Towards a Computational Model to Thematic Typology of Literary Texts: A Concept Mining Approach

Abdulfattah Omar
Department of English, College of Science & Humanities
Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
Department of English, Faculty of Arts, Port Said University, Egypt

***Abstract*—In recent years, computational linguistic methods have been widely used in different literary studies where they have been proved useful in breaking into the mainstream of literary critical scholarship as well as in addressing different inherent challenges that were long associated with literary studies. Such computational approaches have revolutionized literary studies through their potentials in dealing with large datasets. They have bridged the gap between literary studies and computational and digital applications through the integration of these applications including most notably data mining in reconsidering the way literary texts are analyzed and processed. As thus, this study seeks to use the potentials of computational linguistic methods in proposing a computational model that can be usefully used in the thematic typologies of literary texts. The study adopts concept mining methods using semantic annotators for generating a thematic typology of the literary texts and exploring their thematic interrelationships through the arrangement of texts by topic. The study takes the prose fiction texts of Thomas Hardy as an example. Findings indicated that concept mining was usefully used in extracting the distinctive concepts and revealing the thematic patterns within the selected texts. These thematic patterns would be best described in these categories: class conflict, Wessex, religion, female suffering, and social realities. It can be finally concluded that computational approaches as well as scientific and empirical methodologies are useful adjuncts to literary criticism. Nevertheless, conventional literary criticism and human reasoning are also crucial and irreplaceable by computer-assisted systems.**

Keywords*—Computational linguistics; concept mining; data mining; empirical methodologies; semantic annotators; text clustering; typology

I. Introduction

Despite the wide applications of computational and statistical approaches in different disciplines in humanities, many critics still argue against the usefulness of these approaches in literary studies. So far, most literary critics reject the use of computer technology and statistical and computational methodologies in the analysis and interpretation of literary texts [1]. In light of this argument, this study seeks to evaluate the reliability of computational and statistical approaches to literary studies, and more specifically to the thematic typologies of literary texts. In other words, it seeks to see whether computational and statistical approaches, which have long been rejected by many literary critics, can be usefully used in generating typologies that best reveal the thematic features within these texts.

The study takes Thomas Hardy's prose fiction production as an example. The great reputation Hardy received as a novelist and the thematic richness of his texts have always made his novels and short narratives a target for critics and commentators [2]. From the time Hardy published Wessex Edition, where he provided a broad classification of his works, a critical response has focused on the questions of thematic classification of his work. Critics have adopted different approaches for investigating Hardy's thematic treatment of his texts [3-5]. Nevertheless, questions are often raised regarding the reliability of these classifications. In this regard, Hardy's prose fiction represents a good opportunity to test the reliability of computational and statistical methods in the thematic typology applications to literary texts.

To carry out the objectives of the study, concept mining is used. The rationale is that concept mining methods have been usefully used in text clustering and text classification applications. They have been effectively used for generating reliable clustering and classifications that explored the underlying meanings and themes within texts. Unlike conventional text clustering approaches including vector space clustering (VSC), concept mining focuses on identifying the patterns that are associated with concepts in similar texts [6, 7]. The underlying premise is that concept mining can be usefully used for identifying and recognizing the thematic patterns which can thus be used in developing a thematic typology that expresses the thematic interrelationships within literary texts.

The remainder of this article is organized as follows. Section Two surveys the literature on the thematic typology of literary texts. Section Three proposes the research questions. Section Four describes the methodological framework of the study. Section Five describes the data collection and processing procedures. Section Six reports the results of the study. Section Seven is a discussion and interpretation of the results. Section Eight is conclusion.

II. Literature Review

Different critical approaches have been used in the thematic typology studies of literary texts. These have usually been traditionally based on qualitative methods that tended to focus on identifying and revealing the underlying meanings that are

conveyed within the texts [8]. In this regard, thematic classifications have been carried out according to different concepts including purpose, content, and period [9]. Despite the popularity of the concepts of content and purpose, typologies based on these two concepts are always associated with subjectivity being largely based on the intuitive readings and making general observations based on such readings. One major problem with the adoption of these concepts in the identification of themes is that they are based on personal thoughts and emotional reactions, and thus they are subjective.

The period concept, on the other hand, is one of the most dominating methods in the thematic classification of literary texts. The underlying premise of this concept is that every period has its characteristic thematic features that can be usefully used in grouping texts through time. Taking English literature as an example, it is usually classified under the headings the Anglo-Saxon Literature, the Medieval Literature, the Elizabethan Literature, the Neoclassical literature, the Romantic literature, the Victorian literature, the Modernist literature, and the Postmodern literature.

Following this tradition, different thematic typologies have been developed in the critical study of Thomas Hardy's prose fiction. According to Plietzsch [10], the first person to suggest a general classification of Hardy's novels is thought to be Edmund Gosse in 1890. Gosse's classification included only ten novels, which were all that Hardy had written so far [11]. In this regard, Gosse's study misses many of the themes Hardy was to develop in subsequent work. With its limitation, Gosse's classification was nevertheless a step forward in generating a broad classification of Hardy's works.

