Belo Horizonte, v. 12, n. 2, p. 83-101, mai.-ago. 2020 – ISSN 1983-3652
DOI: 10.35699/1983-3652.2020.24396

CLASSIFYING LITERARY GENRES: A METHODOLOGICAL SYNERGY OF COMPUTATIONAL MODELLING AND LEXICAL SEMANTICS

CLASSIFICAÇÃO DE GÊNEROS LITERÁRIOS: UMA SINERGIA METODOLÓGICA DE MODELAGEM COMPUTACIONAL E SEMÂNTICA LEXICAL

Abdulfattah Omar
Prince Sattam Bin Abdulaziz University, Arábia Saudita / Port Said University, Egypt (Egito)
a.abdelfattah@psau.edu.sa

ABSTRACT: Classifying literary genres has always been methodologically confined to philological methods and what is commonly known as Vector Space Clustering (VSC). The problem has been exasperated with the widening gap between computational theory and traditional analysis of literary texts. Towards finding a solution to this problem, the current study utilizes a synergetic approach that brings together two established methods. First, a computational model of genre classification is drawn upon for identifying concept-based, rather than word-bound, topics, where the representation of texts is secured via the 'bag of concepts' (BOC) model as well as the sense-restricted knowledge and meaningful links holding between and among concepts; relatedly, the two model strands of explicit semantic analysis (ESA) and ConceptNet have enacted text classification. Second, a contextual lexical semantic approach (CRUSE, 1986, 2000) is employed so that the contextual variability of word meanings and concepts can be tackled within the confines of the target literary genres classified. The findings of present study have shown that the current composite approach of computational and semantic models has resulted in improved performance in classifying literary genres, especially with respect to delineating the links between each cluster's document-members and generalizing about their unifying genre. Further implications have emerged from the present study, namely, the benefits reserved for digital libraries and the process of archiving, where literary-text classification has proven problematic to both users and readers in many cases.
KEYWORDS: Bag of concepts (BOC). ConceptNet. Explicit Semantic Analysis (ESA). Genre classification. Topic concepts. Vector Space clustering (VSC).

RESUMO: A classificação de gêneros literários sempre se restringiu metodologicamente aos métodos filológicos e ao que é comumente conhecido como *Vector Space Clustering* (VSC). O problema foi exasperado com a crescente lacuna entre a teoria computacional e a análise tradicional de textos literários. Para encontrar uma solução para esse problema, o presente estudo utiliza uma abordagem sinérgica que reúne dois métodos estabelecidos. Primeiro, um modelo computacional de classificação de gênero é utilizado para identificar tópicos baseados em conceito, em vez de vinculados a palavras, em que a representação de textos é protegida por meio do modelo "bolsa de conceitos" (BOC), bem como o conhecimento restrito aos sentidos e os vínculos significativos entre os conceitos; de maneira semelhante, os dois modelos de análise semântica explícita (ASE) e ConceptNet promulgaram a classificação do texto. Segundo, uma abordagem semântica lexical contextual (CRUSE, 1986, 2000) é empregada para que a variabilidade contextual dos significados e conceitos das palavras possa ser abordada dentro dos limites dos gêneros literários alvo classificados. As descobertas do presente estudo mostraram que a atual

abordagem composta de modelos computacionais e semânticos resultou em melhor desempenho na classificação de gêneros literários, especialmente no que diz respeito a delinear os vínculos entre os membros do documento de cada grupo e generalizar sobre seu gênero unificador. Outras implicações emergiram do presente estudo, a saber, os benefícios reservados para as bibliotecas digitais e o processo de arquivamento, em que a classificação de textos literários se mostrou problemática para usuários e leitores em muitos casos.
PALAVRAS-CHAVE: Bolsa de conceitos (COB). ConceptNet. Análise Semântica Explícita (ASE). Classificação de gênero. Conceitos de tópicos. Vector Space Clustering (VSC).

1 Introduction

Different approaches have been proposed for the classification of genre in literary texts. These have been based primarily on textual content and/or biographical considerations using what is referred to as the 'philological method', that is, through an individual researcher's reading of printed materials and their intuitive abstraction of generalisations from that reading. With the advent of electronic texts, however, a large number of literary works have become available to analysis – their electronic format has permitted the application of computational data analysis concepts and procedures. In bridging the gap between literary studies and computational theory, various classification systems, such as vector space clustering and naïve Bayes, have been used in processing literary texts for genre classification, authorship attribution, and purposes of thematic analysis. However, questions of reliability and meaningfulness may be raised in relation to such applications. Although the peculiar nature of literary data is a significant factor, other factors concerning the way classification is carried out need also to be considered. Standard (also referred to as classical) classification systems have two major problems. Firstly, they do not consider the semantic relationships between words and, therefore, the meanings documents have may not be accurately represented. Secondly, there is a gap between extracting the most distinctive features within a given corpus and assigning appropriate descriptions to the clusters generated through this process of extraction. In addressing these challenges, this study proposes an approach based on replacing 'bag of words' (BOW) representation with 'bag of concepts' (BOC) where the most distinctive concepts are extracted. Texts were classified using a hybrid model of Explicit Semantic Analysis (ESA) and ConceptNet.

By way of illustrating this innovative method, this study is based on the automatic classification of 346 American novels written by 86 novelists. The American novel is a rich field of literature and includes numerous thematic genres, such as: adventure novels; children's novels; crime and detective novels; fantasy novels; gothic novels; immigrant/ethnic novels; mystery novels; picaresque novels; political novels; science fiction novels; and war novels. In any given historical period, literary works have traditionally been classified according to the norms of that period and/or the background of the author. As a result, many texts have been classified differently due to a lack of objective and replicable standards. Although standard automatic classification systems have addressed the problem of subjectivity in the genre classification of different literary texts (BIBER, 1986; DOUGLAS, 1992; HOLMES, 1998; JOCKERS, 2009; KESSLER; NUMBERG; SCHTZE, 1997; KOPPEL; ARGAMON; SHIMONI, 2002; RAMSAY, 2005, 2007; RAMSAY; STEGER, 2006; WOLTERS; KIRSTEN, 1999; XIAO; MCENERY, 2005), problems related to the meaning-

fulness, usefulness, and reliability of these classifications remain unresolved. In relation to the classification of literature, standard classification methods are not effective in representing document meanings and assigning appropriate descriptions for the clusters generated through classification procedures, and this has negative impacts on the reliability of classification results. Given the shortness of both philological and standard automatic clustering methods, this article is based on the hypothesis that reconciling literary analysis and computational theory makes it possible to overcome many of the inherent problems within literary studies related to genre classification and analysis. In this context, this study is concerned with exploring alternative methods to traditional classification methods and systems and suggests a more meaningful and reliable genre classification of literature.

