Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327
مجلة اللغة الانكليزية في العالم العربي

Arab World English Journal (AWEJ) Volume 11. Number1 March 2020 Pp. 195-211
(1) 181-194. DOI: https://dx.doi.org/10.24093/awej/vol11no1.16

The Detective and Sensation Fiction of Wilkie Collins: *A Computational Lexical-Semantic Analysis*

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

Abstract
Theme and genre classifications in the works of Wilkie Collins (1824-89) have been extensively investigated using different literary approaches; these are usually based on textual content and biographical considerations. Different critics place Collins' works under the two main headings of detective fiction and sensation fiction. Such analyses have been generated by what is referred to as the 'philological method'; that is, by an individual critic's reading of the relevant material and their intuitive abstraction of generalizations from that reading. A problem with such an approach is that it is not objective, and it is therefore unreliable. The research question is thus asked in response to the subjectivity of previous genre classifications of the novels of Wilkie Collins and the lack of agreement among literary critics and researchers about such classifications. As such, I ask whether an objective and conceptually useful reading of the themes and subjects of Wilkie Collins' prose fiction texts can be developed. *As thus*, computational lexical-semantics is suggested to understand the issues of thematic classification. For this purpose, vector space clustering (VSC) was used for capturing the lexical-semantic features of his novels and linking them explicitly to the relevant themes and genres. It is suggested that through this method, an objective, replicable, and reliable genre classification of Collins' novels is possible. The results of this study can serve as a basis for future studies and criticisms of Wilkie Collins' fiction.

***Keywords*:** computational lexical-semantics; detective fiction; genre classification; sensation fiction; theme analysis; vector space clustering (VSC); Wilkie Collins

Cite as: Omar, A. (2020). The Detective and Sensation Fiction of Wilkie Collins: A Computational Lexical-Semantic Analysis. *Arab World English Journal*, *11* (1) 195-211.
DOI: https://dx.doi.org/10.24093/awej/vol11no1.16

1. Introduction

Theme and subject in the works of Wilkie Collins (1824-89) have been extensively investigated using different literary approaches; these are usually based on textual content and biographical considerations. Different critics place Collins' works under the two main headings of detective fiction and sensation fiction. Such analyses have been generated by what is referred to as the 'philological method'; that is, by an individual critic's reading of the relevant material and their intuitive abstraction of generalizations from that reading. A problem with such an approach is that it is not objective, and it is therefore unreliable. *The idea of objectivity has long been a central issue in the discussion of literary works. Different critics have concerned themselves with identifying new approaches and have suggested new grounds and methodologies for dealing with literary texts. A number of them have found new avenues for addressing problems of traditional literary criticism in the methods of computational linguistic, especially with the development of electronic text formats that* permit the application of computational data analysis concepts and procedures. *Work of this kind is often classified under the broad heading of digital humanities: researchers use computational methods* to either answer existing research questions or to challenge existing theoretical paradigms; in doing so, they generate new questions and pioneer new approaches (Berry, 2012). The purpose of this study thus is to see if such concepts and procedures, and more specifically those which constitute the class of computational lexical-semantic methods, can usefully supplement philological methods in the thematic analysis of the prose fiction writings of Wilkie Collins.

Despite the effectiveness of computational methods in literary studies, including work on thematic analysis, genre classification, authorship attribution, and stylistics, a search of the literature reveals that there has been no computer-aided thematic analysis or classification of the works of Wilkie Collins. This is the case with many writers in English whose review through computational methods is very limited. One reason for this may be due to the opaque language of computational methods for researchers in literature. Ramsay (2003) suggests that "the inability of computing humanists to break into the mainstream of literary critical scholarship may be attributed to the prevalence of scientific methodologies and metaphors in humanities computing research" (167). In light of this limitation, this study seeks to address the gap between literary research and computational analysis by generating an automated classification of Collins' prose fiction works with the purpose of providing an objective analysis of the themes and subjects in his novels. The research question is asked in response to the subjectivity of previous genre classifications of the novels of Wilkie Collins and the lack of agreement among literary critics and researchers about such classifications. As such, I ask whether an objective and conceptually useful reading of the themes and subjects of Wilkie Collins' prose fiction texts can be developed.

The goal of this study is to see if the concepts and procedures of computational linguistics, and more specifically computational lexical semantics, being the process of decoding meanings within texts, can usefully supplement philological methods in the genre classification of the prose fiction of Wilkie Collins. To do this, lexical-semantic analysis using vector space clustering (VSC) methods has been *applied to* find the similarities and/or dissimilarities within the texts; the purpose is to capture the lexical-semantic features of his novels and link them explicitly to the relevant themes and genres.

2. Statement of the problem

The established literature on Collins is saturated with stereotypical criticisms and patterns that are not grounded in empirical criteria. One major problem of such a tendency is that thematic discussions about Collins are mostly confined to a few major themes, such as detective fiction and sensation fiction, paying little attention to other thematic concepts. It has been a commonplace of thematic reviews of Collins to consider that he is the pioneer of detective and sensation fiction (Gasson & Peters, 1998; Mangham, 2008; Page, 2002; Pykett, 2005; Taylor, 2006). Different critics assert that detective and sensation elements are the most dominant themes of his novels and short stories. Others, however, have reflected on his novels and short stories as being a product of nineteenth-century social, economic, and cultural life. The problem with such analyses is that they are not based on empirical grounds. The majority of these criticisms are, to a great extent, stereotypical and have no objective grounds.

Another problem with previous thematic classifications of Collins' fiction is that they are selective. That is to say, many critics select only the novels and short stories that support their arguments and classification. So far, there is no single work that has taken the full measure of the themes of Collins' total corpus of novels and short stories. This may be due to the idea that many classifications of Collins have built on the earlier classification of his work into detective and sensation fiction. The result of this is that many discussions based on philological methods show significant bias. Very often, they are based on certain concepts within the school of the critic or on their personal evaluation. In many cases, the data that critical approaches use is not representative. Rommel (2004) argues that problems and limitations of selectivity and exclusion are integral aspects of traditional approaches to analyzing literary texts. He adds that such approaches lack any empirical evidence, and this is why the main bulk of traditional literary criticisms are located in mainstream literary scholarship, which accepts such problems and limitations as givens. Rettberg (2016) indicates that one way to address such issues and limitations is the integration of computational approaches into critical literary studies. He adds that computational approaches help researchers and critics employ empirical methods and explore much larger amounts of data in systematic ways than traditional approaches.