Following Gosse's attempt, Thomas Hardy provided a broad classification of his prose and verse works. In the Wessex Edition, Hardy [12] classified his novels and short stories into three categories: (1) Novels of Character and Environment, (2) Romances and Fantasies, and (3) Novels of Ingenuity. One obvious observation about Hardy's classification is that there is no clear-cut relationship between the thematic accounts Hardy himself suggested for this classification and the way the texts were finally classified. Plietzsch [13] argues that it seems that Hardy ranked his texts in this particular order as a result of the responses from the public and literary critics which he had received.

The implication for the present study is that the reliability of such a classification is thus questionable. First, Hardy did not provide definite criteria for his classification. Second, some of the texts that used to be regarded as minor by the public and critics in Hardy time are now regarded as major works [14, 15]. Furthermore, the classification does not include all of Hardy' works. A Changed Man and other Stories, for instance, was published one year after the Wessex Edition. In view of this, it excludes some important works that represent the thematic development of Hardy's career as a novelist.

In response to Hardy's work, Abercrombie [16] classified the novels and short stories based on their artistic significance into four categories: (1) Minor Novels, (2) Annexes, (3) Dramatic Form, and (4) Epic Form. In spite of its success in drawing connections between texts, Abercrombie's classification, however, raises many questions concerning replicability and objectivity, since his criteria are subjective and undefined.

Harvey [14] suggested an alternative typology where he divided Hardy's prose works into three main categories: major novels, lesser novels, and short stories. The main criterion of this classification is subject matter. Harvey considered only the social and realistic novels to be major novels. Other novels were classified under the category of minor novels. Once again, the classification lacked any objective criteria [15, 17].

Another approach to the typology of Hardy's prose fiction texts can be traced in grouping Hardy's novels and short stories based on literary criticism perspectives. The underlying principle is that critics have come to classify Hardy's works under different headings including tragedy, women, religion and philosophy, Wessex and regionalism, nature and landscape, social change, and pastoral. One major problem with these critical discussions is that their perception of Hardy's work is very narrow in the sense that they are almost restricted to what Hardy calls 'Novels of Character and Environment'. Furthermore, they ignore important thematic concepts within the texts as they usually focus on just one aspect of his writings. Equally important, such reviews are always based on some biographical elements or historical accounts which again raise questions regarding the objectivity and reliability of such typologies and classifications.

In the face of the problems associated with the conventional classifications of Hardy's prose works, recent studies built on the advances in the application of computational data processing for analyzing and classifying novels and short stories in a way that is both objective and replicable [18, 19]. Motivation has been to understand Thomas Hardy better as a literary artist in an objective, replicable, therefore scientific way. In a recent study, Omar [20] employed centroid-based lexical clustering methods for identifying the thematic structures in Hardy's prose fiction. The novels and short stories were clustered into four classes, where each class or group of texts share the same thematic features based on the lexical profiles of each class.

Despite the reasonable success of such classification in drawing a thematic mapping of Hardy's novels and short stories based on objective grounds, questions regarding the use of vector space clustering (VSC) in exploring the salient thematic features of the texts are often raised. VSC is based on what is known as bag of words techniques where context is not considered at all. In this regard, VSC is not capable of accounting for all the linguistic and contextual features of texts. In the face of these limitations, this study proposes the use of concept mining methods for generating a thematic typology that best captures the underlying meanings, concepts, and thematic patterns of literary texts taking the novels and short stories of Thomas Hardy as an example.

III. Research Questions

Despite the extensive literature on the thematic typology of literary texts including Thomas Hardy's prose fiction writings, almost all of the relevant work is theoretically driven. That is, classification criteria are selected by the critic based on some critical theory or framework supported by personal knowledge

and evaluation of the texts. Moreover, many existing accounts follow the stereotypical classifications of what might be called Hardy Critical Industry. In other words, many of Hardy's commentators are willing to agree with conventional, well-known evaluations of Hardy even though such evaluations conflict with their critical presuppositions. Two examples of this are given. First, many commentators have favored the idea of classifying Hardy's works into major and minor novels in relation to subject-matter, and many studies use such dichotomy without giving reasons for its adoption. Second, many thematic reviews of Hardy use the term 'Wessex novels' in reference to nine or ten of Hardy's novels without explaining why these nine or ten texts should constitute variants of the same theme apart from the fact that they are about Wessex [21].

It can also be claimed that many of the classifications of Hardy's work followed Hardy's own classification of his works. The problem is that Hardy did not set clearly defined criteria for his classification. Furthermore, some classifications based on philological methods are greatly biased. In the face of this problem, this study seeks to answer the following research questions:

- Can computational linguistic methods be usefully used in addressing the limitations of the conventional approaches of literary criticism regarding thematic typologies of literary texts?
- How can computational models in general and concept mining methods in particular be used in developing a thematic typology of literary texts with reference to the prose fiction writings of Thomas Hardy?
- What is the future of computational approaches and scientific and empirical methodologies in literary studies?