2 Literature review

The genre classification of literary texts—the process of classifying texts according to what they have in common, either in their formal structure or in their treatment of subject matter, or both—has always been a controversial issue in literary debates (BHATIA, 2014; WILDER, 2012). Throughout its long history as a discipline, it has often been argued that genre classification is significant to critical literature studies in a number of ways. Firstly, identifying the genre of a text gives us an opportunity to gain a better understanding of its intended overall subject. Secondly, grouping similar texts together can deepen our sense of the value of any single text by allowing us to view it comparatively, alongside the many other texts of its type (FOWLER, 1982; SARICKS, 2009). Thirdly, with so many texts available today, sometimes the information needed is not something specific. Rather, we seek answers to general questions, which are typically answered by techniques that look at an entire document, or set of documents, through clustering, categorization, or genre classification.

Although the idea of genre classification is as old as Aristotle' *Poetics*, so far there has been no systematic classification approach to literary works (UTAS, 2006). There are no fixed or universal formulas that can be adopted in an objective and replicable manner in defining literary genres or classifying literary texts. With this lack of standards, genre classification has traditionally been based partly on textual content and partly on biographical considerations. These have been generated by what may be defined as the 'philological method', that is, by an individual researcher's reading of printed materials and his/her intuitive abstraction of generalisations from that reading. As a result, the categories and genres suggested are always evolving and tend to be inconsistent. In other words, the idea that literary texts are usually classified on the basis of *external* criteria leads to unreliable genre classifications.

The inconsistencies associated with genre classifications based on philological methods have opened up pathways for the development of more objective methods. For example, the 2015 edition of The Best American Poetry included a poem by the Chinese poet Yi-Fen Chou. It was, however, later revealed that Chou was a white male American poet named Michael Derrick Hudson who wrote under a Chinese pseudonym in order to get his poems published. According to Hudson, his poems were rejected over 40 times purely on biographical considerations. The editor admitted that his classification of the poem under the Chinese American category was solely based on the author's biographical information (CARISSIMO, 2015; FLOOD, 2015). Apart from ethical considerations of the issue,

this problem brought to the fore one of the problems with the classification of ethnic American literature. Ethnic American literature refers to a class of literature where speakers are conscious of being members of a people sharing a common and distinctive racial, national, religious, or cultural heritage (FRANCO, 2006; GRICE, 2001). It represents the body of literature that was written in the United States by writers of African, Arab, Indian, and Chinese descent. Works of this kind are usually concerned with concepts such as national consciousness, time, space, and belonging (BAYM, 2007; FRANCO, 2006; GOMAA, 2016; GRICE, 2001). Nevertheless, the classifications of such works is usually based on the ethnic background and biographical information of the authors, even if they do not actually have these literary and artistic features (GOMAA, 2016; NELSON, 2015). The implication here is that classifications based on subjective criteria are not reliable (DUNN; ARGAMON, RASOOLI; KUMAR, 2016).

The idea of subjectivity has always been a central issue in discussing literary works. The Russian Formal School, which developed at the beginning of the 20th century, is an obvious example. It was developed with the purpose of framing and regulating guidelines for the objective analysis of poetry and literature (ERLICH, 1981; KARCZ, 2002; MANDELKER, 1983; STEINER, 1995; TOBIN, 1988). For the first time, literary analysis came to be considered a science. The formalists developed what they considered a scientific method for studying poetic language (ERLICH, 1981). In the face of severe criticism, the school initiated debates about the importance of adopting objective grounds when evaluating poetry and literature. Today, there is a tendency to reconstruct literary studies based on scientific grounds where objectivity and empirical determination are given the highest priority. Recent years have witnessed the development of what can be considered, in English, literary science. This is a translation of a German term ‘Literaturwissenschaft’, which refers to all disciplines of literature that adopt a scientific method (BAASNER; ZENS, 2005).

Over the last two decades, the development of computational approaches has encouraged scholars to think about novel methods that address many of the inherent problems within literary studies. Works of this kind are classified under the broad headings of ‘digital humanities’ and/or ‘literary computing’ studies – researchers use computational methods either to answer existing research questions or to challenge existing theoretical paradigms. In doing so they generate new questions and pioneer new approaches (BERRY, 2012). Classification systems and, particularly, vector space clustering (VSC) methods and naïve Bayes remain among the most widely used computational approaches in literary studies in general and in the genre classification of literature in particular. VSC is an approach whereby similar texts are first clustered or grouped together based on the extraction of keywords from them; subsequently, labels or genres are suggested for each group (CHAKRABORTY; PAGOLU, 2014; CHAKRABORTY; PAGOLU; GARLA, 2014; MANNING; RAGHAVAN; SCHÜTZE, 2008; RIESEN; BUNKE, 2010).

VSC ignores word order and the context in which words are used. Each document is represented by the number of occurrences of each word in the document in Euclidean vector space where each token in the vector corresponds to a unique/given word in the matrix (JOACHIMS, 2002; OZGUR, 2006). Naïve Bayes classification, on the other hand, is a supervised learning approach that is based on the assumption of independence. It estimates the probabilities of each attribute based on its frequencies over the course of processing the training data. The Naïve Bayes classifier assumes that the presence/absence of a particular feature of a class is unrelated to the presence/absence of any other feature. In this way, when the features of a given attribute depend on each other, or upon the exis-

tence of other features, the classifier considers these attributes to belong to a given class. Both VSC and naïve Bayes are based on what is known as the bag of words (BOW) representation model and it is considered today to be a classical classification system. In BOW methods, documents are represented in the form of term vectors, where a term is a morphological normal form of the corresponding word (the words *love*, *loves*, *loved*, *lover*, and *lovers*, for instance, are represented as just one normal form, namely *love*). The assumption is that meaning is carried by the vocabulary with no regard to syntax, semantic relationships among words, or the context in which these words are used.

In spite of the effectiveness of these systems in different applications and disciplines, the peculiar nature of literary texts needs to be taken into account and standard classification systems based on word similarity may not be feasible when dealing with literary data. Traditional VSC methods, those based on keyword indexing, have serious shortcomings that need to be considered when used in literary studies. In such methods, classification is based on word similarity so that users can form impressions of what texts are about thematically. In genre classification studies, however, critics are rarely interested in lexical databases. Rather, they need concepts that are often more difficult to abstract, represent, and process. In spite of the reasonable success of recent classification systems, their effectiveness in dealing with literary texts is still very limited (OLMOS; LEÓN; JORGE-BOTANA; ESCUDERO, 2013). As thus, literary studies, including genre classification, authorship attribution, and thematic analysis, have tended to use the BOW model, which has often yielded controversial and even confusing results.

