disgrace such an illicit relationship causes him. Finally, Group D includes just one text, *The Unconquerable* (Hardy062). The most important variables in 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, possibly indicating why it has unique lexical features making it distinct from the other texts.

5. Discussion

Based on the results presented in section 4, it can be claimed that the clustering structures are meaningful. Each cluster or group has its own distinctive lexical profile distinguishing it from other groups or clusters. It may be argued 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. This study suggests that the two texts are related to each other in terms of dealing with class consciousness. In his introduction to the New Wessex Edition of *The Hand of Ethelberta*, for example, Gittings (1978) underestimates the novel, classifying it as "the joker in the pack" of Hardy's novels. Widdowson (1998), in contrast, 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 provides 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 (1998) 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, with the thematic classification pointing to tragic, historic, and fantasy elements in the texts. Consequently, as far as thematic interrelationships are concerned, it is clear that the texts exhibit clear features for genre classification. This can be a starting point for a comprehensive genre classification of the novels and short stories of Hardy, making it possible for texts to be classified under the main categories of tragedy, comedy, romance, epic, fantasy, history, pastoral history, etc. One advantage of such a classification is that it can hone the ways in which we think about the works. 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.

6. Conclusions

This paper has addressed the issue of whether thematic concepts can be identified in literary texts using computational models. It reports on a study in which document clustering methods were used to group the selected texts into distinct classes based on their semantic similarity. The results clearly indicate that such methods can usefully be employed to generate distinct and meaningful classes expressing certain thematic concepts, such as class status, sex, marriage, love, romance, and the English countryside. The results of the analysis are objective in that they are generated through the application of mathematically based computational methods to empirically derived data; as such, they are not open to any influence from the theoretical presuppositions that the researcher might have. Unlike the results of philological methodologies, computational results are replicable, testable, and thus scientifically respectable. This is not to disparage the subtle elaborations of literary criticism of Thomas Hardy that have been undertaken over many years. Indeed, the results of the study largely agree with non-computational philological classifications of Hardy's texts, which have generated a number of hypotheses that have not been empirically confirmed; rather, the contribution is that the results obtained from this study are objective.

7. Ethics Committee Approval

The authors confirm that ethical approval was obtained from Prince Sattam Bin Abdulaziz University (Approval Date: January 17, 2021).

Acknowledgements

This publication was supported by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia

References

Abdalgader, K. (2018). Centroid-based lexical clustering. In H. Pirim (Ed.), *Recent applications in data clustering* (pp. 378–403). London: IntechOpen.

Aggarwal, C. C., & Reddy, C. K. (2016). *Data clustering: algorithms and applications*. London; New York: Chapman and Hall/CRC Press.

Balossi, G. (2014). *A Corpus linguistic approach to literary language and characterization: Virginia Woolf's The Wave*. Amsterdam, Netherlands: John Benjamins Publishing Company.Amsterdam, Netherlands

Bevis, M. (2013). *The Oxford handbook of Victorian poetry*. Oxford: Oxford University Press.

Bownas, J. L. (2012). *Thomas Hardy and Empire: The Representation of Imperial Themes in the Work of Thomas Hardy*. Farnham: Ashgate

Bozdogan, H., & Gupta, A. K. (2012). Multivariate statistical modeling and data analysis. Proceedings of the *Advanced Symposium on Multivariate Modeling and Data Analysis May 15–16, 1986*. Netherlands: Springer.

Brady, K. (1982). *The short stories of Thomas Hardy*. New York, NY: St. Martin's Press.

Brantlinger, P., & Thesing, W. (2008). *A companion to the Victorian novel*. Oxford: Wiley.

Burrows, J. (2004). Textual analysis. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 88–97). Oxford: Blackwell.

Chakraborty, G., Pagolu, M., & Garla, S. (2014). *Text mining and analysis: Practical methods, examples, and case studies using SAS*. Cary, North Carolina: SAS Institute.

Cox, R. G. (1970). *Thomas Hardy: The critical heritage*. New York, NY: Barnes & Noble.

Dalziel, P. (1992a). Hardy's unforgotton "indiscretion": The centrality of an uncontrolled work. *Review of English Studies, XLIII* (171), 347–366. doi:10.1093/res/XLIII.171.347

Dalziel, P. (Ed.) (1992b). *Thomas Hardy: The excluded and collaborative stories*. Oxford: Clarendon Press.