3. Literature Review

Computational lexical semantics has been used in a number of different applications, including information retrieval, text documentation, and authorship detection (Pustejovsky, 2012; Saint-Dizier et al., 1995; Storjohann, 2010). In spite of its effectiveness in addressing various problems, applications of computational lexical semantics are still very limited. This may be attributed to the unfamiliarity of the field of computational theory and its methodologies to literary critics. This unfamiliarity has led many literary scholars and critics to consider them somehow alien to literary studies. This can explain the gap we see between critical literary theory, on one hand, and computer-based text analysis and quantitative approaches on the other: the majority of critical theory researchers have never argued for the need of computational approaches to supplement widely used critical approaches (Finneran, 1996; Potter, 1989; Schreibman, Simens, & Unsworth, 2013). With increasing access to e-texts and greater availability and power of computational tools, however, there has been growing interest in literary computing studies for text analysis and interpretation with methods of computational lexical semantics being widely used. There have been a number of different applications, including authorship attribution, stylometric analysis,

thematic analysis, genre classification, characterization, and textual analysis. The focus of this study is thematic analysis.

Thematic analysis is a fundamental discipline in literary criticism studies. Nevertheless, the definition and practice of thematic analysis are not yet settled. There is no agreement on the definition of the term 'theme' itself. Some think of the theme as the moral or lesson of a literary work. Others hold to the idea that the theme represents the main idea in a literary work. While a variety of definitions of the term have been suggested, this study considers the theme as a pattern of meaning or threads of ideas giving an overall picture of a literary work. Thematic reviews have previously been done using non-computational methods. With the development of computational approaches, scholars have come to think about how effective computational approaches are in identifying meanings within texts and, it is suggested by literary computing researchers that computational lexical-semantic approaches have proved effective in improving the understanding of literary texts (Rockwell, 2003). Despite the relative success of studies of this kind, they have been met with strong objections from a number of critics and scholars. Some think that such methods of interpreting texts are still far from successful at detecting what a text is about exactly (Corns, 1991; Rommel, 2004).

The literature suggests that computational lexical-semantic methods of analyzing texts are central to computer-based applications on thematic analysis (Argamon & Olsen, 2006; Yu, 2008). The main assumption is that methods of computational lexical-semantic analysis are effective at identifying what a text is about. Consequently, thematic hypotheses can be based on clustering results. Two good examples supporting this position are Plaisant et al. (2006) and Horton et al. (2006). Plaisant et al. offer a thematic analysis of eroticism in Emily Dickinson's correspondence. The discussion on eroticism in the author's correspondence has been one of the most controversial and important debates about Dickinson of the twentieth century. The study analyzed a corpus of about 300 XML-encoded letters comprising nearly all the correspondence between Emily Dickinson and her sister-in-law Susan Huntington. The authors conclude that computational techniques were effective in generating new insights and ideas. This study, above all, offers a system to support humanities scholars in their interpretation of literary works (Plaisant, Rose, & Yu, 2006). Horton et al. (2006) undertake a thematic review of sentimentalism in early American novels. The researchers looked at how a collection of texts can exhibit certain thematic features using five American novels in the case study, including Stowe's *Uncle Tom's Cabin* (1897) and *The Minister's wooing* (1859); Jacobs *Incidents in the Life of a Slave Girl* (1861); and Rowson's *Charlotte: a Tale of Truth* (1794) and *Charlotte's Daughter* (1828). The novels were divided into 184 chapters and the level of sentimentality evident was assessed in each. The authors concluded that 95 chapters were highly sentimental while 89 showed low sentimentality.

Similarly, a number of studies have adopted computational lexical-semantic methods for investigating thematic interrelationships in Shakespearean texts (Jockers, 2009; Ramsay, 2005, 2007). Ramsay (2005) holds that computational tools provide objective criteria that can serve to adjudicate some of the problems of thematic analysis and objectively assign appropriate themes to literary texts. He used computational lexical-semantic methods to generate hypotheses about the thematic interrelationships in Shakespeare's plays. The plays were grouped into four distinct clusters: comedy, tragedy, history, and romance. He reported that comedies and histories clustered

together very well, but it was hard to distinguish romance from tragedy. Ramsay admitted that the results were not wholly convincing. However, he also stressed the importance of thinking about objective criteria in the thematic analysis of literature.

The main problem with such applications relates to feature selection—the use of distinctive lexicons that can express the meanings of texts. In the face of this limitation, this study proposes the use of a hybrid complex of statistical measures, including frequency analysis, variance analysis, term frequency-inverse-document frequency (TF-IDF), and principal component analysis (PCA) in sequence to selecting the most distinctive features for generating reliable document clustering, which can be usefully used in thematic analysis tasks.

4. Methodology

4.1. Methods

There are several approaches that have already been explored in computational lexical-semantic analysis, particularly in the study of fictional and prose literary works. These include non-negative matrix factorization, vector space clustering, latent semantic analysis, self-organizing maps, and locality-preserving projection. In this study, the vector space clustering (VSC) model was used. The rationale for using VSC is that it is one of the most popular methods for data representation in document clustering applications and is still suitable for the majority of clustering purposes.

In VSC applications, a vector space model (VSM) is usually built. This is a technique whereby documents are compared to each other then indexed or classified in terms of their similarity or distance based on the words they contain. The underlying procedure in VSM involves the initial extraction of all useful information within a document collection and its recording in an index known as a vector space. A proximity measurement is then used to compute the semantic similarity among the documents with the purpose of grouping similar documents together. This is done by measuring the relative distances between the row vectors. The distance between any two vectors in a space is jointly determined by the size of the angle between the lines joining them to the origin of the space's coordinate system and by the lengths of those lines.

