On the Digital Applications in the Thematic Literature Studies of Emily Dickinson's Poetry

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

***Abstract*—Thematic studies in literature have traditionally been based on philological methods supported by personal knowledge and evaluation of the texts. A major problem with studies in this tradition is that they are not objective or replicable. With the development of digital technologies and applications, it is now possible for theme analysis in literary texts to be based at least partially on objective replicable methods. In order to address issues of objectivity and replicability in thematic classification of literary text, this study proposes a computational model to theme analysis of the poems of Emily Dickinson using cluster analysis based on a vector space model (VSM) representation of the lexical content of the selected texts. The results indicate that the proposed model yields usable results in understanding the thematic structure of Dickinson's prose fiction texts and that they do so in an objective and replicable way. Although the results of the analysis are broadly in agreement with existing, philologically-based critical opinion about the thematic structure of Dickinson's work, the contribution of this study is to give that critical opinion a scientific, objective, and replicable basis. The methodology used in this study is mathematically-based, clear, objective, and replicable. Finally, the results of the study have their positive implications to the use of computational models in literary criticism and literature studies. The success of computer-aided approaches in addressing inherent problems in the field of literary studies related to subjectivity and selectivity argues against the theoretical objections to the involvement of computer and digital applications in the study of literature.**

Keywords*—Cluster analysis; digital applications; Emily Dickinson; lexical content; philological methods; thematic studies; Vector Space Model (VSM)

I. Introduction

The analysis of literary texts according to thematic criteria has long been central to literary criticism. There is an established discipline in literary criticism, here referred to as thematic literary criticism (TLC), which studies literary texts in terms of their assessed themes [1-7]. TLC has traditionally been carried out on the basis of philological criteria and/or according to predefined templates or stereotypical classifications. Missing from such studies, however, is any discussion of the issues of objectivity and replicability, or indeed evidence of any awareness that these are issues at all. They are, however, fundamental to all areas of science [8-12]. This study addresses these issues in a literary context in relation to the poetry of Emily Dickinson (1830–1886). The aim is to make some progress towards developing an objective and replicable method for the thematic studies of Emily Dickinson's poetry that can be extended to more general literary classification, overcoming the subjectivity of traditional philological methods. This study builds on work undertaken on Information Retrieval (IR), Automated Text Classification (ATC), and related technologies with the ultimate aim of developing an effective framework for thematic literature studies based on empirical grounds.

In order to identify the thematic structures in the poetry of Emily Dickinson, vector space clustering (VSC) methods are used. VSC is an effective tool for identifying and forming meaningful groups of the objects. The hypothesis thus is that VSC methods can be used in generating an experimentally replicable, objective and conceptually useful analysis based on empirical evidence abstracted from Emily Dickinson's poetry. The remainder of the article is organized as follows. Section 2 is a brief survey of the thematic studies of the poetry of Emily Dickinson. Section 3 describes the methods and procedures of carrying out the computational thematic analysis of the data. Section 4 is analysis and discussions. Section 5 is the conclusion.

II. Literature Review

Different approaches have been developed in the critical study of the thematic structures of literary texts. These include New Criticism, Phenomenology, Structuralism, Deconstructionism, Post-structuralism, Psychoanalysis, Post-Colonialism, Marxism, Feminism, and Historicism [13]. Critics and researchers are usually free to adopt any of these approaches or even adopt their individual style of analysis. In identifying the thematic significance of a given text, a critic may focus on the text, or views it within its larger historical or sociocultural framework. Another critic focuses primarily on economic critique, often exploring how identity is related to social class [14]. Apart from these methodologies, numerous thematic discussions rely heavily either on the author's biographical considerations or even the critic's personal anecdotes, voice, and experience. The problem with such studies is that they are neither objective nor replicable. Regardless of the adopted critical approach, thematic studies of literary works in the philological tradition are in one way or other reflections of the critics' own judgments, which can be affected by personal feelings, emotions, impressions, or prejudices. Moreover, a critic cannot set definite criteria he

*Corresponding Author
Submission Date: May 31, 2020
Acceptance Date: June 13, 2020

used for his classification so that it can be replicated or repeated by another researcher. Even worse, it cannot guarantee that two critics following the same approach. As a result, two readings of a given text can result in completely different interpretations of the same text. It is true thus to suggest that thematic studies in literature have traditionally been based on philological methods supported by personal knowledge and evaluation of the texts [15].