Dillion, J. (2016). *Thomas Hardy: Folklore and resistance*. London: Palgrave Macmillan UK.

Eaton, M. L. (2007). *Multivariate statistics: A vector space approach* (Vol. 53). Beachwood, OH: Institute of Mathematical Statistics.

França, F. M. G., & de Souza, A. F. (2008). *Intelligent text categorization and clustering*. Berlin, Heidelberg: Springer.

Gani, A., Siddiqa, A., Shamshirband, S., & Hanum, F. (2016). A survey on indexing techniques for big data: Taxonomy and performance evaluation. *Knowledge and Information Systems, 46*(2), 241–284.

Gilmartin, S., & Mengham, R. (2007). *Thomas Hardy's shorter fiction: A critical study*. Edinburgh: Edinburgh University Press.

Gittings, R. (Ed.) (1978). *An introduction to The Hand of Ethelberta* (New Wessex edition ed.). New York, NY: St. Martin's Press.

Glynn, D., & Robinson, J. A. (2014). *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*. Amsterdam ; Philadelphia: John Benjamins Publishing Company.

Gold, M. K., & Klein, L. F. (2016). *Debates in the digital humanities*. Minneapolis: University of Minnesota Press.

Headrick, P. P. (2013). *The Wiley guide to writing essays about literature*. New York: John Wiley & Sons.

Hodson, J. (2017). *Dialect and literature in the long nineteenth century.* New York: Routledge.

Hofmann, T. (2017). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 50-57.

Hoover, D. L., Culpeper, J., & O'Halloran, K. (2014). *Digital literary studies: Corpus approaches to poetry, prose, and drama.* New York: Routledge.

Ireland, K. (2014). *Thomas Hardy, time and narrative: A narratological approach to his novels.* London: Palgrave Macmillan UK.

Jockers, M. L., & Thalken, R. (2020). *Text analysis with R: For students of literature.* Cham, Switzerland: Springer International Publishing.

Kachuck, B. (1995). Feminist social theories: Theme and variations. *Sociological Bulletin, 44*(2), 169–193.

Kassambara, A. (2017). *Practical guide to principal component methods in R.* Statistical Tools for High-Throughput Data Analysis (STHADA).

King, J. (1978). *Tragedy in the Victorian novel: Theory and practice in the novels of George Eliot, Thomas Hardy and Henry James.* Cambridge: Cambridge University Press.

Kogan, J. (2007). *Introduction to clustering large and high-dimensional data*. Cambridge: Cambridge University Press.

Mallett, P., & Maier, S. E. (2013). *Thomas Hardy in context*. Cambridge: Cambridge University Press.

Mani, I. (2013). *Computational modeling of narrative.* San Rafael, California: Morgan & Claypool Publishers.

Moisl, H. (2015). *Cluster analysis for corpus linguistics.* New York: Walter De Gruyter.

Mulhern, F. (2014). *Contemporary Marxist literary criticism.* London: Routledge.

Mullings, C., Kenna, S., Deegan, M., & Ross, S. (2019). *New technologies for the humanities.* London: De Gruyter Saur Verlag.

Nemesvari, R. (2011). *Thomas Hardy, sensationalism, and the melodramatic mode*. New York, NY: Palgrave Macmillan US.

Omar, A. A. (2010). 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, 1*(2).

Omar, A. A. (2020a). Feature selection in text clustering applications of literary texts: A hybrid of term weighting methods. *International Journal of Advanced Computer Science and Applications, 11*(2), 99–107.

Omar, A. A. (2020b). On the digital applications in the thematic literature studies of Emily Dickinson's poetry. *International Journal of Advanced Computer Science and Applications, 11*(6), 361–365.

Page, N. (2000). *Oxford reader's companion to Hardy.* Oxford: Oxford University Press.

Pugh, T., & Johnson, M. E. (2013). *Literary studies: A practical guide.* London; New York: Routledge.

Purdy, R. L. (1979). *Thomas Hardy: A bibliographical study.* Oxford*:* Oxford University Press.