Although the BOW model has proved to be effective when dealing with a number of different classification tasks, including classifying news and scientific articles (ADOLPHS; KNIGHT, 2020), it is suggested that it is not appropriate for applications in literary studies. The general assumption in using ATC is that extraction of the most distinctive or key features within a given corpus leads to a successful grouping together of similar texts. This can be illustrated by the following example. A corpus of news reports may be processed by ATC and be classified into two main clusters or groups. The most distinctive features of Group 1 include the terms: *Obama*, *Pentagon*, *Congress*, *policy*, *politics*, *issue*; *administration*, *peace*, *terrorism*, *China*, *Arab*, *Middle*, *East*, *Israel*, *United*, *Nations*, *Security*, *Council*; while those of Group 2 are *Messi*, *Ronaldo*, *World*, *Cup*, *FIFA*, *champions*, *league*, *premier*, and *football*. It is easy to assign an appropriate and meaningful description to each cluster or group. It may be suggested that texts in Group 1 can meaningfully be classified together under the broad heading “World Politics”, while those in Group 2 can classified under the heading “Football”. The same argument extends to other document types, including scientific articles and abstracts, as well as business and technical documents.

In literature, however, the most distinctive terms or words are not sufficient by themselves to generate successful and reliable classifications, as the lexical relations of these terms are not considered. Ramsay (2005) used a VSC based on the BOW model for objective genre classification of Shakespeare’s plays. The plays were grouped into 4 distinct clusters. These were: 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 could not be wholly convincing. However, he also stressed the importance of thinking about objective criteria in genre classification. Similarly, Jockers (2009) attempted a genre classification using the Naïve Bayes method and BOW model to generate a classification of 37 plays of Shakespeare. The plays were clas-

sified into three distinct groups, which Jockers labelled as comedies, histories, and tragedies. Again, there are some issues with Jockers' analysis. For some texts, there was no justification for classifying some texts under one category or group as opposed to another. Although some classifications of literary texts seem reasonable and are supported by internal and external evidence, the distinctive lexical features of these literary texts are not sufficient for generalizing information about the generated groups. In his classification, for instance, of the prose fiction of Thomas Hardy, Omar (2010; 2015) classified the novels and short stories of Hardy into four distinct groups. The most distinctive lexical features of Group 2 are the words: *husband*, *captain*, *squire*, *France*, *Casterbridge*, *sergeant*, *curate*, *dance*, *countess*, *knight*, *duke*, *shepherd*, *architect*, and *horse*. These words cannot by themselves express or indicate the themes or genre they represent. In this way, literary texts need to be dealt with differently and more reliable methods need to be developed for the classification of literary texts.

3 Statement of the problem

Recent years have witnessed great developments in the quality and effectiveness of automatic text classification (ATC). ATC is now widely used in different applications, including information retrieval (IR) and text mining. In the field of IR, for instance, search engines, such as Google and Yahoo, have made considerable achievements in developing classification methods and approaches with the purpose of successfully grouping similar texts together. Such developments have had positive implications on both the IR performance of these search engines and the reliability of ATC in general (LIAO; CHU; HSIAO, 2012; WEI; LUC; CHANGB, 2015). The success of automatic classification methods has encouraged researchers in many disciplines including medicine, engineering, and linguistics, to adopt such classification systems in dealing with their data. In literature, unfortunately, applications are still lacking. This may be down to the idea that literary analysis and computational theory are poles apart. Hammond and Brooke (2013) assume that there are cultural barriers that make it difficult for literary scholars and researchers to make use of computational theory. Kessler et al. (1997) assert that genre classification has long been ignored in computational linguistics. Furthermore, the ineffectiveness of computational methods in dealing with literary texts has raised a number of doubts concerning their usefulness. This can be attributed in part to the problems with the BOW model and standard classification methods that have had negative impacts on the reliability and performance of these methods. These problems include misrepresentation of the meaning of documents as well as a failure to assign meaningful description to clusters generated from the classification procedures used. In light of the limitations of ATC systems in relation to literary studies, this study proposes the integration of semantic relations between words using bag of concepts (BOC) representation, where the text acts as a vector in the space of concepts. The rationale here is that these concepts are semantic sources and rich veins of knowledge – processing these can have positive impacts on the generation of meaningful categories that better describe collections of documents. The research questions are thus asked in response to the effectiveness of semantic mapping using both Explicit Semantic Analysis (ESA) and ConceptNet methods for more reliable genre classification of literature.

4 Methodology

4.1 Methods

In the proposed system, each document is represented as a set of concepts where word sets, as well as phrases, are mapped onto the concepts they represent. It is therefore referred to as the bag of concepts (BOC) model (BELLEGARDA, 2008; MAJKIĆ, 2014). Concept meanings are represented in terms of a set of probabilistic topics for building or generating probabilistic semantic maps. The function of the proposed system, then, is to identify the most important features within the data and predict the associations between these features to group documents with similar concepts together (GRIFFITHS; STEYVERS, 2007). In order to do this, a semantic representation hybrid model, including both Explicit Semantic Analysis (ESA) and ConceptNet, is used in order to capture adequate details of the semantics of the texts. The combined model is considered useful in supplementing text representation with the rich knowledge and information available in global online encyclopaedias. Linking concepts with web-based wikis has the advantage of supplementing text representation with related content and significant amounts of world language.

ESA is a clustering approach developed by Gabrilovich and Markovitch (2006) based on computing semantic similarity and relatedness within texts by means of representing the meaning of texts in a high dimensional space of concepts derived from digital knowledge bases or knowledge platforms. The main assumption behind ESA is that computing the degree of semantic relatedness between fragments of natural language in texts can be improved by explicitly representing the meaning of any text in terms of encyclopaedic or knowledge-based concepts (GABRILOVICH; MARKOVITCH, 2006, 2007). ESA is generally based on Wikipedia, which is currently the largest and most diverse online knowledge base and has the greatest amount of well-organized human knowledge. As such, concepts can easily be related to articles in a useful manner (GABRILOVICH; MARKOVITCH, 2006, 2007).