4.2. Data collection

To support reliable generalization about the thematic content of the novels of Collins, the corpus has to be both large and representative. As such, this study is based on a corpus comprising all the novels of Wilkie Collins. Thirty texts were downloaded from two online sources: (1) the Gutenberg Project and (2) the Wilkie Collins Pages Website. Prior to data processing, the texts were subjected to four processes: cleaning up, removing function words, executing stop word lists, and stemming. The texts were first cleaned of all punctuation marks and non-alphabetic characters. All function words were then removed because the focus is on keywords, which are lexical items. Thirdly, the texts were stemmed, i.e. suffixes were removed from the ends of words. Finally, stop word lists were generated and executed. The lists included many recurring lexical items, with little to add to thematic categorization such as proper names, titles, and reporting verbs.

4.3. Procedures

A data matrix, M, is abstracted from the corpus: the rows, *i*, represent texts; the columns, *j*, represent lexical types occurring across all the texts; and the value at M*ij* is the frequency of

occurrence of lexical type *j* in text *i*. Each matrix row vector, therefore, represents a lexical frequency profile for the corresponding text. Because each lexical variable in the profile has semantic content, the profile gives a representation of what a text is about, what it is not about, and gradations in between. The matrix, M, consists of 30 rows and 22,801 columns. At this stage, the matrix suffers from two problems. First, some texts are very long, while others are very short. Second, the dimensions are so big that it is impossible for any cluster analysis to produce meaningful results. As such, M needs to be transformed.

4.3.1. Compensation for variation in text length

In clustering applications, document length plays an important role in grouping similar texts together. Measuring the similarity within texts can be greatly influenced by vectors that have the largest values. It is expected that the proximity measurements will be dominated by longer documents. In a VSM, the distance between any two documents is determined by their length and the magnitude of the angle between the vectors. This means that if the length of the document increases, the number of times a particular term occurs in the document will also increase. Consequently, length becomes an increasingly important determinant of vector clustering in the space and, vice versa—if the documents are short, the angles between the vectors become smaller and short documents will be clustered together.

One way to resolve this problem is through the normalization of text length: the length of shorter texts is compensated for by means of standardizing the vectors, so that all documents in a matrix are represented equally. This is a way of penalizing the term weights for a document in accordance with its length (Amati & Rijsbergen, 2002; Robertson & Walker, 1994; Singhal, Chris, & Mandar, 1996). For the purposes of this study, the cosine normalization method and Pearson's correlation analysis were used. Cosine normalization is one of the most commonly applied techniques in the vector space model (Rijsbergen, 1979; Salton & Buckley, 1987; Singhal, Chris, et al., 1996; Singhal, Salton, Mitra, & Buckley, 1996). The underlying principle of cosine normalization is that all documents in a given collection are represented equally. In this process, all row vectors of the matrix are transformed so that they have unit length and are made to lie on a hypersphere of radius 1 around the origin, ensuring that all vectors are equal in length. Accordingly, variation in the length of documents and, correspondingly, of the vectors that represent them is no longer a factor (Moisl, 2009).

Pearson's correlations analysis of words sampled from the 30 texts was undertaken to develop the results as shown in Table One.

Table 1. *Results of Pearson's Correlation Coefficient Analysis*

		Uniqueness	**Words**
Uniqueness	**Pearson's Correlation**	**1**	**0,691**
	Sig. (2-tailed)		**,000**
	N	**29**	**29**
Words	**Pearson's Correlation**	**0,691**	**1**
	Sig. (2-tailed)	**,000**	
	N	**29**	**29**

The correlation analysis used vectorised measurements of words chosen from 30 texts; it also represents each text's uniqueness against its word count, in terms of length and frequency. As such, Pearson's method of correlation provides an approximate word measure for Willkie Collins' novels in each category. This measure can then be correlated to the length of each text: the longer the text, the lower the relative occurrence of a particular set of words. This phenomenon is explained further in the next section on clustering validation.

4.3.2. Dimensionality Reduction

In VSM, the high dimensionality of text data is a major issue and has a negative impact on the reliability and accuracy of almost all document clustering applications. The larger the data dimensionality, the more difficult it becomes to define the manifold sufficiently well to achieve reliable analytical results. A good clustering should be based only on the most important terms "because irrelevant and redundant words often degrade the performance of classification algorithms both in speed and classification accuracy" (Novovičová, Malík, & Pudil, 2004, p. 1010). The analysis is concerned only with the distinctive lexicons that can represent the texts on semantic grounds. This means that if a word is used frequently, even for more than a thousand times in one text, it cannot be taken into account in the analysis. In order to select only the distinctive lexical features that can help in generating hypotheses about the theme and subject of the selected texts, a hybrid approach of reductive methods is used. These include frequency analysis, variance analysis, term frequency-inverse-document frequency (TF-IDF), and principal component analysis (PCA).

A frequency analysis was carried out. The hypothesis was that very infrequent words are of little importance in making generalizations about the selected texts. It was decided, therefore, that words with frequencies of less than ten should be removed. This reduced the number of vectors from 22,801 to 13,450. This was followed by a variance test. It was revealed that the first 1,000 columns were the most varied. As such, the vectors from 1-2,000 were retained and the other vectors were deleted, as is shown in Figure One.