Referring to the literature on the poetic production of Emily Dickinson, thematic studies have been given due attention. Emily Dickinson is one of the most important American poets of the nineteenth century and is considered by many critics as one of America's greatest and most original poets of all time [16]. For many critics, Dickinson's poetry is widely regarded as a milestone in American literature [17]. Dickinson wrote forty volumes of almost 1,800 poems [18, 19]. Many critics argue that Dickinson's poems speak of love, death and nature [20, 21]. One major problem with these studies is that critics have been generally selective in their treatment of the thematic analysis of her poems. They have directed their attention towards particular thematic aspects of her work. For instance, there is a strong body of criticism that confines the works of Dickinson to the subject of death [20, 22-24]. Evidently, many commentators stress the preoccupation morbid in her poetry. Critics generally have focused on her most celebrated works, classified as death poetry. The work on Dickinson is thus best described in terms of its 'selectivity'. Critics have been concerned with particular issues in Dickinson's work and to that end they have tended to select particular pieces of writing for criticism and investigation.

Rommel [25] argues that the problem and limitation of exclusion is accepted as an integral aspect of traditional approaches to textual analysis, and for this reason most literary critics deal with representative textual phenomena when they talk about the surface features of a text. He points out that in the majority of literary critical studies that adopt traditional methods, some kind of textual sampling takes place and critics occasionally make judgments according to the frequency of occurrence or absence of certain textual features. He makes it clear that traditional philological methods are insufficient when dealing with literary texts, since their length makes it too difficult for any traditional approach to measure the frequency of an element efficiently. Rommel concludes that empirical evidence that is truly representative of the whole text is extremely difficult to come by, and mainstream literary scholarship has come to accept this limitation as a given fact. In the face of this limitation, quantitative and computational approaches have been suggested to address the problems of selectivity and lack of objectivity in literary studies. Although these approaches have been most naturally associated with applications related to authorship and style, "they can also be used to investigate larger interpretive issues like plot, theme, genre, period, tone, and modality" [26]. This study seeks to bridge this gap in the literature by looking into computational approaches that address the problems of philological methods of thematic analysis and classification.

III. Methods and Procedures

Recent years have witnessed the development of different computational approaches in document clustering theory. This is a broad framework that includes numerous methods for grouping similar texts together [27-30]. These methods include: vector space clustering (VSC); latent semantic indexing (LSI); concept mining; explicit semantic analysis (ESA); and Netword. The approach that seems most theoretically consistent with our goal, however, is VSC. This is a clustering method whereby texts are clustered into distinct sets based on their semantic similarity [27, 29, 30]. This approach has two steps. Firstly, the relevant documents are mathematically quantified as vectors in high-dimensional space using the vector space model (VSM). Secondly, the similarity between documents is computed using exploratory multivariate analysis (EMVA) methods and hierarchical cluster analysis methods [31, 32]. The rationale is that: (1) the research question directing the present discussion is exploratory, since it is concerned with generating hypotheses about the conceptual structure of Dickinson's corpus; (2) the discussion is concerned with grouping texts of identical/similar themes into distinct sets, which suggests that the idea of analysis is a multivariate data-solving problem [33]; and finally EMVA methods have proved successful in many VSC applications [34]. EMVA encompasses numerous techniques, but for the present purposes cluster analysis is the most appropriate. This is a multivariate mathematical technique for finding relatively homogeneous clusters of cases based on proximity measurements. The rationale of using cluster analysis is that it is the most appropriate technique for organizing a collection. More importantly, cluster analysis methods are used when we do not have any prior hypotheses about the data [29, 35-38]. This serves the principle of objectivity, a primary concern of this research.

In order to support objective and reliable generalizations about Emily Dickinson's poetry, a corpus of all Dickinson's poems (recently collected in *Emily Dickinson's Poems As She Preserved Them* by Cristanne Miller) is built. These are 1775 poems. Dickinson's letters to Susan Gilbert (the woman who was her friend, her muse, mentor, primary reader and editor, fiercest lifelong attachment, and Only Woman in the World) were not included in this study. This study is only concerned with the poetic production of Emily Dickinson. One requirement, however, is that the texts must be pre-processed prior to their representation as data in the corpus. In the current case, the poems were reduced to lists of tokens with only content words retained. That is, function words, like determiners and prepositions, have been removed. 59,378 content-type words were identified in this way, forming the basis for analysis.