IV. Method

In different natural language processing (NLP) applications including text clustering and text classification, concept mining is a process that has been used to provide an automated categorization of documents based on their content [22, 23]. It is a workflow that is used to discover implicit and explicit relationships, useful associations and groupings in a set of documents or data collection with the purpose of detecting similar documents in a large corpora and classifying them by topic [24, 25]. It can provide thus powerful insights into the meaning, provenance, and similarity of documents [26-28]. The assumption is that each word in a given document relates to several possible concepts which make it possible to cluster documents based on their content. The underlying principle of concept mining is the conversion of words into concepts. This is done in two subsequent steps. First, documents are reduced into a sequence of words that describes the content. Second, these words are mapped into concepts [29].

In this way, given that we have a number of documents on generative grammar; concept mining is possible by identifying relationships and generating facts based on the data within collection and the dimensions of the subject. These can be something like Chomsky and generative grammar, theoretical linguistics and generative grammar, Phrase Structure Rules (PSR) and Generative grammar, deep and surface structures in generative grammar, etc. Documents can also be classified by topic as WH-movement, linguistic competence, etc.

In this way, concept mining is based on clustering or grouping semantically-similar texts together. Text clustering is the process of automatically grouping natural language texts according to an analysis of their information/semantic content. In other words, clustering is a task of dividing given data into defined set of clusters and it is the task of classification to structure these clusters and sort them into categories according to a group structure known in advance [30, 31]. In concept mining processes, text clustering starts by discovering and finding groups that have similar content, and then organizing our perceptions of these groups into categories. In other words, clustering places documents into natural classes and generating taxonomies that best describe the patterns within the datasets [32].

V. Data Collection Procedures

For generalizability purposes, the study is based on all the novels and collections of short stories written by Thomas Hardy. Three sources were used for data collection. These are shown as follows.

- Chadwyck-Healey Literature Collections is a commercial product with authoritative full-text databases that offers coverage of English literary works from 1477 to the present.
- The Gutenberg Project is the oldest producer of free e-books on the Internet with a large volume of collections produced by thousands of volunteers. The project was founded in 1971 by Michael Hart.
- The Thomas Hardy Short Story Page (http://darlynthomas.com/hardyshortstories.htm-) includes all collection of short stories and the individual, excluded and collaborative stories written by Thomas Hardy.

Hardy has 14 published novels. These are listed below.

1) Desperate Remedies
2) Under the Greenwood Tree
3) A Pair of Blue Eyes
4) Far from the Madding Crowd
5) The Hand of Ethelberta
6) The Return of the Native
7) The Trumpet-Major
8) A Laodicean
9) Two on a Tower
10) The Mayor of Casterbridge
11) The Woodlanders
12) Tess of the D'Urbervilles
13) Jude the Obscure
14) The Well-Beloved

In his life time, Hardy also published four collections of short stories. These are A Group of Noble Dames [33], Life's Little Ironies [34], Wessex Tales [35] and A Changed Man and

other stories [36]. These account for "thirty-seven stories in all [37]. For the first three collections, texts were abstracted from the Wessex Edition [12]. A Changed Man and other stories was published one year after the publication of the Wessex Edition. So it was not included in that edition. The data was abstracted from Macmillan & Co 1913 edition which includes as well the novella The Romantic Adventures of the Milkmaid.

A. Wessex Tales

The short stories used in this study are the contents of the Wessex Tales collection of the 1912 *Wessex Edition*. These are shown as follows.

The Three Strangers
A Tradition of Eighteen Hundred and Four
The Distracted Preacher
The Withered Arm
Fellow-Townsmen
Interlopers at the Knap

B. Life's Little Ironies

The stories in this collection had been written at different periods but were assembled in 1893 for publication under the title given. In 1912, Hardy reassembled the stories for the Wessex Edition as indicated below. The texts used for the data of the study are the ones in the Wessex Edition. The stories in this collection are listed below.

An Imaginative Woman
For Conscience's Sake
The Fiddler of the Reels
To Please His Wife
On the Western Circuit
A Few Crusted Characters
A Tragedy of Two Ambitions
The Son's Veto

C. A Group of Noble Dames

In his preface to *A Group of Noble Dames*, Hardy [38] indicates that the tales were first published in periodicals six or seven years before being collected and published in 1891. Hardy published the tales once again in the same form in which they appeared in the 1891 edition for Wessex Edition in 1912. The stories in this collection are shown below.

The First Countess of Wessex
Barbara of the House of Grebe
The Marchioness of Stonehenge
The Lady Icenway
The Duchess of Hamptonshire
Anna, Lady Baxby
Lady Mottisfont
Squire Petrick's Lady
The Honourable Laura
The Lady Penelope

D. A Changed Man and other Stories

Although the tales in this collection represent an important stage in Hardy's development, they were not collected in one volume until 1913. Interestingly, Hardy used the expression minor novels instead of short stories. This may suggest that the short story was not a fully-fledged literary genre yet. Stories in this collection are shown below.

A Changed Man
Alicia's Diary
A Tryst at an Ancient Earthwork
A Committee-Man of 'The Terror
The Waiting Supper
The Grave by the Handpost
What the Shepherd Saw
Master John Horseleigh, Knight
A Mere Interlude
The Duke's Reappearance
Enter a Dragoon
The Romantic Adventures of a Milkmaid

E. Excluded and Collaborative Stories

These are the stories which were not included in the collected volumes published during Hardy's life. They were collected and edited by Pamela Dalziel [39] in Thomas Hardy: The Excluded and Collaborative Stories. Dalziel argues that although the stories occupy a significant position in the professional career of Hardy as a novelist, they have not received due critical treatment from critics and biographers. She also stresses that these stories are not thematically coherent: "Each story is an individual work, meriting treatment as such, and its unique conditions of composition and publication (or non-publication) have been carefully considered, in so far as they are now recoverable, when making editorial decisions" [39]. In making these stories available in one volume, Dalziel addresses some limitations in Hardy scholarship.