Riesen, K., & Bunke, H. (2010). *Graph classification and clustering based on vector space embedding.* New Jersey, United States: World Scientific Publishing Company.

Shanahan, J. G., Qu, Y., & Wiebe, J. (2005). *Computing attitude and affect in text: Theory and applications.* Heidelberg: Springer Netherlands.

Siemens, R., & Schreibman, S. (2013). *A companion to digital literary studies.* Oxford: Blackwell.

Somani, A. K., Shekhawat, R. S., Mundra, A., Srivastava, S., & Verma, V. K. (2019). *Smart systems and IoT: Innovations in computing. Proceeding of SSIC 2019.* Springer Singapore.

Srivastava, A. N., & Sahami, M. (Eds.). (2009). *Text mining classification, clustering, and applications* (1st ed.). Boca Raton, Florida: Chapman and Hall/CRC.

Tullis, T., & Albert, B. (2008). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (2nd ed.). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Vigar, P. (2014). *The novels of Thomas Hardy: Illusion and reality.* London: Bloomsbury Academic.

Wellek, R., & Warren, A. (1963). *Theory of literature.* Harmondsworth, London: Penguin.

Widdowson, P. (1998). *On Thomas Hardy: Late essays and earlier*. Basingstoke: Macmillan.

Wilson, K. (2010). *A companion to Thomas Hardy.* New York: Wiley-Blackwell.

Wu, W., Xiong, H., & Shekhar, S. (2013). *Clustering and information retrieval.* Heidelberg: Springer.

Zyngier, S. (2008). *Directions in empirical literary studies: In honor of Willie Van Peer.* Amsterdam, Netherlands: John Benjamins Publishing Company.

Appendix A. Texts and naming 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

Kurgudaki temaları belirleme: Centroid tabanlı sözcüksel kümeleme yaklaşımı

Özet

Son yıllarda, beşeri bilimler ve edebi çalışmalarda yaygın olarak kullanılan sayısız hesaplama yönteminin gelişimine tanık oldu. Seçicilik, nesnellik ve tekrarlanabilirlik dahil olmak üzere bu alanlardaki farklı içsel sorunlara uygulanabilir çözümler sağlama açısından bu tür yöntemlerin potansiyellerine rağmen, literatürdeki tematik çalışmalar üzerinde çok az şey yapılmıştır. Neredeyse tüm çalışmalar, bireysel araştırmacıların metinleri okumasına ve bu okumadan genellemelerin sezgisel soyutlamasına dayanan geleneksel yöntemlerle yapılır. Bu yaklaşımların nesnellik ve tekrarlanabilirlik konularında olumsuz etkileri vardır. Dahası, bu tür geleneksel yöntemlerin her yıl yayınlanan yüzbinlerce yeni romanı etkili bir şekilde ele alması zor. Bu sorunlar karşısında, bu çalışma sözcüksel kümeleme yöntemlerine dayalı olarak edebi metinlerin tematik sınıflandırmaları için entegre bir hesaplama modeli önermektedir. Örnek olarak, bu çalışma Thomas Hardy'nin romanlarını ve kısa öykülerini içeren bir külliyat üzerinden yapılmıştır. Metinlerin sözcüksel içeriğinin vektör uzayı modeline (VSM) dayalı olarak hesaplamalı anlambilimsel analizi kullanılmıştır. Sonuçlar, seçilen metinlerin anlamsal içeriklerine göre tematik olarak gruplandırıldığını göstermektedir. Hesaplama teorisinde ve veri madenciliği uygulamalarında uzun süredir kullanılan metin kümeleme yaklaşımlarının edebi çalışmalarda yararlı bir şekilde kullanılabileceği nihayet iddia edilebilir.

Anahtar *sözcükler*: hesaplamalı modeller; hesaplamalı anlambilim, sözcüksel kümeleme; sözcük içeriği; filolojik yöntemler; Thomas Hardy; Vektör Uzayı Modeli

AUTHOR BIODATA

Dr. Abdulfattah Omar is an Associate Professor of English language and linguistics at Prince Sattam Bin Abdulaziz University (KSA). He received his PhD degree in 2010 from Newcastle University, UK. Dr. Omar's research interests include computational linguistics, digital humanities, and translation studies.