Documents were then represented using the vector space model (VSM). This model is both conceptually simple as well as convenient for computing semantic similarities within documents. A data matrix, D*ij*, was built where the rows D*i* represent the documents; the columns D*j* represent the lexical-type variables; and the value of the matrix D*ij* encompasses the frequency of lexical type *j* in document *i*. The data matrix D*ij* was constructed from 59,378 variables representing 1775

poems. As such, each row of the matrix represents a lexical frequency profile for the corresponding text. Because each lexical variable in the profile has a semantic set, the profile gives a representation of what the text is about; what it is not about; and gradations of meaning in-between. However, it should be noted that the matrix 59,378 has some characteristics that could adversely affect the validity of clustering results. Firstly, some poems are long while others are very short. Secondly, its data space dimensionality is so large as to be unwieldy.

To address the variation in text length, the row vectors of the matrix were normalized to compensate for variations in length among texts so that their lexical frequency profiles could be meaningfully clustered. This normalization was related to mean text length using the function:

$$Freq = Freq\ F_i(\frac{\mu}{length\ (i)})$$

The effect of this is to reduce the values in the vectors that represent long documents, while increasing the values of the vectors representing the shorter ones. For documents that are near to or at the mean value, little or no change occurs in the corresponding vectors. The overall effect is to make all the corresponding documents similar in length for the purposes of analysis. As a next step, the problem of high dimensionality was considered. To achieve this, two simple methods of dimensionality reduction were applied. These were: the elimination of relatively low-variance variables; and the retention of highest TF-IDF (term frequency-inverse document frequency) variables.

As shown in Fig. 1, relative variance can now be clearly seen with variables of high variance on the left and variables of low variance on the right. The high-variance variables have to be retained, since they are the main criteria by which the texts can be distinguished from one another. The flat area on the right represents the low-variance variables that contribute little or nothing to distinction between texts—these variables, starting at about 1001 and moving to the right, can be discarded. Variables 1001–59378 were eliminated because of their relatively low variance. The reason for retaining the first 1,000 variables is that these were thought to be the most important for the current analysis. This indicates that a certain amount of subjectivity was at play in determining the number of variables to retain. Finally, TF-IDF was used to identify the most distinctive variables within the dataset. Given that the highest TF-IDF variables are the most important, each column was calculated by means of TF-IDF using the function:

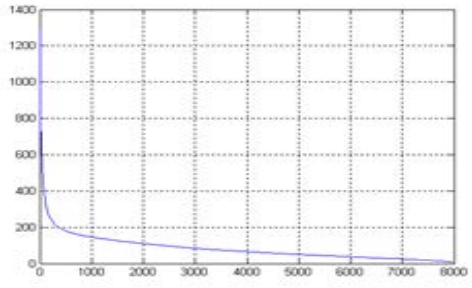

Fig 1. Term Weighting by Variance for the Matrix 59,378.

$$tfid(t_j) = tf(t_j)log_2(\frac{\mathrm{D}}{df_j})$$

Where $tf(t_j)$ is the frequency of *term* t_j across all documents in the data matrix. Using this formulation, the TFIDF of a lexical type A that occurs once in a single document is 1 x $\log_2$ (1000/1) = 9.97; and the TFIDF of a type B that occurs 400 times across 3 documents is 400 x $\log_2$ (1000/3) = 3352. As can be seen in this example, B is far more useful for document differentiation than A, which is more intuitively satisfying than the alternative. The variables are sorted in descending order, as shown in Fig. 2.

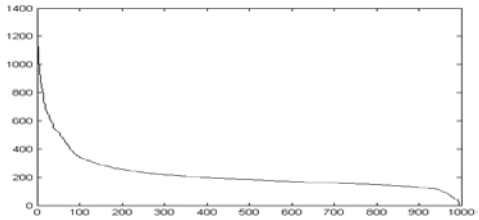

Fig 2. An Illustration of TF-IDF Term Weighting.