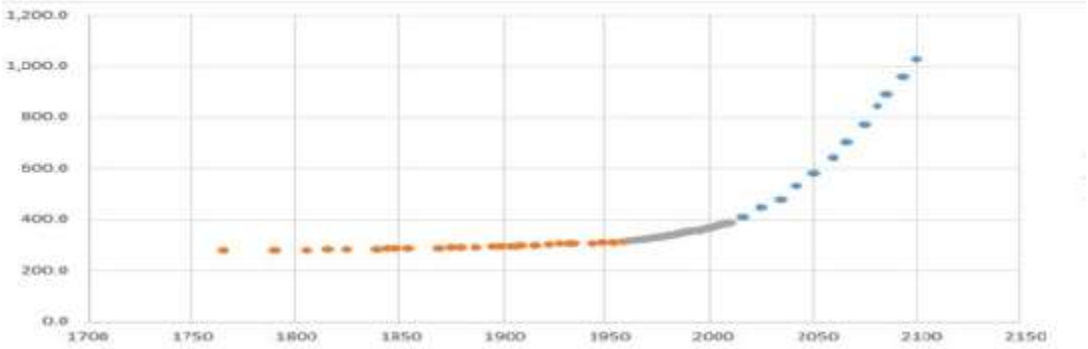

Figure 1. Variance Analysis of the Matrix Collins 30, 13450

Next, a TF-IDF test was carried out. This is the most common method of calculating term frequency. In our case, it was decided that only the highest TF-IDF values should be retained, allowing us to reduce the original matrix to 400 variables. This is shown in Figure Two.

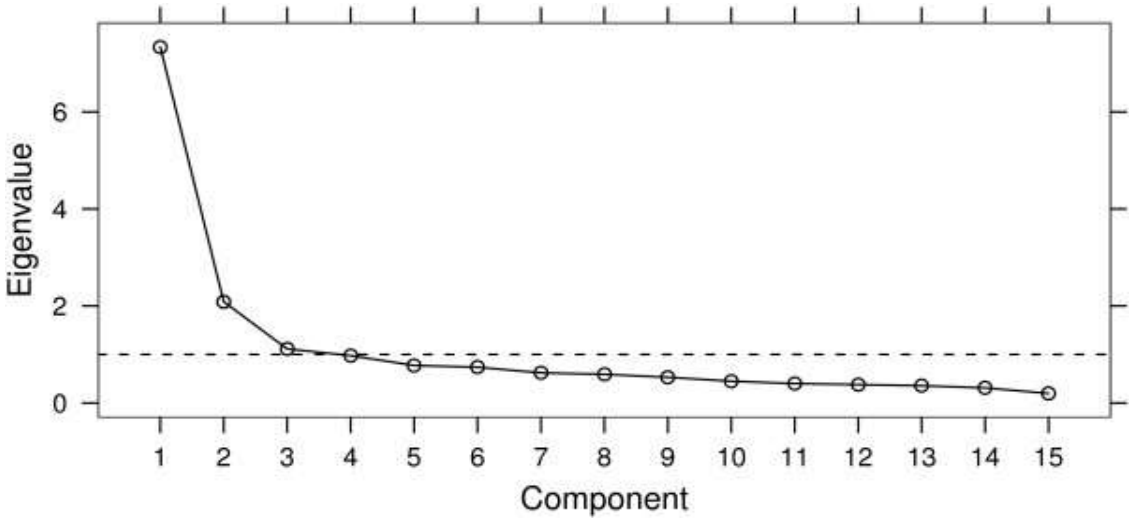

Figure 2.A TF-IDF Analysis of Collins Matrix 30, 2000

Finally, PCA was applied. PCA is a basic geometric tool used to produce a lower-dimensional description of the rows and columns of a multivariate data matrix (Härdle & Simar, 2003; Jackson, 1991). The main function of PCA is to find the most informative vectors within a data matrix. As Jolliffe (2002) explains: "The central idea of PCA is to reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data sets" (p. 1). It can be described as a technique for data quality (Jackson, 1991). Using a scree plot representation of the highest TF-IDF values (as shown in Figure Three), it may be agreed that components 1-70 are probably meaningful and components 71-400 are probably trivial. As such, variables 1-70 were retained and 71-400 deleted.

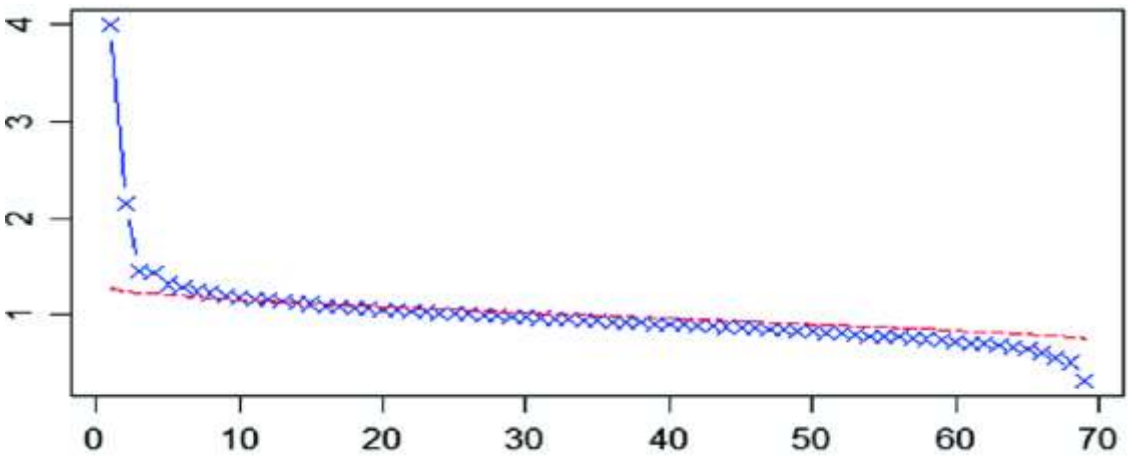

Figure 3. A TF-IDF Analysis of Collins 30, 400

5. Analysis and Discussions

VSC methods were used to group the 30 texts according to the most distinctive lexicons generated above. Figure Four presents a tree with the document titles corresponding to the row vectors and its leaves uniting into large clusters containing subordinate clusters. The large clusters are blended

into a single category containing all or some part of one or more row vectors. The lengths of each horizontal line presented in each category are related to the text clusters or the number of words—the longer the line, the greater the dissimilarity. The hierarchical clustering shown in Figure Four gives the assumption that texts in each group or cluster will have something in common that makes them similar to each other and different from other texts in other groups. Hierarchical cluster analysis is the most common method of generating a model of clusters; this method works successfully even with variables in opposition. This method can cluster variables together just like factor analysis.