F. A List of Hardy's Excluded and Collaborative Stories

This collection consists of 10 stories including both excluded and collaborative stories of Hardy. The Spectre of the Real is the only story acknowledged by Hardy to be a collaborative tale [39]. It was written in collaboration with Florence Henniker. Blue Jimmy: the Horse Stealer and The Unconquerable were written in collaboration with his wife Florence Dugdale-Hardy. However, Hardy never admitted Florence's role in the two stories [39]. The texts in this collection are listed below.

How I Built Myself a House
The Thieves Who Couldn't Help Sneezing
The Doctor's Legend
Our Exploits at West Poley
Destiny and a Blue Cloak
Old Mrs. Chundle
The Spectre of the Real
The Unconquerable
An Indiscretion in the Life of an Heiress
Blue Jimmy: The Horse Stealer

G. Unpublished Work

The Poor Man and the Lady is Hardy's first novel. He sent the manuscript of the novel to Macmillan who refused to publish it on the grounds that the book creates a world which is entirely dark. Accordingly, Hardy took it to Chapman & Hall where George Meredith recommended him not to publish it. Meredith thought that the text was socialist, injudiciously

provocative, and full of indiscriminate satire. As a result, Hardy gave up the idea of publishing it and it is commonly said that he burned it [40]. Nevertheless, many critics often have argued that the novel may have been partly drawn on for the short story An Indiscretion in the Life of an Heiress [41, 42]. However, Hardy always stressed that the short story is different from the novel [43]. The text of this work is based on Weber's edition of Hardy' lost novel. Weber [44] claimed that there was an earlier record for the novel Hardy destroyed. He realized that the novel deserves a critical attention; therefore, he revived it. He gathered information from six sources together which equipped him, as he claims, with a detailed knowledge of The Poor Man and the Lady and used it for the synopsis of the novel he produced.

A corpus was thus built from the electronic texts of the novels and short stories. Codes were used in reference to the texts as shown in Table I.

TABLE I. THE CORPUS

Code	Title
Hardy01	Desperate Remedies
Hardy02	Under the Greenwood Tree
Hardy03	A Pair of Blue Eyes
Hardy04	Far from the Madding Crowd
Hardy05	The Hand of Ethelberta
Hardy06	The Return of the Native
Hardy07	The Trumpet-Major
Hardy08	A Laodicean
Hardy09	Two on a Tower
Hardy10	The Mayor of Casterbridge
Hardy11	The Woodlanders
Hardy12	Tess of the D'Urbervilles
Hardy13	Jude the Obscure
Hardy14	The Well-Beloved
Hardy15	The Poor Man and the Lady
Hardy16	Wessex Tales
Hardy17	Life's Little Ironies
Hardy18	A Group of Noble Dames
Hardy19	A Changed Man and other Stories
Hardy20	Excluded and Collaborative Stories

It was decided to consider each of the collection of short stories as a single document. Short stories were not considered as separate documents. The rationale is that concept mining is more effective with long documents. In this regard, it was thought that concept mining would work better with the collections of short stories than individual stories.

VI. DATA ANALYSIS

For the extraction of the concepts, concept-based text representation was used. The documents were converted into clauses or what is referred to as 'bag of concepts'. Documents were represented as strings of concepts, where each document was represented by a given number of vectors. For this purpose, the study adopts the extraction model developed by Kim, et al. [45] shown in Fig. 1.

Given the high dimensionality of the corpus, it becomes impossible for any concept extraction or mining system to deal with these huge datasets effectively. In concept extraction applications, just like other clustering applications, high dimensionality is a serious problem that has adverse impacts on the reliability of the clustering performance [46]. With high dimensionality data in the clustering applications to literary texts, semantic similarity or relatedness is not accurately computed [47].

In the face of this problem, dimensionality reduction is carried out. The purpose is to keep only the distinctive features or variables. Also, concept-frequency inverse document frequency (CF-IDF) is used. CF-IDF is a weighting scheme for discovering the key concepts within datasets that is based on term-frequency inverse document frequency (TF-IDF) that tends to rank the concepts based on their frequency in relation to document frequency [48-50].

As a final step, semantic similarity between the texts is computed. Thus is a process whereby metrics are used for weighting or ranking similar concepts based on a concept taxonomy [51]. Semantically similar concepts can be thus grouped or classified together as shown in Fig. 2.