As can be seen in the Fig. 2, variables 1–200 were retained and variables 201–1000 were removed. The result is a transformed data matrix 200, which provided the basis for subsequent analysis.

IV. Analysis and Discussions

Agglomerative hierarchical cluster analysis methods were used to find meaningful clusters in the data, which can be used to empirically derive the structure of the thematic concepts of the poetry of Emily Dickinson. The data matrix was hierarchically analyzed first using Ward linkage clustering (or what is usually referred to as the increase in the sum of squares) with the Euclidean distance between points. This is the most suitable method for our analysis because it allows the clearest partitioning of the matrix rows. Ward's method of clustering allows us to discover useful associations and meaningful groupings in the dataset. Hierarchical cluster analyses are presented in the form of diagrams known as dendrograms. These are visual representations of cluster structures that show how clusters are related to each other, which clusters are merged or fused at each stage of the analysis, and how the distance between them is calculated at the time of their merging or fusion [39].

One advantage of this clustering method shown in Fig. 3 is that it offers a solution to a common problem in cluster analysis—how to decide on the optimal number of clusters to fit a dataset. The strong tendency towards left branching associated with other clustering methods is avoided in Ward clustering. The matrix rows are assigned to three main groups, which are assigned as groups A, B, and C. For clustering validation purposes, a cross-validation approach was used. The texts were randomly divided into two subsets, *A* and *B*, and cluster analysis was carried out separately on each. The level of similarity in the results indicates validity [40]. Comparison shows a close fit between the results of hierarchical analysis. There is no contradiction between the results of the two clustering structures.

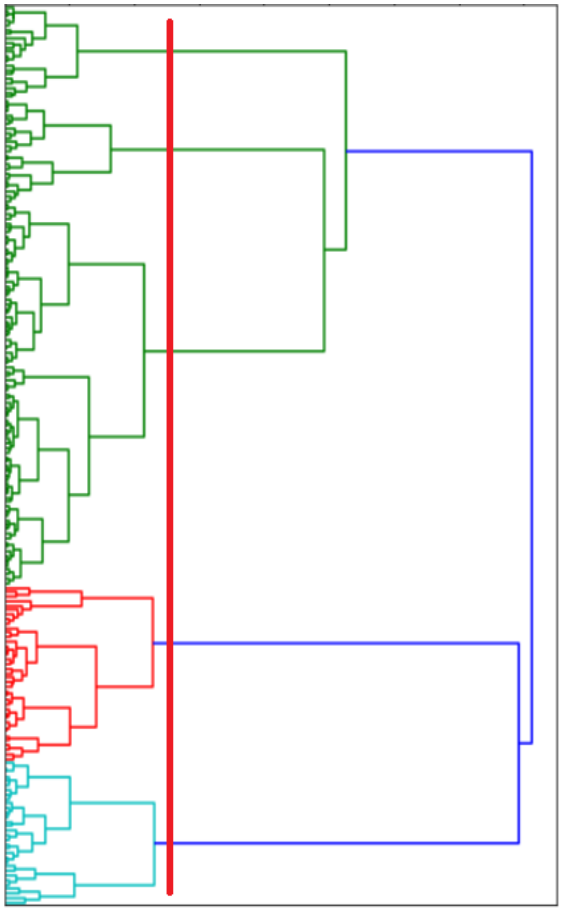

Fig 3. Hierarchical Cluster Analysis of Dickinson's Matrix using Euclidean Distance & Ward Linkage Clustering.

Given that the texts were clustered on the basis of lexical frequency vectors, each cluster has a characteristic lexical frequency profile that distinguishes it from the others. Based on this assumption, it should be possible to identify the most important variables for each group, and, on the basis of the lexical semantics of these items, to infer thematic characteristics of the respective groups. To do this, a centroid analysis was carried out. A centroid is the center of a given geometric figure. Centroid vectors were constructed by means of the vectors in the Dickinson matrix constituting the four groups A-C in accordance with the function:

$$V_i = \frac{\sum_{i=1}^{m} Dij}{m})$$

Where:

V $_j$ is the $_j$th element of the centroid vector (for $_j$ = 1 … the number of columns in D);

D is the Dickinson data matrix, and;

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

The resulting vector groups were compared to show how, on average, the three groups differed for each of the 180 lexical variables. The aim was to identify the variables in which the difference was greatest and the thematic characteristics of each group can then be inferred.