ESA is different to conventional clustering methods (such as vector space clustering) in the sense that it represents the meaning of a text, not just the meaning of a text's vocabulary items. The text is seen as a combination of concepts, rather than just a sequence of words. In this way, ESA reduces dependence on key words (the distinctive feature of the BOW model) and thus improves clustering performance. This study makes use of ConceptNet, rather than WordNet, as the former has more semantic relationships and a greater vocabulary available. Furthermore, due to the limitations of WordNet, many relationships and concepts are missing making it inappropriate for literary data—literary data requires more concepts and relations if one wishes to gain a general idea as to what a text is about.

Although ESA has proven efficacy in the semantic analysis of textual data—especially for short texts—experimental results indicate that ESA has some problems. Weiping et al. (2008) have argued that ESA is not effective in dealing with long documents. Shalaby and Zadrozny (2017) add that ESA has some problems, such as sparse vectors, which cause there to be low similarity between texts, and that it is restricted to explicit meanings derived from knowledge bases, such as Wikipedia, as the main source of concepts. The problem is that ESA disregards the idea that concepts have implicit associations that cannot be generated by relying on Wikipedia, or a different knowledge base, as the main or only source. To counteract this, ConceptNet is used to support the clustering performance

of ESA since it can effectively deal with both explicit and implicit meanings. ConceptNet is a process that has long been used to provide automated categorization of documents based on extracting concepts embedded in documents. It is a workflow that is used to discover implicit and explicit relationships and useful associations and groupings in a set of documents or collection of data with the purpose of detecting similarities between documents in a large corpora and classifying them by topic. It can thus provide powerful insights into the meaning, provenance, and similarity of documents (FANG; MEHLITZ; LI; SHENG, 2006; HAN; KAMBER, 2001; LOOKS et al., 2007). It is assumed that each word in a given document relates to several possible concepts and this makes it possible to cluster documents based on their content. In ConceptNet, texts are processed and general concepts are extracted over two stages. Firstly, documents are reduced to a sequence of words that describes their content. Secondly, these words are mapped onto concepts. For example, if we had a number of documents on generative grammar, ConceptNet would be used to identify relationships and generate facts based on the data within the collection and the dimensions of the subject. That is, a concept may carry different levels of importance in different sentences and/or documents. For example, these could include things 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, for example wh-movement; linguistic competence, etc. In our case, ConceptNet was used together with ESA as an effective method for making sense of texts.

ConceptNet is a semantic resource for connecting common sense knowledge representations to one another (GELBUKH, 2007; LIU; SINGH, 2004). According to Speer and Havasi (2013), ConceptNet can be usefully used in different natural language processing tasks, including document classification, as it effectively captures a wide range of commonsense concepts and relations that are structured in a simple and easy to use semantic network. ConceptNet is used along with ESA, since both are concerned with computing and determining semantic similarity between texts, with a focus on concepts rather than words. ConceptNet is based on diverse relational ontologies that makes it, along with ESA, useful in making practical, context-oriented, common sense inferences from real-world texts. I suggest that the integration of ESA and ConceptNet will lead to better classification performance, which will be useful when classifying texts of the same genre and assigning appropriate categories to these texts (HAVASI; SPEER; ALONSO, 2007; SPEER; HAVASI, 2013).

However, when it comes to identifying semantic similarity between texts, computational tools need to be enhanced with some semantically analytic methods that qualitatively specify one significant aspect, that is, contextual variability of word meanings and concept meanings. This aspect can be dealt with in the current methodological framework should one draw upon D. A. Cruse's (1986; 2000) contextual lexical-semantic approach. The approach is sufficiently practical as it captures the semantic nature of both words and their respective senses within contextual constraints. Towards fulfilling this process, Cruse focuses on two dimensions of lexical-semantic analysis: (i) paradigmatic sense relations of inclusion and identity and (ii) paradigmatic relations of exclusion and opposition. Such a paradigmatic dimensions of inclusion and exclusion entail an answer to the following question: "what sort of entities do sense relations relate?" (CRUSE, 2000, p. 147). At this point, Cruse has methodologically highlighted the paradigmatic relations reflecting the semantic choices available at a particular structure point in a sentence; this has facilitated the se-

mantic process of identifying two broad classes of sense relations: first, those expressing identity and inclusion between word meanings (hyponymy, meronymy, and synonymy); and, second, those expressing opposition and exclusion (complementaries, antonymy, reversives, converses, and polarity). Indeed, for the sake of genre classification and their semantic networking, it is also significant to view Cruse's paradigmatically oriented sense relations of inclusion and exclusion within the conceptual boundaries of what Coulson (2006, p. 73-74) terms "frame-shifting in text processing." This is especially so if one considers each sense relation as being framed and potentially shifted across different text types; and this "underscores the roles of conceptual frames in structuring experience and making inferences that go beyond what's immediately present" (COULSON, 2006, p. 74). Simultaneously, the same aspect of frame-shifting may potentially create a context where the meaning of text and its truth can accord with the genre-based use of language (PREDELLI, 2005).

4.2 *Data representation*

To support reliable generalizations about the data, a corpus of 346 American novels written by 86 novelists from different historical periods was created. These were randomly selected from the lists of American novels and novelists generated through the genre discussions of the American novel in Bendixen's (2012) A Companion to the American Novel and Cassuto and Reiss (2011) *The Cambridge History of the American Novel*. Texts were then downloaded mainly from Project Gutenberg, Literature Online, and the Internet Archive, which are all trustworthy resources. A list of the selected texts is given in Appendix 1. For the semantic representation of the selected texts, a concept map, a graphical tool for visually representing the relationships between concepts and ideas, was built so as to abstract all the comprehensible concepts within the documents. In this method, each concept is depicted in the form of a node, has a semantic interpretation, and is associated with Wikipedia and ConceptNet, which is a large semantic network consisting of a large number of common sense concepts (HAVASI et al., 2007; LIU; SINGH, 2004). One major problem with the corpus, however, was the high dimensionality of the data, which had implications for the interpretability of the results generated and the reliability of the classification. Prodigious amounts of irrelevant information arising from high-dimensionality data make effective clustering difficult.

In ATC applications, the larger the data dimensionality, the more difficult it becomes to define the manifold sufficiently well to achieve reliable analytical results. As such, it is necessary to base classification applications or tasks only on the most important or distinctive features available because irrelevant and redundant information has a fundamental bearing on the reliability and accuracy of classification performance. In order to address this problem, distinctive concepts were selected by means of weighing vector concepts using TF-IDF (term frequency inverse document frequency), as developed by Spärck Jones (1972).