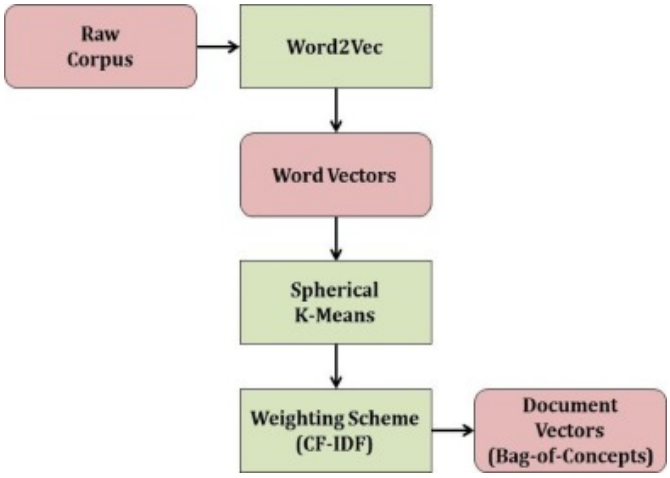

Fig. 1. Bag of Concepts Model Developed by Kim, et al. [45].

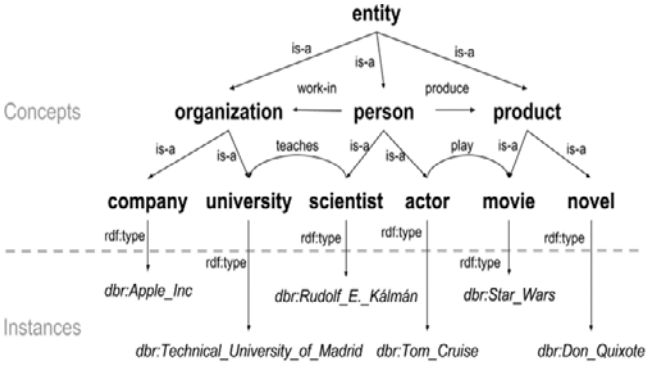

Fig. 2. An Example of Computing Semantic Similarity [51].

Most of metrics are designed for computing the semantic similarity with focus on the structure of the semantic network between concepts (e.g., path length and depth), or only on the Information Content (IC) of concepts [51]. These metrics, however, are not appropriate for the thematic typologies of literary texts which require identifying the conceptual similarities within texts. In so doing, the method developed by Zhu and Iglesias [51] for computing the conceptual similarity within texts is used for clustering and categorizing the selected texts based on their type concepts.

The selected texts are categorized based on their conceptual similarity into five categories. These are class conflict, Wessex, religion, female suffering, and social realities. First, the concept of class conflict is represented through romantic relations, mismatched marriages, elopement, and inequalities. Texts in this group included The Poor Man and the Lady, Life's Little Ironies, Far From the Madding Crowd, and A Group of Noble Dames. Second, the concept of Wessex is associated with a number of other concepts including rural life, local traditions, and industrialization. The concept of Wessex is extensively represented in Desperate Remedies, Under the Greenwood Tree, A Pair of Blue Eyes, The Hand of Ethelberta, The Trumpet Major, A Laodicean, and The Well-Beloved. Third, the concept of religion is represented through Christian beliefs and rituals, faith, morality, human existence, spiritual doubt, Evangelicalism and biblical references. Texts grouped under the concept of religion included Jude the Obscure, Tess of the D'Urbervilles, and Wessex Tales. Fourth, the concept of female suffering is represented through divorce, sexuality, struggle of women, oppression, and loss of chastity. Texts in this group included Two on a Tower, An Indiscretion in the Life of an Heiress, and A Changed Man and Other Tales. Finally, the concept of social realities is represented through poverty, social security, inequalities, and forces of oppression. Texts categorized under the heading of social realities included The Woodlanders. The Return of the Native and The Mayor of Casterbridge.

VII. Analysis and Discussions

Based on computing the conceptual similarity within the datasets, the texts were grouped into five main categories. These included class conflict, Wessex, religion, female suffering, and social realities. These are shown as follows.

A. Class Conflict

Class conflict is one of the central concepts in Hardy's novels and short stories including The Poor Man and the Lady and A Group of Noble Dames. In The Poor Man and the Lady, for instance, Hardy describes the ugly face of class conflict. This is represented in the love story between Miss Allamont and Will Strong. Miss Allamont is the squire's daughter and his heiress and Will Strong is a son of a peasant working on the estate of the Squire. In spite of the class gap between both, Miss Allamont takes a romantic interest in Will, and this is strongly rejected by her parents. Being rejected, Will moves to London where he achieves a striking success and becomes a public figure. However, he is still rejected by the family. This leads the two lovers to marry in secret and live away from her family. Soon her life is endangered and she dies. It was thus obvious that Hardy did not like class differences of his age and tended to represent the hypocrisy of the Victorian age in his books. Hardy describes the sufferings of the lower classes and the severe laws that threaten their lives in these books [52].

B. Wessex

Wessex is a dominating theme in many of Hardy's novels and short stories. In these texts, Hardy stressed the death of England's rural life along with its old customs and local traditions. Many critics consider Hardy as the greatest novelist in the form of regional fiction, and they think that the best example of regional fiction is Hardy's Wessex novels [53-55]. In the Wessex texts, Bullen [56] argues, Hardy succeeded at linking human behavior with the physical world. He adds that the works of Thomas Hardy to the historic place he was concerned with in his writings indicating that he was nostalgic for the past of England and that he distrusted modern civilization.