Group A comprises 894 poems including, "Because I could not Stop for Death, It was not Death, for I stood up, I Heard a Fly Buzz when I Died, and I felt a Funeral in my Brain." This group is characterized by words like *die*, *death*, *funeral*, *soul*, *Heaven*, *clergyman*, *Father*, *Christ*, *God*, and *immortality*. These are frequently used in the poems of Group A. Correlating the results of the lexical profiles above with some knowledge about these texts, it can be observed that they are concerned with idea of death.

The most distinctive lexical features of Group B, in turn, are *sea*, *feathers*, *bird*, *storm*, *wild*, *light*, *woods*, *valley*, *world*, *nature*, *dew*, *flower*, *summer*, *shower*, *bee*, *garden*, *Grass* as well as colors names such as *yellow* and *purple*. Poems of this group include, "A Dew Sufficed Itself, A Service of Song, May Flower, My Garden, Summer Shower, The Bee is not afraid, The Grass, The Purple Clover, and The Sea of Sunset." Based on the lexical-semantic features of these words, it can be suggested that they are concerned with nature.

Finally, Group C included 628 poems including, "That I did always love, Heart We Will forget him, I Cannot Live Without You, You Left Me, and I know that he exists." The most distinctive lexical features of this group are *sweet*, *love*, *heart*, *beloved*, and *charm*. It can be suggested thus that these poems are centered on the theme of love as reflected in the lexical-semantics of the words.

It can be finally concluded that the clustering structures identified in this study correspond in principle to the classification of Dickinson's poetry in the philological tradition of literary criticism outlined earlier. It can be claimed however that quantitative and computational approaches to literature provide accurate and acceptable methods of classification and analysis [41]. Furthermore, these approaches, using scientific and objective methodologies, can be used in the service of traditional literary studies to help critics cope with the huge amount of electronic text now becoming available [42, 43].

V. Conclusion

Computational analysis of Emily Dickinson's poetry has yielded a replicable, objective, and conceptually useful thematic structuring of her works. Although the results of the analysis are broadly in agreement with existing, philologically-based critical opinion about the thematic structure of Dickinson's work, the contribution of this study is to give that critical opinion a scientific, objective, and replicable basis. The methodology used in this study is mathematically-based, clear, objective, and replicable. It has been shown to be effective in the literary study of Dickinson's work and is thus potentially applicable in literary scholarship more generally. Quantitative and computational methods can be used to empirically derive taxonomies of thematic concepts of the poetry of Emily Dickinson.

Equally important, nonetheless, computers and machines cannot be replacements for humans in terms of reading and interpreting literature. I suggest that the computational element in literary criticism can develop concrete evidence to support or refute hypotheses or interpretations that have in the past been based on personal readings and the somewhat

serendipitous noting of interesting features. In other words, what computational methods give us is an objective clustering giving insight into alternative interpretations based on criteria that are definitely in the text and which constrain our subjective interpretations. This is the main point of this study. It does not claim that this method is better or replaces all human interpretations of literary texts, but rather it constrains human subjective interpretation by presenting classification criteria that must be taken seriously precisely because they are objective and replicable. The clustering results of this study can serve as a base for future studies and criticisms of the thematic analysis of Emily Dickinson's poetry.

Finally, the results of the study have their positive implications to the use of computational models in literary criticism and literature studies. The success of computer-aided approaches in addressing inherent problems in the field of literary studies related to subjectivity and selectivity argues against the theoretical objections to the involvement of computer and digital applications in the study of literature.

ACKNOWLEDGMENTS

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

REFERENCES

[1] M. E. Atwood, Survival: a thematic guide to Canadian literature. Toronto: Anansi, 1972.

[2] V. H. Brombert, Novels of Flaubert: A Study of Themes and Techniques. Princeton: New Jersey: Princeton University Press, 2015.

[3] F. Hammill, Canadian literature (Edinburgh critical guides to literature). Edinburgh: Edinburgh University Press, 2007.