This is one of the most common weighting methods and it is widely used for identifying the most important variables within datasets (ROBERTSON, 2004). The underlying principle of this method is the adoption of a certain set of effective terms that collectively characterize the set of documents. In TF-IDF, the most discriminant terms are the highest TF-IDF variables. With document clustering, if the highest TF-IDF variables, which are taken to be the most discriminant terms, are identified, then unimportant variables can be

deleted, reducing the dimensionality of the data. In our case, TF-IDF was used to identify the most statistically significant concepts, rather than words, since this study uses concepts as variables.

Given that the highest TF-IDF variables are the most important, each column was calculated. Based on the TF-IDF measurement, only the highest 142 concepts were retained. This resulted in removing the noisy concepts, which were not significant for the classification task, and retaining only the most distinctive concepts. TF-IDF helps in suggesting only the distinctive features within a dataset. The expressions and phrases *It was raining*, *It was a queer summer*, *once upon a time*, and *In the town there were*…, for instance, are very frequently used in almost all the selected texts. In text clustering, however, these phrases are not useful because they are not distinctive. The advantage of TF-IDF is that it helps to identify only the most frequent and salient features (words or phrases) in each cluster or group, which are infrequent in other clusters or groups.

In order to support the lexical semantics of the extracted concepts, Encyclopedia Britannica-based ESA was first used. In spite of the popularity and very frequent use of Wikipedia in ESA applications, it was thought that Encyclopedia Britannica was more appropriate to the purposes of the study as it is a rich digital knowledge platform that provides relevant and authentic information. It is also constantly updated in order to maintain the value of the platform and its reputation for reliability. As a second step, ConceptNet was used in order to improve and support clustering performance.

5 Analysis

With no prior assumptions about the data, the 346 texts fell into 7 distinct groups. Group 1 included 86 novels; Group 2 included 36 novels; Group 3 included 51 novels; Group 4 included 48 novels; Group 5 included 41 novels; Group 6 included 63 novels; and finally Group 7 included 21 novels. Classification was generated based on the idea that each group had a number of distinctive features and that members of the same group were semantically and conceptually similar, making it possible to divide the texts into sections by genre. Texts in Group 1, for instance, are best described as immigrant or ethnic American novels where the most distinctive features include: *immigration*, *immigrants*, *immigrated to America*, *home*, *return home*, *a wish to return*, *yearning to return*, *tradition*, *family traditions*, *culture shock*, *Indian food*, *Indian music*, *Indian community*, *China*, *Chinese food*, *Chinese culture*, *Arabs*, *Arab world*, *Arabic*, *Islamic*, and *Muslim*. Among the texts included in Group 1 are: Abu-Jabr's *Crescent* (2003) and *The Language of Baklava* (2006); Bulosan's *America is in the Heart* (1946); Chua's *Battle Hymn of the Tiger Mother* (2011); Desai's *The Inheritance of Loss* (2006); Hua's *Deceit and Other Possibilities* (2016); and Tan's *The Joy Luck Club* (1989). The immigrant novels are generally concerned with topics related to homeland tales and diaspora experience (BOELHOWER, 1981; WALKOWITZ, 2010). Words like *return home*, *a wish to return*, *yearning to return* tell cross border tales as stark and dark stories of wrath, oppression, humiliation, and identity crisis as presented by an author in a very stylized approach. The text's coherence aided by words like *immigration*, *immigrants*, *immigrated to America*, *home*, *return home* in such narratives constitutes the genesis of a genre, that is, a new kind of genre born from depictions of the sufferings of immigrants. The immigrant is no more a historical protagonist, but an alien with multicultural visitations to a dystopian land

of nightmare where his suffering goes beyond *culture shock*.

Texts in Group 2 included: May's *Little Women* (1868); Gag's *Millions of Cats* (1928); Beverly Cleary's *Henry Huggins* (1950); *Mitch and Army* (1967) and *Strider* (1991); and Erin Hunter's *Into the Wild* (2003). The most distinctive words and phrases of this group included: *fairy tale*, *marvelous tale*, *Barbie for girls*, *Barbie dream house*, *stories about*, *seemed impossible*, *how impossible it was*, *castle hall*, *castle in the air*, *struggling against wizards*, *magic box*, *magic country*, *magic kingdom*, *darkness*, *burning eyes*, *river was burning*, *hopes and fears*, *pink mountain*, *sudden and mysterious disappearance*, *bed in heaven*, *adventure*, and *struggled against the strong arm*. Based on this categorization and the salient features of this group, it may be suggested then that these novels can be described as children's novels. Mickenberg and Vallone (2012) argue that the children's novel has kept the tradition of tales of simple fantasy, folklore, and folk tales alive, as can be seen in the salient word categories like *fairy tale*, *marvelous tale*, *Barbie for girls*, *Barbie dream house*; or descriptions of crime, ghost and horror stories with juvenile protagonists, or mystery and adventure as represented in words like *struggling against wizards*, *magic box*, *magic country*, *magic kingdom*, and *mysterious disappearance.* This is a vast genre covering many thrillers, mysteries, and adventures.

In Group 3, the most distinctive concepts included: *planet*, *gravities of Earth*, *robots*, *robot ship*, *against the stars*, *climbed into the stars*, *ship's brain*, *hydrogen bombs*, *atomic bombs*, *rocket system*, *planets in this system*, *unknown spacecraft*, *alien spacecraft*, *destroy another planet*, *alien creatures*, *army of creatures*, *mythological creatures*, *strange creatures*, *fighting machine*, *sensation*, *alien passengers*, *alien spaceship*, *transport of troops between planets*, *biologically destroyed*, *everyone on this planet destroyed*, *center of gravity*, *missiles*, *death of the planet*, *superhero*, and *virus*. This group included works such as: John Campbell's *Islands of Space* (1931); Murray Leinster's *First Contact* (1945); and William Gibson's *Neuromancer* (1984). It makes sense, therefore, to assign the genre of science fiction to this group of novels. The genre of science fiction tells tales about science and technology, worlds in space, and environments set in the future. Though these tales are fictitious, they often draw closely on scientific theories. Science fiction also closely resembles the genre of fantasy and words like *against the stars*, *unknown spacecraft*, *alien spacecraft, climbed into the stars*, take the reader to distant regions in space. At the same time, such descriptions like *biologically destroyed*, *center of gravity*, *virus*, *hydrogen bombs*, *atomic bombs* etc. speak of the misuse of science by man. Grim it may be, but such expressions make this genre powerful and distinct from the closely aligned genres of dystopian fiction or fantasy.