C. Religion

Religion is one of the central themes in Hardy's prose fiction. In his novels and short stories, the Bible and biblical names are obviously frequent. Influenced by controversies of the age, Hardy used what came to be known as the evolutionary narrative envisaged an alternative to a narrative which assumes that God created the world in its present state. Hardy expressed morality in a unique way. Morality is not based on traditional Christian beliefs. Rather, it is a social construct enforced by human intelligence rather than divine authority [57].

Many critics claim that any thematic discussion of Hardy's works has to consider religion as a crucial element in understanding and interpretation [58]. Hardy's texts cannot be fully understood without critical considerations of religious background and influence. According to Fergusson [59], the avoidance of such religious dimensions in thematic discussions results in interpretative gaps and loss of many thematic concepts in the literary texts.

D. Female Suffering/Tragedy

The texts included in this category reflect Hardy's preoccupation with the Victorian women and their sufferings. Hardy's women are destined to suffer. They are victims to the merciless conditions of the age. Women's suffering was deeply rooted in the hypocrisies of the Victorian society which was male dominated and obsessed with the idea of woman virginity. Many critics have advocated the idea that the texts of Hardy address the low position of women in the Victorian society and the strict laws that tended to deprive them of their independence. Morgan [60] argues that Hardy's texts reflect his sympathy towards women and his deep concern with their sufferings. Hardy introduces the sufferings of his women with a peculiar pathos and shows them as victims of male-dominated society. Likewise, in her book The Feminist Sensibility in the Novels of Thomas Hardy, Kaur [61] stresses that Hardy had sympathy for the women's cause and their sufferings. She stresses that Hardy is 'feministic', an artist with feminist sensibilities.

The results of the study agree with different feminist readings of Thomas Hardy that consider the writings of Hardy as a cry against the injustices done to the Victorian woman,

and an assertion her rights. The feminist investigations of Hardy's work often involve a discussion of sexuality and the sensual dimensions of the texts [60, 62, 63]. The main assumption of feminist readings of Hardy is that his texts reflect in one way or another the individual and social pressures the Victorian woman had to experience [61, 64-67]. The central concept in such reviews is that Hardy's works both depict and resist the male-centered culture and the oppression of the Victorian woman. At this point, much of the feminist reading of Hardy's prose fiction praises his progressive exploration, understanding, and support of women issues at a time of social crisis and change [66, 68].

E. Social Realities

In this class of novels and short stories, Thomas Hardy was concerned with depicting the contemporary social issues of his age. Levine [69] argues that 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. That is, Hardy's mind was preoccupied with improving conditions of society. He marks Hardy as a realistic writer who thought his role to express the joys and woes of the victims of merciless conditions of life. Likewise, Reid [70] argues that Hardy's works represent a cry against the excesses of modern civilization and injustices of modern societies.

It can be concluded that concept mining methods can be used for identifying the conceptual similarities within texts and generating reliable typologies of the novels and short stories of Thomas Hardy. Texts were successfully categorized based on their thematic concepts of class conflict, Wessex, religion, female suffering, and social realities. Although the thematic typology of Thomas Hardy based on the concept mining methods agrees in principle with the previous classifications based on fundamental philological approaches, the results reported here are testable, replicable, and thus reliable.

VIII. Conclusion

This study reviewed the literature regarding the thematic typologies of literary texts focusing on the thematic classification of Thomas Hardy's novels and short stories. It was obvious that the typologies of literary texts including those of Thomas Hardy have been traditionally based on philological methods with no consideration of empirical methodologies. With the development of computational approaches, quantitative and statistical methods have come into use. The majority of these typologies, however, are largely based on standard clustering methods or more specifically VSC theory. Despite the effectiveness of this methodology in generating classifications that are based on objective and replicable methods, underlying meanings and concepts were not fully explored. This is attributed to the absence of context in VSC applications which are largely based on 'bag of words' methods. In the face of these limitations, this study proposed a thematic typology based on concept mining methods. The findings indicated that concept mining was usefully used in generating a thematic typology of the novels and short stories of Thomas Hardy that revealed the thematic patterns of the texts. These thematic patterns would be best described in these categories: class conflict, Wessex, religion, female suffering, and social realities. Although the study was limited to the novels and short stories of Thomas Hardy, the results can be extended to other literary texts. It can be finally suggested that the computational and quantitative methods will be central components in the future of thematic typology research.

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 fulfilment of the current research project.

References

[1] R. G. Potter, "Literary Criticism and Literary Computing: The Difficulties of a Synthesis," Computers and the Humanities, vol. 22, no. 2, pp. 91-97, 1988.

[2] S. Gilmartin, Thomas Hardy's Shorter Fiction: A Critical Study. Edinburgh: Edinburgh University Press, 2007.

[3] R. D. Morrison, Thomas Hardy: A Companion to the Novels. McFarland, Incorporated, Publishers, 2021.

[4] S. Gatrell, Thomas Hardy's vision of Wessex. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan, 2003, pp. xvii, 264 p.

[5] P. Gossin, Thomas Hardy's Novel Universe: Astronomy, Cosmology, and Gender in the Post-Darwinian World. London; New York: Routledge, 2017.