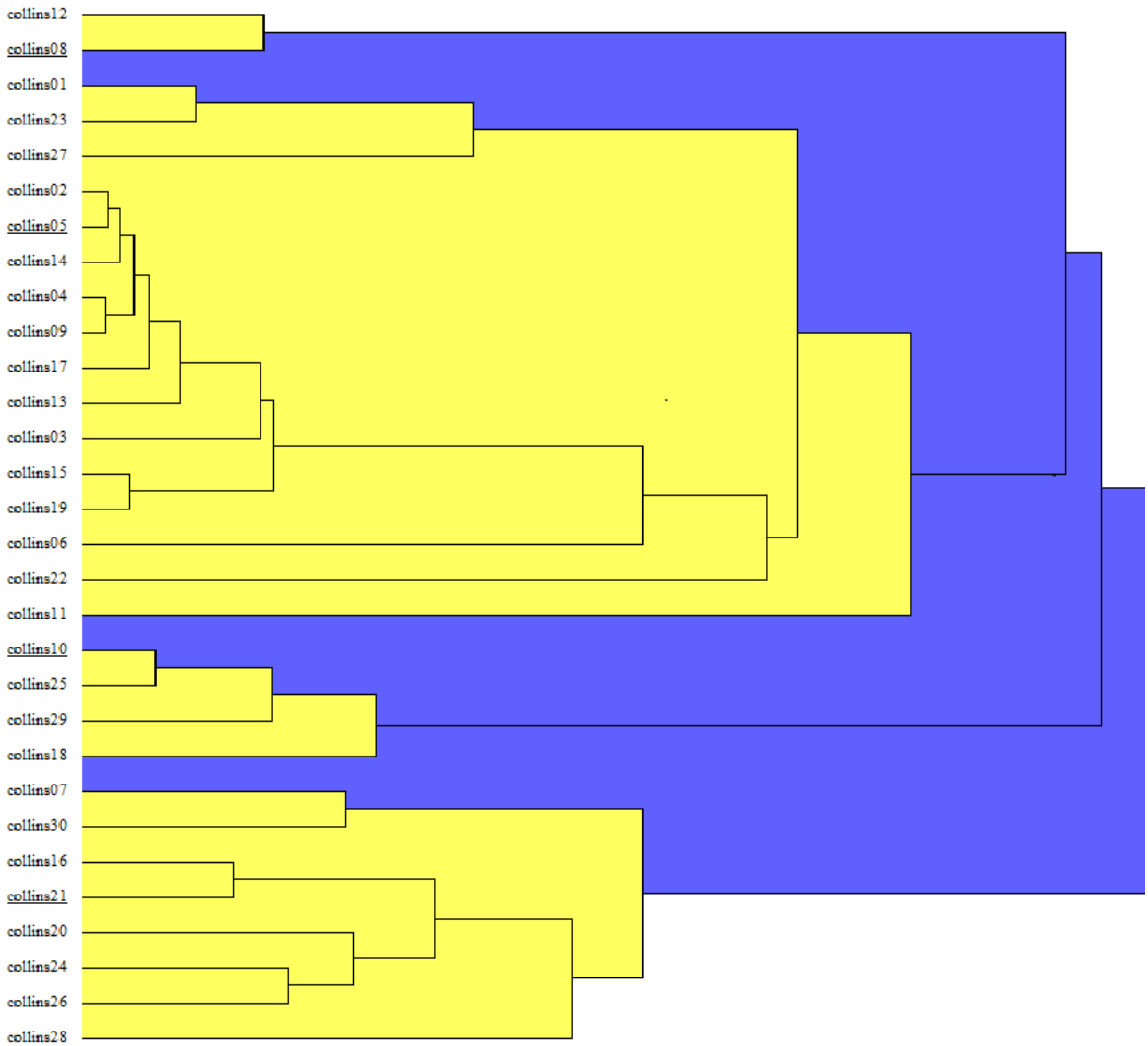

Figure 4. VSC of Collins' Novels Using Hierarchical Cluster Analysis Methods

The validity of any type of cluster analysis depends primarily on the methodology adopted. In literary cluster analysis, varied clustering methods may generate different structures potentially affecting the reliability of the results. To ensure validity of the clustering, cross-validation or relative comparison methods are used. In a cross-validation approach, the texts are randomly

divided into subsets or groups and the cluster analysis is carried out separately on each group. The results of such cluster analysis can indicate its validity (Rencher, 2002) if the relative approach is based on comparing the clustering structure, generated by the same algorithms, but using an alternative representation of the data. Cross-validation shows a close fit between the clustering structures. Specifically, there is a total correspondence between the structures based on the data matrix composed of all the 30 rows and the structures based on the random distribution of these 30 rows into four groups, as shown in Table Two.

The hierarchical clusters are shown in Table Two as four sub-clusters, labeled as Group 1, Group 2, Group 3, and Group 4.

Table 2. *Cluster Analysis of Wilkie Collins' Fictional Prose Texts*

Cluster	Members
Lexicon Group 1	Collins 12, Collins 08
Lexicon Group 2	Collins 1, Collins 23, Collins 27, Collins 02, Collins 05, Collins 14 Collins 04, Collins 09, Collins 17, Collins 13, Collins 03, Collins 15, Collins 19, Collins 06, Collins 22, Collins 11
Lexicon Group 3	Collins 10, Collins 25, Collins 29, Collins 18
Lexicon Group 4	Collins 7, Collins 30, Collins 16, Collins 21, Collins 20, Collins 24, Collins 26, Collins 28

Table Two illustrates the thematic features of each group. These thematic features are *broad categories* or *clusters* that are the focus of this study and can be considered predictors of Collins's thematic structure. Using the semantics of these lexical features, it becomes easy to assign labels to the groups, such as love, detective, romance, etc.