In Group 4, the most distinctive variables were: *The President*, *the U.S. President*, *Confederate Government*, *overthrow of the government*, *government officials pledge*, *presidential advisor*, *presidential campaign*, *presidential candidate*, *Presidential elections*, *American politics*, *the White House*, *Vice-President*, *government agencies*, *military advisor*, *military aides*, *military leaders*, *military intelligence*, *opposing the treaty*, *the Jews*, *the American Jews*, *majority of Americans*, *Israel*, *peace and war*, *battle for freedom*, *freedom of speech*, *preserve our democracy*, *American democracy*, *the U. S. Congress*, *the Congress passed*, *party leaders*, *the Republican and Democratic leaders*, *the U. S. Constitution*, *political propaganda*, *ideology*, *the free world*, and *the civilized world*. This group included works such as: Henry Brooks Adams' *Democracy: An American Novel* (1880); Jed Mercurio's *American Adulterer* (2009); Jeff Greenfield's *The People's Choice* (1995); Edward Klein's *The Obama Identity* (2010); Philip Roth's *Our Gang* (1971); and Robert

Penn Warren's *All the King's Men* (1946). Based on the most distinctive concepts of this group, the novels grouped together here can be described as political novels. Political fiction has been recognized as a distinct genre in American literature since the beginning of the 20th century (SPEARE, 1924). In spite of the fact that the political novel is a dominant genre in contemporary world literature, the definition of the genre remains vague and unclear. While the term may be considered complex and vague, the political novel can be seen as a political instrument that reflects national character (BLOTNER, 1966, 1977), and the political milieu is the dominant setting (HOWE, 1992). It can also be a tool of *political propaganda* for a particular *ideology* of *parties*, *leaders*, *presidents*, or even *citizens*. The political novelist may also be seen as a political historian who reflects on the political history of a particular period or event, such as *treaties*, *world struggles*, and *presidential elections*.

In Group 5, the most distinctive words and phrases were: *some money*, *a lot of money*, *lend him money*, *out of money*, *spending too much money*, *marrying for money*, *demands for money*, *make money*, *love*, *love affairs*, *true love*, *make(s)/made love*, *marriage and divorce*, *extremely/very poor*, *poor boy/girl*, *American women/girls*, *American world*, *American students*, *loose women*, *unmarried women*, *unhappy women*, *get into trouble*, *hope*, *hope to marry*, *poor/large family*, *happy life*, *death*, *aristocracy*, *wealth and social position*, *Washington/New York society*, *social and political changes*, *dirty and ugly*, *poverty and ignorance*, *poverty and vice*, *poverty and homelessness*, *the slums*, *prostitutes*, *became a prostitute*, *companies*, *frustration*, *(complete) loss and tragic*. This class of works included: John Barth's *Floating Opera* (1979); Mark Twain's *Adventures of Huckleberry Finn* (1884); Edith Wharton's *The House of Mirth* (1905); Stephen Crane's *A Girl of the Streets* (1983); Henry James' *Daisy Miller* (1879) and *The Portrait of a Lady* (1881); Paul Laurence Dunbar's *The Love of Landry* (1900) and *The Sport of the Gods* (1902); and Norman Mailer's *The Naked and the Dead* (1998). The texts in this group can best be described as social realism or realistic fiction. Texts in this category usually highlight social events and issues relevant to contemporary life, such as falling in love, marriage, finding a job, divorce, alcoholism, etc. Social realist novels usually have elements of romance (CLAYBAUGH, 2018; KAPLAN, 1992), which is why the most distinctive words in this group are those related to *love*, *marriage*, and *affairs*.

In Group 6, the most distinctive concepts were: *concentration camps*, *barbed wire*, or *battleships*, *missiles*, *tanks*, *World War I*, *World War II*, *atrocities*, *win the war*, *soldiers*, *lieutenant/s*, and *troops*. Texts in this group included: Ernest Hemingway's *A Farewell to Arms* (1929) and *For whom the Bell Tolls* (1940); Ralph Peters' *Red Army* (1989); and Kevin Powers' *The Yellow Birds* (2012). An appropriate heading or label for this group may be war novels. The genre of the war novel has given authors the opportunity to examine both the best and the worst of human nature. Words such as *concentration camps*, *barbed wire* or *battleships*, *missiles*, and *tanks* speak of this genre as penetrating into popular consciousness and words like *gas chambers* and *Nazi atrocities* remind people of the savagery of the world wars.

In Group 7, the most distinctive variables were: *villain*, *save the world*, *destroy the evil*, *spill/s*, *curse*, *race*, *power*, *witch*, *battle creatures*, and *dragons*. This group included: Katherine Arden's *The Bear and the Nightingale* (2017); Glen Cook's *The Black Company* (1984); and David Eddings's *The Belgariad* (1993). These works may be classified under the genre of fantasy fiction. Fantasy novels generally overlap with science fiction novels. This may be attributed to the fact that so far there is no complete definition of either that is

unanimously agreed upon. For many critics, however, fantasy novels are normally imaginary tales about the struggle between good and evil in non-real or fictional settings (CHESTER, 2016). Heroes usually have superpowers that enable them to destroy villainous creatures and save the world and the human race.

It is evident then that this categorization of the texts was not based in anyway on any biographical considerations, but that it was only the semantics of texts that were used. Works such as Sutton Griggs' *Unfettered* (1902), and Paul Laurence Dunbar's *The Love of Landry* (1900) and *The Sport of the Gods* (1902), for instance, which have been traditionally classified under the genre of African American literature, were not classified under the Immigrant novels in Group 1. It is true that such works deal with the social conditions and realistic aspects of American life such as unemployment, crime, prostitution, and poverty. However, these works are not limited to the culture and problems of black Americans in the United States. Similarly, Anita Diamant's novel *The Boston Girl* was grouped under social realist novels, although her works are traditionally classified under Jewish literature. The idea that Diamant is concerned with Jewish issues in different books, including *Choosing a Jewish Life* (1998) and *Living a Jewish Life* (2007), has led critics to classify her novels under the heading of Jewish novels. In *The Boston Girl*, however, Diamant is concerned more with reflecting on the complicated life American women had to lead in the twentieth century. Likewise, Jin's *War Trash* (2005) and David Anthony Durham's *The Sacred Band* (2011), which are usually classified as ethnic American literature, are classified in this study as war novels. Race and ethnicity are not good indicators for genre analysis and there is always a need for internal evidence for genre classification and analysis in literary studies.