[6] W. Lu, Y. Zhou, J. Yu, and C. Jia, "Concept Extraction and Prerequisite Relation Learning from Educational Data," in The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-19), 2019, pp. 9678- 9685: Association for the Advancement of Artificial Intelligence.

[7] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," Journal of Biomedical Informatics, vol. 100, p. 100057, 2019/01/01/ 2019.

[8] M. M. Louwerse and W. van Peer, Thematics: Interdisciplinary Studies. John Benjamins Publishing Company, 2002.

[9] A. García-Berrio, A Theory of the Literary Text. De Gruyter, 2016.

[10] B. Plietzsch, "Hardy's Classifcation of his Works " 2003.

[11] E. Gosse, "Thomas Hardy," in Thomas Hardy; The Critical Heritage, R. G. Cox, Ed. (Critical heritage series, New York: Barnes & Noble. First published in The Speaker (13 September 1890). , 1970, pp. 167-172.

[12] T. Hardy, The works of Thomas Hardy in prose and verse. With prefaces and notes. (Wessex edition.). London: Macmillan & Co, 1912, p. 23 vol.: plates; maps. 23 cm.

[13] B. Plietzsch, The Novels of Thomas Hardy as a Product of Nineteenth Century Social, Economic, and Cultural Change. Berlin: Tenea Verlag Ltd, 2004.

[14] G. Harvey, The complete critical guide to Thomas Hardy (The complete critical guide to English literature). London: Routledge, 2003, pp. x, 228 p.

[15] P. Widdowson, Hardy in history : a study in literary sociology. London ; New York: Routledge, 1989, p. 260.

[16] L. Abercrombie, Thomas Hardy. A critical study. Martin Secker: London, 1912, p. 8º.

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

[18] A. Omar, "Addressing Subjectivity and Replicability in Thematic Classification of Literary Texts: Using Cluster Analysis to Derive Taxonomies of Thematic Concepts in the Thomas Hardy's Prose Fiction," Journal of the Chicago Colloquium on Digital Humanities and Computer Science, vol. 1, no. 2, pp. 1-12, 2010.

[19] A. Omar, Addressing Subjectivity in Thematic Classification of Literary Texts: A Fresh Look at the Prose Fiction of Thomas Hardy. Berlin: Lap Lambert Academic Publishing, 2015.

[20] A. Omar, "Identifying Themes in Fiction: A Centroid-Based Lexical Clustering Approach," Journal of Language and Linguistic Studies, vol. 17, no. Special Issue 1, pp. 580-594, 2021.

[21] M. Ford, Thomas Hardy: Half a Londoner. Harvard University Press, 2016.

[22] S. Shehata, Concept Mining: A Conceptual Understanding Based Approach. Waterloo, Ontario: University of Waterloo, 2009.

[23] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "Using text mining techniques for extracting information from research articles," in Intelligent natural language processing: Trends and Applications: Springer, 2018, pp. 373-397.

[24] G. M. Borkar, L. H. Patil, D. Dalgade, and A. Hutke, "A novel clustering approach and adaptive SVM classifier for intrusion detection in WSN: A data mining concept," Sustainable Computing: Informatics and Systems, vol. 23, pp. 120-135, 2019.

[25] M. Mittal, L. M. Goyal, D. J. Hemanth, and J. K. Sethi, "Clustering approaches for high ‐ dimensional databases: A review," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 3, p. e1300, 2019.

[26] M. Looks, A. Levine, G. A. Covington, R. P. A. L. R. P. Loui, and J. W. Lockwood, "Streaming Hierarchical Clustering for Concept Mining," in Aerospace Conference, 2007 IEEE, 2007, pp. 1-12.

[27] L. Fang, M. Mehlitz, F. Li, and H. Sheng, "Web Pages Clustering and Concepts Mining: An approach towards Intelligent Information Retrieval," Cybernetics and Intelligent Systems, 2006 IEEE Conference, pp. 1-6, 2006.

[28] J. Han and M. Kamber, Data mining : concepts and techniques. San Francisco, Calif. ; London: Morgan Kaufmann, 2001, pp. xxiv, 550 p.

[29] K. Li, H. Zha, Y. Su, and X. Yan, "Concept Mining via Embedding," in 2018 IEEE International Conference on Data Mining (ICDM), 2018, pp. 267-276: IEEE.

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

[31] C. D. Manning, P. Raghavan, and H. Schütze, An Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

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

[33] T. Hardy, A group of Noble Dames. New York: Harper and brothers, 1891, pp. 5 p.l., 292 p.

[34] T. Hardy, Life's Little Ironies. Leipzig: B. Tauchnitz, 1894, p. 295 p.

[35] T. Hardy, Wessex Tales. New York: Harper & brothers, 1896, pp. vii, 290, [1] p.

[36] T. Hardy, A Changed Man, The Waiting Supper, and Other tales, Concluding with The Romantic Adventures of a Milkmaid (Thomas Hardy's works. The Wessex novels.). London: Macmillan & co., 1913, pp. vii, 412, [1] p.

[37] P. Widdowson, "Into the Hands of Pure-minded English Girls": Hardy's Short Stories and the Late Victorian Literary Marketplace," in A companion to Thomas Hardy, K. Wilson, Ed. no. Blackwell companions to literature and culture) Malden, MA: Wiley-Blackwell Pub., 2009, pp. 364-378.