Table 3. *Lexical-sematic Clusters of Groups Identified for Thematic Cluster Analysis*

Lexicon Group 1 Emotions and pathos Collins 12, Collins 08	*Anguish, sorrow, pity, love, despair, passion, excitement anger contempt apathy pretty fury.*
Lexicon Group 2 Family relationships Collins 1, Collins 23, Collins 27, Collins 02, Collins 05, Collins 14 Collins 04, Collins 09, Collins 17, Collins 13, Collins 03, Collins 15, Collins 19, Collins 06, Collins 22, Collins 11	*Daughter, family, son, juvenile wife, mother, widow, senile, grandson, niece, paternity, young, acquaintance, maternal, brother, parent, sibling, husband, wife, nephew, lineage.*
Lexicon Group 3 Detective Collins 10, Collins 25, Collins 29, Collins 18	*Mystery, secret, anonymous, unknown, covert, clandestine detection.*

ISSN: 2229-9327

Lexicon Group 4 Human character Collins 7, Collins 30,	*Asylum, eccentricity, infidel, solitary, jealousy, hypocrite, betray, revenge,*
Collins 16, Collins 21, Collins 20, Collins 24, Collins 26, Collins 28	*jagged waves, rays, dusky, withered boughs, scenery, wreck, streets, flashed, dreary.*

Table Three presents content words with lexical variety from across Collins' prose writings. For instance, words like *infidel*, *asylum* or *eccentricity* appear very frequently in the context of human character; epithets like *affectionate*, *anxiety* reflects the author's emotional patterns; *narrow banks barren of verdure* shows his versatile imagery, much emphasized repeatedly in words such as *jagged*, *waves*, *rays* and *dusky.* These examples reveal a great deal of linguistic variation in Collins' fictional prose works that cannot be attributed to a limited number of variables of time using linguistic measures. These findings reveal that Collins's prose fictional texts use linguistic features that can be clustered in lexical-semantic fields frequently appearing in his texts.

Table Two and Table Three along with their descriptions are thus sufficient to determine thematic similarity among the words occurring in the same text. This is described as 'topical similarity' due to its linkages with the context (Clark, 2015), it is often a challenge to determine which words appear in the same text owing to its length. The solution offered is to reduce the context to a single sentence or only to a few words so that similar words do not appear in the same contextual window (Clark, 2015). In the context of the sampled texts for this study, this suggestion would not work since almost synonymous words, such as *dark*, *stagnant* or *barren*, occur in the same sentence and in the same context.

Sahlgren (2006) makes a distinction between two alternatives: one uses "syntagmatic" relations to the context while the other uses "paradigmatic" relations. The syntagmatic words are those that co-occur in the same text region, whereas paradigmatically related words are surrounding words that are often not the same (Sahlgren, 2006). Theoretically, it is recommended that each word is considered as a single context and the number of times such a context word occurs in a text is counted. The matrix may not require a target word used to calculate the context vectors in this study, but such words provide the context. For instance, Collins frequently uses the terms *secret* and *mystery* together, which are sometimes similar in meaning. Hence, there are many words surrounding other groups of words, including human emotions, imagery, and family relationships, sampled for this study. The clustering structure shown in Figure Four*Figure* exemplifies that the row corresponding to detective novels (*mystery*, *detection*, *secret*) in the term-term matrix displays a frequent numerical overlap with its corresponding vector. In other words, words surrounding the context of *secret* are the same as those surrounding *mystery.* A similar pattern is seen for *passion* and *love*; or *pathos* and *pity*. Such a pairing of target words also hints at the syntagmatic and paradigmatic relations of terms to their context.

Based on the hierarchical clustering shown in Figure Four of Collins' 30 novels, a lexical-semantic analysis can be suggested to identify the dominant themes. The thematic categorization

of Willkie Collins' novels on the basis of the semantic content using vector space clustering methods was, however, a big challenge. Firstly, the clustering required semantic intuition and splitting of the cluster into similar or contrasting words. In order to distinguish them thematically, a centroid vector for each of the four groups was constructed by quantifying the means of the vectors constituting them. Having calculated the mean, the differentiating variables between all groups were investigated in order to suggest a thematic categorization for each group. The codes or themes were emphasized in making the lexical and sematic selection. This process is similar to that used by structuralist researchers such as Moretti (2011), Bakhtin (1981) and Propp (1968) who reduced the clustering to variables like plot, characters, and setting. For example, Moretti (2011) carried out a structural analysis of William Shakespeare's *Hamlet* where plot and characters were seen as nodes and these nodes were isolated and then reconnected together again in order to show how the plot changes with the structural transformation of the characters. Elson, Dames, and McKeown (2010) justified the approach of structural analysis by making use of plot and characters to build clusters. Jayannavar, Agarwal, Ju, and Rambow (2015) revisited the hypotheses of Elson et al. (2010) and validated them by recommending dialogic interactions and semantic orientation to formulate clusters. Though these approaches succeeded in producing static networks or clusters for a piece of literature, they failed to recommend a thematic analysis or to build a similar static network to examine the themes of a novel or prose works, as is presented in this study.

Secondly, the process of clustering revealed that some words that were close to each other semantically created a challenge in how to suggest a dominant theme that can include all the texts within the same group and distinguish them from texts in other groups. This observation suggests that collocation analysis may be used for building thematic clustering. The commonality between words and their semantic values can also be seen as intrinsic properties. This suggestion is confirmed by concept mapping for each thematic category, presenting a network analysis of the co-occurrence patterns for each of the four groups. Each cluster represents one group consisting of words used as nodes. The concept mapping of similar words joined as nodes hints at their uniqueness in a particular corpus. This further illustrates the lexical-semantic qualities of Collins' texts. For the purpose of quantifying different categories of words in Collins' texts, it was necessary to make a word-by-word textual analysis.

As indicated in the clustering structure, it seems that the majority of Collins' work falls into two categories. The first category includes texts, such as *The Woman in White*, of sensation and detective fiction. The results are, therefore, broadly in agreement with the existing, philologically-based critical opinion on the thematic structure of Collins' work. The contribution of this study, however, is that it gives that critical opinion a scientific, that is an objective and replicable, basis. The methodology used in this study has been shown to be effective in the literary analysis of Collins' work and is thus potentially applicable to literary scholarship more generally. Computational analysis methods have been used here to empirically derive taxonomies of thematic concepts in the novels of Wilkie Collins. The implication is that the computational element in literary criticism provides what Hockey (2000) describes as "concrete evidence to support or refute hypotheses or interpretations which have in the past been based on human reading and the somewhat serendipitous noting of interesting features" (p. 66).