[4] M. L. Jockers and D. Mimno, "Significant themes in 19th-century literature," Poetics, vol. 41, no. 6, pp. 750-769, 2013/12/01/ 2013.

[5] W. R. Sanborn, The American Novel of War: A Critical Analysis and Classification System. Jefferson, North Carolina; London: McFarland Incorporated Publishers, 2012.

[6] W. Sollors, The return of thematic criticism (Harvard English studies). Cambridge, Mass.: Harvard University Press, 1993.

[7] T. Todorov, The fantastic : A Structural Approach To A Literary Genre. Ithaca, N.Y.: Cornell University Press, 1975.

[8] B. L. Berg, Qualitative research methods for the social sciences. Boston: Allyn and Bacon, 1998.

[9] I. Holloway, Basic concepts for qualitative research. Oxford: Blackwell Science, 1997.

[10] R. Gomm, Key concepts in social research methods (Palgrave key concepts). Basingstoke: Palgrave Macmillan, 2009.

[11] G. Payne and J. Payne, Key concepts in social research. London: SAGE, 2004.

[12] M. Q. Patton, Qualitative Research & Evaluation Methods, 3rd ed ed. London: Sage, 2002.

[13] L. Tyson, Critical Approaches to Literature. London; New York: Routledge, 2018.

[14] W. Sollors, The Return of Thematic Criticism. Harvard University Press, 1993.

[15] 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," Journal of the Chicago Colloquium on Digital Humanities and Computer Science, vol. 1, no. 2, pp. 1-14, 2010.

[16] B. Steffens, Emily Dickinson. Lucent Books, 1998.

[17] G. Grabher, R. Hagenbüchle, and C. Miller, The Emily Dickinson Handbook. University of Massachusetts Press, 1998.

[18] R. Gray, A Brief History of American Literature. Wiley, 2010.

[19] C. Miller, "Emily Dickinson's Poems As She Preserved Them." Harvard University Press, 2016.

[20] W. Martin, The Cambridge Companion to Emily Dickinson. Cambridge Cambridge University Press, 2002.

[21] N. Tandon and A. Trevedi, Thematic Patterns of Emily Dickinson's Poetry. Atlantic Publishers & Distributors, 2008.

[22] M. Dauben, "Emily Dickinson" - The Death Motif in the Poetry of Emily Dickinson. GRIN Verlag, 2010.

[23] N. Dietrich, Emily Dickinson's Death Poetry. GRIN Verlag, 2003.

[24] B. Lindberg, "The theme of death in Emily Dickinson's poetry," Studia Neophilologica, vol. 34, no. 2, pp. 269-281, 1962.

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

[26] D. L. Hoover, "Quantitative Analysis and Literary Studies," in A Companion to Digital Literary Studies, R. G. Siemens and S. Schreibman, Eds. Malden, MA: Blackwell Publishers, 2013, pp. 517-533.

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

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

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

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

[31] F. Husson, S. Le, and J. Pagès, Exploratory Multivariate Analysis by Example Using R. CRC Press, 2017.

[32] A. Lüdeling and M. Kytö, Corpus Linguistics (no. v. 2). De Gruyter, 2009.

[33] R. Adams, "Perceptions of innovations: exploring and developing innovation classification," PhD, School of Management Cranfield University, 2003.

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

[35] M. R. Anderberg, Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Elsevier Science, 2014.

[36] E. J. Bynen, Cluster analysis: Survey and evaluation of techniques. Springer Netherlands, 2012.

[37] B. S. Duran and P. L. Odell, Cluster Analysis: A Survey. Springer Berlin Heidelberg, 2013.

[38] C. Hennig, M. Meila, F. Murtagh, and R. Rocci, Handbook of Cluster Analysis. CRC Press, 2015.

[39] A. Fielding, Cluster and Classification Techniques for the Biosciences. Cambridge, UK; New York: Cambridge University Press, 2007.

[40] A. C. Rencher, Methods of Multivariate Analysis, Second Edition ed. John Wiley & Sons, INC, 2002.

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

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

[43] S. Zyngier, M. Bortolussi, A. Chesnokova, and J. Auracher, Directions in Empirical Literary Studies: In honor of Willie van Peer. John Benjamins Publishing Company, 2008.