[38] T. Hardy, A Group of Noble Dames, Wessex Edition ed. (The works of Thomas Hardy in Prose and Verse. With prefaces and notes. (Wessex Edition.). [The Wessex novels. II. Romances and fantasies.]). London: Macmillan and Co, 1912, p. 235.

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

[40] E. Gosse. (1928, January 22) Thomas Hardy's Lost Novel. London Times.

[41] T. Coleman, "An Indiscretion in the Life of an Heiress." London: Hutchinson 1976, p.^pp. Pages.

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

[43] R. L. Purdy and M. Millgate, "The Collected Letters of Thomas Hardy (hereafter Collected Letters)." Oxford: Clarendon Press, 1978, p.^pp. Pages.

[44] C. J. Weber, "An Indiscretion in the Life of an Heiress. Hardy's "lost novel" now first printed in America and edited with introduction and notes by Carl J. Weber." Baltimore, MD.: The Johns Hopkins Press. , 1935, p.^pp. Pages.

[45] H. K. Kim, H. Kim, and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation," Neurocomputing, vol. 266, pp. 336-352, 2017/11/29/ 2017.

[46] 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.

[47] A. Omar, "Classifying literary genres: a methodological synergy of computational modelling and lexical semantics," Texto Livre: Linguagem e Tecnologia, vol. 13, no. 2, pp. 83–101, 2020.

[48] S. Agarwal, A. Singhal, and P. Bedi, "Classification of RSS feed news items using ontology," in 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 491-496: IEEE.

[49] F. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong, "News recommendations using CF-IDF," in Proceedings of the International Conference on Web Intelligence, Mining and Semantics 2011 (WIMS 2011), 2011.

[50] F. Goossen, W. IJntema, F. Frasincar, F. Hogenboom, and U. Kaymak, "News personalization using the CF-IDF semantic recommender," in Proceedings of the International Conference on Web Intelligence, Mining and Semantics, 2011, pp. 1-12.

[51] G. Zhu and C. A. Iglesias, "Computing Semantic Similarity of Concepts in Knowledge Graphs," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 1, pp. 72-85, 2017.

[52] I. Clark, Thomas Hardy's Pastoral: An Unkindly May. Palgrave Macmillan UK, 2016.

[53] C. Pickford, The Rural-urban Bind in Thomas Hardy's Regional Novels. Manchester Metropolitan University, 2012.

[54] R. Pite, Thomas Hardy: Selected Writings. Oxford: Oxford University Press, 2021.

[55] K. Snell, The Bibliography of Regional Fiction in Britain and Ireland, 1800–2000. London; New York: Routledge, 2017.

[56] J. B. Bullen, Thomas Hardy: The World of his Novels. Frances Lincoln, 2013.

[57] N. Vance, Bible and Novel: Narrative Authority and the Death of God. Oxford: Oxford University Press, 2013.

[58] R. Franklin, Thomas Hardy and Religion: Theological Themes in Tess of the D'Urbervilles and Jude the Obscure. Eastbourne: East Sussex: Sussex Academic Press, 2021.

[59] D. Fergusson, Faith and Its Critics: A Conversation. Oxford: Oxford University Press, 2011.

[60] R. Morgan, Women and Sexuality in the Novels of Thomas Hardy. London; New York: Routledge, 2006.

[61] M. Kaur, The Feminist Sensibility in the Novels of Thomas Hardy. New Delhi: Sarup & Sons, 2005.

[62] T. R. Wright, Hardy and the Erotic (Macmillan Hardy Studies). Palgrave Macmillan 1989.

[63] M. R. Higonnet, The Sense of sex: feminist perspectives on Hardy. Urbana: University of Illinois Press, 1993, p. 270 p.

[64] P. Ingham, Thomas Hardy: A Feminist Reading Hemel Hempstead: Harvester Wheatsheaf, 1989.

[65] J. Thomas, Thomas Hardy, Femininity and Dissent: Reassessing the Minor Novels. New York: Macmillan, 1999.

[66] M. Jacobus, "Tess's Purity," Essays in Criticism, vol. 26, pp. 318-38, 1976.

[67] M. Jacobus, " Women Writing and Writing about Women." London: Croom Helm, 1979, p.^pp. Pages.

[68] P. Boumelha, Thomas Hardy and Women: Sexual Ideology and Narrative Form. Brighton: Harvester Wheatsheaf, 1982.

[69] G. Levine, Reading Thomas Hardy. Cambridge: Cambridge University Press, 2017.

[70] F. Reid, Thomas Hardy and History. Springer International Publishing, 2017.

AUTHORS' PROFILE

Abdulfattah Omar is an Associate Professor of English Language and Linguistics in the Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University (KSA). Also, he is a standing lecturer of English Language and Linguistics in the Department of English, Faculty of Arts, Port Said University, Egypt. Dr. Omar received his PhD degree in computational linguistics in 2010 from Newcastle University, UK. His research interests include computational linguistics, digital humanities, discourse analysis, and translation studies. ORCID: 0000-0002-3618-1750.