The essence of the proposed methodology is the use of the lexical-semantic content of texts in terms of frequency to categorize them. This method is mathematically-based, clearly understood, objective, and replicable. The human interpretation of texts is based on both lexical content and on higher levels. Still, it is, as pointed out previously, non-objective, non-replicable, and highly subjective. This study aimed to see whether applying computational-based methods to literary texts may constrain a subjective human interpretation by injecting elements of objectivity and replicability. The findings in most cases support the non-computational interpretation of the selected texts. However, in certain cases, the findings of the computational methods are at variance with the human interpretation and classification of texts.

For example, the two texts *Jude* and *Tess* were grouped together, although they are traditionally perceived as thematically different in that they depict two different realms. The two texts are computationally clustered together since they share the most distinctive lexical variables. In such a case, the question is: which is correct? The answer is: neither. Computational methods provide an objective clustering that gives us insight into an alternative interpretation based on criteria that can definitely be found in the texts and which constrain our subjective interpretations. This is the point of the study: not to claim that this method is better than or replaces all human interpretations of literary texts, but rather that it constrains subjective human interpretations by presenting classification criteria that are objective and replicable.

In spite of its success in grouping semantically-related texts together, text clustering based on word-level representation suffers from some limitations. Features like metaphor, irony, and humor still represent real challenges to text clustering applications. The reason for this is that there is nothing intrinsic to individual words to characterize or suggest any form of metaphor, irony, or sarcasm. At this final stage of this research, a number of limitations need to be considered.

First, the analysis was limited to word-level representations. These are lexical frequencies within documents. For analysis, the study depended on what is called the ‘bag of words’ method. Vector representation did not consider the ordering of words in a document. While many researchers accept that word-level representation is very effective in IR and text categorization applications, it is thought, however, that through the combination of word-level and phrase-level representations, a higher level of representation can be achieved leading to better and more accurate results. Furthermore, this can reduce the number of inaccuracies resulting from polysemy. As a result, it is recommended that data representation is based on both word and phrase levels for future automated thematic classification applications of literary texts.

Second, whereas this study has suggested some unifying themes that could well accommodate the texts of each group, it was not concerned with text summarization. It is recommended that computer-based studies of Wilkie Collins are focused on developing methods for text summarization. All the work done concerning providing summaries of Collins’ novels relies heavily on a critic’s own understanding and impressions of the text. In the face of this, it is becoming imperative to make use of effective algorithms available for extracting information and understanding discourse structure so that reliable summaries of Collins, and literary production in general, are generated.

Third, this study is concerned with identifying the thematic relationships within texts. This opens up some discussions on the ideas of the historical, tragic, etc. elements in the works of Collins. Nevertheless, the analysis was not concerned with providing a full charting of the genre classification of the prose fiction works of Collins. It is recommended that we create a genre classification of Collins' novels by means of exploratory multivariate analysis. With this, we could identify tragedies, comedies, and histories, etc. within the literary corpus of Collins.

7. Conclusion

Based on the preceding vector space analysis of Wilkie Collins texts, conclusions can be drawn by looking at the clusters and their lexical-semantic features. It is not a complete analysis as the contents of each cluster and the role of each feature may lead to further thematic classifications. A logical next step to be taken is to analyze all clusters and their features, both content-based and cluster-based, to see whether any new results can be formulated. A similar approach can be taken of another suitable corpus by including Collins' short stories as well or by choosing a different author. This study also notes the challenge of classifying a literary canon that broke with literary traditions making a quantitative literary analysis difficult. This study has attempted to build clusters from Collins' novels to represent their thematic categories. The main objective of this study was to understand the approach that a cluster analysis should rightly take in determining themes and key subjects in this kind of writing. A secondary objective was to understand how this clustering presented an indication of the author's style. Vector analysis was done for 30 texts of Collins' prose to understand the issues of thematic classification. The results of this study can serve as a basis for future studies and criticisms of Wilkie Collins' fiction. The study has gone some way towards resolving issues of authorship attribution and genre classification. Computational lexical-semantic methods can be used to identifying authors and characterizing texts by genre. As such, they are recommended as effective methods for authorship problems and genre classification applications.

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.

About the Author:
Abdulfattah Omar is an Associate Professor of linguistics at Prince Sattam Bin Abdulaziz University. He finished his PhD in linguistics at Newcastle University in 2010. His research interests include computational linguistics, digital humanities, and literary computing.
ORCID: https://orcid.org/0000-0002-3618-1750

References

Amati, G., & van Rijsbergen, C. J. (2002). Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems, 20*(4), 357-389.

Argamon, S., & Olsen, M. (2006). Toward meaningful compting *Commun ACM, 49*(4), 33-35.

Bakhtin, M. (1981). Discourse in the Novel *The Dialogic Imagination: Four Essays* (pp. 259-422). Austin: University of Texas Press.