6 Conclusion

The system in this study proved itself to be useful in improving classification performance of literary texts in assigning appropriate and meaningful attributes to each group or category. It may be claimed that an objective, replicable, and thus reliable genre classification of literature is now possible. The findings of this study support the growing body in the literature indicating how the continuing development of ATC algorithms offers new pathways for the classification of literature. With the increase in the number of literary texts available and the emergence of numerous literary subgenres, it may be suggested that conventional methods are no longer appropriate to genre classification studies. The findings of this study can be extended to other literary disciplines, including thematic analysis, automatic summarization, stylometry, and authorship attribution, which are based on classification applications. The implications of the study are also useful for digital libraries and archives whose classifications of literary texts are in many cases misleading to readers and users.

7 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 unstintin-

gly provided towards the fulfillment of the current research project.

References

ADOLPHS, S.; KNIGHT, D. *The Routledge Handbook of English Language and Digital Humanities*: Taylor & Francis, 2020.

BAASNER, R.; ZENS, M. *Methoden und Modelle der Literaturwissenschaft:* Eine Einführung (Methods and Models of Literary Studies: An Introduction (Revised Edition). Berlin: Erich Schmidt, 2005.

BAYM, N. *The Norton Anthology of American Literature* (7th ed.). New York; London: W.W. Norton & Co., 2007.

BELLEGARDA, J. *Latent Semantic Mapping: Principles And Applications.* Morgan & Claypool Publishers, 2008.

BENDIXEN, A. *A Companion to the American Novel.* 1st ed.. Chichester, West Sussex: Wiley-Blackwell, 2012.

BERRY, D. M. *Understanding Digital Humanities*. Basingstoke: Palgrave Macmillan, 2012.

BHATIA, V. K. *Analysing Genre : Language Use in Professional Settings*. London: Longman, 2014.

BIBER, D. *Spoken and Written Textual Dimensions in English*: Resolving the Contradictory Findings. *Language, 62*(2), p. 384-413, 1986.

BLOTNER, J. *The Modern American Political Novel, 1900-1960*. Austin: University of Texas Press, 1966.

BLOTNER, J. *The Political Novel.* Norwood Editions, 1977.

BOELHOWER, W. Q. The Immigrant Novel as Genre, *MELUS, 8*(1), p. 3-13, 1981.

CARISSIMO, J. *Yi-Fen Chou*: White author under fire after using Asian pen name to be published more often. *The Independent*, 8 September 2015.

CASSUTO, L.; REISS, B. *The Cambridge History of the American Novel*. Cambridge: Cambridge University Press, 2011.

CHAKRABORTY, G.; PAGOLU, M. *T*ext Mining and Analysis: Practical Methods*, Examples, and Case Studies.* SAS Institute, 2014.

CHAKRABORTY, G.; PAGOLU, M.; GARLA, S. Text Mining and Analysis: Practical Methods*, Examples, and Case Studies Using SAS*. Cary, North Caroline: SAS Institute, 2014.

CHESTER, D. *The Fantasy Fiction Formula.* Oxford University Press, 2016.

CLAYBAUGH, A. *The Novel of Purpose:* Literature and Social Reform in the Anglo-American World. Cornell University Press, 2018.

COULSON, S. *Semantic Leaps:* Frame-shifting and Conceptual Blending in Meaning Construction. Cambridge: Cambridge University Press, 2006.

CRUSE, D. A. *Lexical Semantics*. Cambridge: Cambridge University Press, 1986.

CRUSE, D. A. *Meaning in Language:* An Introduction to Semantics and Pragmatics. Oxford: Oxford University Press, 2000.

DOUGLAS, D. *The multi-dimensional approach to linguistic analyses of genre variation*: An overview of methodology and findings. *Computers and the Humanities,* 26(5-6), p. 331-345, 1992.

DUNN, J.; ARGAMON, S.; RASOOLI, A.; KUMAR, G. Profile-based authorship analysis. *Digital Scholarship Humanities*, *31*(4), p. 689-710, 2016.

ERLICH, V. *Russian Formalism:* History- Doctrine. New Haven; London: Yale University Press, 3rd ed., 1981.

FANG, L.; MEHLITZ, M.; LI, F.; SHENG, H. Web Pages Clustering and Concepts Mining: An approach towards Intelligent Information Retrieval. Cybernetics and Intelligent Systems, *IEEE Conference*, 2006, p. 1-6.

FLOOD, A. *White poet used Chinese pen name to gain entry into Best American Poetry*. The Guardian, 8 September 2015.

FOWLER, A. *Kinds of Literature:* an introduction to the theory of genres and modes. Harvard Univ Press, 1982.

FRANCO, D. J. *Ethnic American Literature:* Comparing Chicano. Jewish, and African American Writing. Charlottesville; London: University of Virginia Press, 2006.

GABRILOVICH, E.; MARKOVITCH, S. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. P*roceedings of the Twenty-First National Conference on Artificial Intelligence*, 2006, p. 1301-1306.

GABRILOVICH, E.; MARKOVITCH, S. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis*. Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007, p. 1606-1611.

GELBUKH, A. *Computational Linguistics and Intelligent Text Processing*: Springer, 2007.

GOMAA, D. *The Non-National in Contemporary American Literature:* Ethnic Women Wri-

ters and Problematic Belongings. Palgrave Macmillan, 2016.

GRICE, H. *Beginning Ethnic American Literatures*. Manchester: Manchester University Press, 2001.

GRIFFITHS, T. L.; STEYVERS, M. Topics in Semantic Representation. *Psychological Review, 114*(2), p. 211-244, 2007.

HAMMOND, A.; BROOKE, J. *A* Tale of Two Cultures: Bringing Literary Analysis and Computational Linguistics Together*. Proceedings of the Second Workshop on Computational Linguistics for Literature*, Atlanta, Georgia, June 14, 2013.

HAN, J.; KAMBER, M. *Data mining:* concepts and techniques. San Francisco, Calif.; London: Morgan Kaufmann, 2001.

HAVASI, C.; SPEER, R.; ALONSO, J. Conceptnet 3: A Flexible Multilingual Semantic Network for Common Sense Knowledge. *Recent Advances in Natural Language Processing*, 2007, p. 27-29.

HOLMES, D. I. The Evolution of Stylometry in Humanities Scholarship. *Lit Linguist Computing, 13*(3), p. 111-117, 1998. doi:10.1093/llc/13.3.111

HOWE, I. *Politics and the novel*. New York; London: Columbia University Press, 1992.

JOACHIMS, T. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Kluwer Academic Publishers, 2002.