Berry, D. M. (ed.) (2012). *Understanding Digital Humanities*. Basingstoke: Palgrave Macmillan.
Clark, S. (2015). Vector Space Models of Lexical Meaning. In S. Lappin & C. Fox (Eds.), *The Handbook of Contemporary Semantic Theory* (pp. 493-522): Wiley-Blackwell.
Corns, T. N. (1991). Computers in the Humanities: Methods and Applications in the Study of English Literature. *Literary and Linguistic Computing, 2* (2), 127-130. DOI:10.1093/llc/2.2.127
Elson, D. K., Dames, N., & McKeown, K. R. (2010). *Extracting Social Networks from Literary Fiction.* Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden.
Finneran, R. J. (ed). (1996). *The Literary Text in the Digital Age*: Ann Arbor: University of Michigan Press.
Gasson, A., & Peters, C. (1998). *Wilkie Collins: An Illustrated Guide*. Oxford: Oxford University Press.
Härdle, W., & Simar, L. (2003). *Applied multivariate statistical analysis*. Berlin; New York: Springer.
Hockey, S. M. (2000). *Electronic Texts in the Humanities: Principles and Practice*. Oxford: Oxford University Press.
Horton, T., Taylor, C., Yu, B. & Xiang, X. (2006). '*Quite Right, Dear and Interesting': Seeking the Sentimental in Nineteenth Century American Fiction*. Paris-Sorbonne: Digital Humanities
Jackson, J. E. (1991). *A User's Guide to Principal Components*. New York: John Wiley & Sons.
Jacobs, H. A. (1861). *Incidents in the Life of a Slave Girl*. Boston: Jacobs.
Jayannavar, P. A., Agarwal, A., Ju, M., & Rambow, O. (2015). Validating Literary Theories Using Automatic Social Network Extraction. *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, 32–41.
Jockers, M. L. (2009) 'Machine-Classifying Novels and Plays by Genre,' Matthew L. Jockers blog, 13 February. Retrieved from: http://www.stanford.edu/~mjockers/cgi-bin/drupal/node/27
Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Berlin ; Springer Verlag.
Mangham, A. (2008). *Wilkie Collins: Interdisciplinary Essays*: Cambridge Scholars Pub.
Moisl, H. (2009). Using Electronic Corpora in Historical Dialectology Research: The Problem of Document Length Variation. In M. Dossena & R. Lass (Eds.), *Studies in English and European Historical Dialectology* (98, pp. 67-90). Bern: Peter Lang
Moretti, F. (2011). Network Theory, Plot Analysis. *New Left Review, 68*(March-April), 80-102.
Novoviˇcov´a, J., Mal´ık, A., & Pudil, P.(2004). Feature Selection Using Improved Mutual Information for Text Classification. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (pp. 1010-1017). Springer.
Page, N. (2002). *Wilkie Collins: The Critical Heritage*: London: Routledge.
Plaisant, C., Rose, J., Yu, B., Auvil, L., Kirschenbaum, M. G., Smith, M. N., Clement, T., & Lord, G. (2006). Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces**.** Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries**,** 141–150. ACM.

Potter, R. G. (1989). *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*. Philadelphia: University of Pennsylvania Press, Incorporated.
P Propp, V. (1968). *Morphology of the Folktale*. Austin: University of Texas Press.
Putejovsky, J. (2012). The Semantics of Functional Spaces. In A. C. Schalley (ed.), *Practical Theories and Empirical Practice: A Linguistic Perspective* (pp. 307-324). John Benjamins Publishing
Pykett, L. (2005). *Wilkie Collins*. Oxford: Oxford University Press.
Ramsay, S. (2003). Special Section: Reconceiving Text Analysis: Toward an Algorithmic Criticism. *Lit Linguist Computing, 18*(2), 167-174. DOI:10.1093/llc/18.2.167
Ramsay, S. (2005). In Praise of Pattern. *TEXT Technology: the Journal of Computer Text Processing, 14*(2), 177-190.
Ramsay, S. (2007). Algorithmic Criticism. In R. G. Siemens & S. Schreibman (Eds.), *A companion to digital literary studies*. Malden, MA: Blackwell Publishers. Oxford: Blackwall
Rencher, A. C. (2002). *Methods of Multivariate Analysis* (2nd ed.): John Wiley & Sons, INC.
Rettberg, S. (2016). Electronic Literature as Digital Humanities. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A New Companion to Digital Humanities* (1st ed., pp. 127-137). West Sussex (UK): Wiley Blackwell.
Rijsbergen, C. J. V. (1979). *Information Retrieval* (2nd ed.). London: Butterworth.
Robertson, S. E., & Walker, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 345–354. Springer-Verlag, Retrieved from http://portal.acm.org/citation.cfm?id=188490.188561#
Rockwell, G. (2003). What is Text Analysis, Really? *Literary and Linguistic Computing, 18*(2), 209-219. DOI:10.1093/llc/18.2.209
Rommel, T. (2004). Literary Studies. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 88-97). Oxford: Blackwell.
Rowson, S. (1794). *Charlotte. A Tale of Truth*. Philadelphia: Printed by D. Humphreys, for M. Carey.
Rowson, S. (1828). *Charlotte's Daughter; or, The Three Orphans. A Sequel to Charlotte Temple*. Boston: Richardson & Lord.
Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* (Unpublished Doctoral Thesis). Stockholm University, Stockholm, Sweden
Saint-Dizier, P., Viegas, E., issues, L. s., Bird, S., Boguraev, B., & HIndle, D. (1995). *Computational Lexical Semantics*: Cambridge University Press.
Salton, G., & Buckley, C. (1987). *Term Weighting Approaches in Automatic Text Retrieval*. Retrieved from
Siemens, R., & Schreibman, S. (2013). *A Companion to Digital Literary Studies*. West Sussex: Blackwell Wiley.
Singhal, A., Chris, B., & Mandar, M. (1996). Pivoted Document Length Normalization. *Proceedings of the 19th annual international ACM SIGIR conference on Research and*

development in information retrieval, 21-29. DOI:http://doi.acm.org/10.1145/243199.243206
Singhal, A., Salton, G., Mitra, M., & Buckley, C. (1996). Document length normalization. *Information Processing & Management, 32*(5), 619-633.
Storjohann, P. (ed). (2010). *Lexical-Semantic Relations: Theoretical and practical perspectives*: John Benjamins Publishing Company.
Stowe, H. B. (1859). *The Minister's Wooing*. New York: Derby and Jackson.
Stowe, H. B. (1897). *Uncle Tom's Cabin*. New York: T. Y. Crowell & company.
Taylor, J. B. (2006). *The Cambridge Companion to Wilkie Collins*. Cambridge: Cambridge University Press.
Yu, B. (2008). An Evaluation of Text Classification Methods for Literary Study. *Literary and Linguistic Computing, 23*(3), 327-343.