JOCKERS, M. L. *Machine-Classifying Novels and Plays by Genre*. Retrieved from: https://www.stanford.edu/~mjockers/cgi-bin/drupal/node/27, 13 February 2009.

KAPLAN, A. *The Social Construction of American Realism*. Chicago: University of Chicago Press, 1992.

KARCZ, A. *The Polish Formalist School and Russian Formalism*. Rochester, N.Y. ; Woodbridge: University of Rochester Press, 2002.

KESSLER, B.; NUMBERG, G.; SCHTZE, H. Automatic Detection of Text Genre. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics.* Madrid, Spain, 1997.

KOPPEL, M.; ARGAMON, S.; SHIMONI, A. R. Automatically Categorizing Written Texts by Author Gender. *Lit Linguist Computing, 17*(4), p. 401-412, 2002. doi:10.1093/llc/17.4.401.

LIAO, S.-H.; CHU, P.-H.; HSIAO, P.-Y. Data Mining Techniques and Applications – A Decade Review from 2000 to 2011. *Expert Systems with Applications, 39*(12), p. 11303-11311, 2012.

LIU, H.; SINGH, P. ConceptNet- A Practical Commonsense Reasoning Tool- Kit. *BT TEch-*

nology Journal, 22(4), p. 211-226, 2004.

LOOKS, M.; LEVINE, A.; COVINGTON, G. A.; LOUI, R. P. A. L. R. P.; LOCKWOOD, J. W. A. L. J. W.; CHO, Y. H. A. Streaming Hierarchical Clustering for Concept Mining. *Aerospace Conference*, 2007 IEEE.

MAJKIĆ, Z. *Big Data Integration Theory:* Theory and Methods of Database Mappings. Programming Languages, and Semantics, Springer, 2014.

MANDELKER, A. Russian Formalism and the Objective Analysis of Sound in Poetry. *The Slavic and East European Journal, 27*(3), p. 327-338, 1983.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

MICKENBERG, J.; VALLONE, L. *The Oxford Handbook of Children's Literature*. Oxford University Press, 2012.

NELSON, E. S. *Ethnic American Literature*. California; Oxford: Greenwood Press, 2015.

OLMOS, R., LEÓN, J. A.; JORGE-BOTANA, G.; ESCUDERO, I. Using latent semantic analysis to grade brief summaries: A study exploring texts at different academic levels. *Lit Linguist Computing, 28*(3), p. 388-403, 2013.

OMAR, A. 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. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science, 1*(2), p. 1-14, 2010.

OMAR, A. *Addressing Subjectivity in Thematic Classification of Literary Texts:* A Fresh Look at Thomas Hardy's Prose Fiction. Berlin: Lambert, 2015.

OZGUR, Y. *Empirical selection of nlp-driven document representations for text categorization.* Syracuse University, 2006.

PREDELLI, S. *Contexts:* Meaning, Truth, and Use of Language. Oxford: Oxford University Press, 2005.

RAMSAY, S. In Praise of Pattern. *TEXT Technology: the Journal of Computer Text Processing, 14*(2), p. 177-190, 2005.

RAMSAY, S. Algorithmic Criticism. In: SIEMENS, R. G.; SCHREIBMAN, S. (eds.), *A companion to digital literary studies* (Vol. A companion to digital literary studies, pp. xx, 620 p.). Malden, MA: Blackwell Publishers, 2007.

RAMSAY, S.; STEGER, S. Distinguished Speakers: Keyword Extraction and Critical Analysis with Virginia Woolf's The Waves. *Digital Humanities*, Sorbonne, Paris, 2006.

RIESEN, K.; BUNKE, H. *Graph classification and clustering based on vector space embedding*. Singapore; London: World Scientific, 2010.

ROBERTSON, S. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation, 60*(5), p. 503-520, 2004.

SARICKS, J. G. *The Readers' Advisory Guide to Genre Fiction*. Chicago; London: American Library Association, 2009.

SHALABY, W.; ZADROZNY, W. Semantic Representation Using Explicit Concept Space Models. *Proceddings of the 31 AAAI Conference on Artificial Intelligence*, 2017, p. 4983-4984.

SPÄRCK JONES, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*, p. 11-21, 1972.

SPEARE, M. E. *The Political Novel:* Its Development in England and in America. New York: Oxford University Press, 1924.

SPEER, R.; HAVASI, C. ConceptNet 5: A Large Semantic Network for Relational Knowledge. In GUREVYCH, I.; KIM, J. (eds.). *The People's Web Meets NLP- Collaboratively Constructed Language Resources*: Springer, 2013, p. 161-176.

STEINER, P. Russian Formalism. In: SELDEN, R. (ed.), *The Cambridge History of Literary Criticism*. Cambridge: Cambridge University Press, 8 ed., 1995, p. 11-29.

TOBIN, Y. *The Prague School and its Legacy n Linguistics, Literature, Semiotics, Folklore, and the Arts*. Amsterdam; Philadelphia: J. Benjamins, 1988.

UTAS, B. Genres in Persian Literature 900-1900. In: LINDBERG-WADA, G.; PETTERSSON A.; PETERSSON, M.; HELGESSON, S. (eds.), *Literary history:* towards a global perspective. Berlin; New York: W. de Gruyter, Vol. 2, p. 199-242, 2006.

WALKOWITZ, R. *Immigrant Fictions:* Contemporary Literature in an Age of Globalization. Madison: Wisconsin University of Wisconsin Press, 2010.

WEI, T.; LUC, Y.; CHANGB, H. A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications, 42*(4), p. 2264-2275, 2015.

WEIPING, W.; PENG, C.; BOWEN, L. A Self-Adaptive Explicit Semantic Analysis Method for Computing Semantic Relatedness Using Wikipedia. *Proceedings of the 2008 International Seminar on Future Information Technology and Management Engineering*, 2008.

WILDER, L. A. *Rhetorical Strategies and Genre Conventions in Literary Studies:* Teaching and Writing in the Disciplines. Southern Illinois University Press, 2012.

WOLTERS, M.; KIRSTEN, M. Exploring the Use of Linguistic Features in Domain and Genre Classification. *Proceedings of the ninth conference on European chapter of the As-*

sociation for Computational Linguistics, Bergen, Norway, 1999.

XIAO, Z.; MCENERY, A. Two Approaches to Genre Analysis: Three Genres in Modern American English. *Journal of English Linguistics, 33*(1), p. 62-82, 2005. doi:10.1177/0075424204273957.

Recebido em dia 14 de abril de 2020.
Aprovado em dia 11 de junho de 2020.