A Hybrid Semantic Mapping Approach to the Genre Classification of Literature

The American Novel as an Example

Dr. Abdulfattah Omar
Assistant professor of English language and linguistics
Department of English, College of Science and Humanities, Prince Sattam Bin Abdulaziz University, Al-Kharj, 11942, Saudi Arabia
Department of English, Faculty of Arts, Port Said University, Egypt

Abstract

Egypt, as an African country, was tied over its long history with the people of Nile Basin and Horn of Africa. Therefore, cooperation with the countries of Africa aims at consolidating the relations by using and effectuating mechanisms of cooperation in the field of international relations and boosting the role of Egyptian private sector by driving the fields of Egyptian investment in infrastructure projects.

Development and growth of relations depended on internal variables and their external impacts and growth of the power of Egypt to be reflected at the end on strengthening the relations between Egypt and the countries of Africa.

Egypt's relation to the countries of Africa shall not be reduced to trade, economy or Egypt's need to Nile water. It is time to revive the soft power
of Egypt represented in education, scientific research, media, culture and joint work with civil community organizations to demonstrate the real image of Egypt and Africans and their civilizations. Problem of the study lies in focus of President Abdulfattah Al Sisi on Africa and its development, and the strategic depth of the content in Egypt since he came to power, particularly after tension of relations between Egypt and Africa in 1980s and 1990s and the fact that they were limited to regional and internal official meetings, and after Ethiopia established the Renaissance Dam following January 2011 Revolution. President Sisi endeavored since he came to power to focus on Africa and Development of African countries. The president assures in all his interviews the importance of linking communication between Egypt and Africa, and stresses on the importance of Africa for Egypt and importance of Egypt for Africa.

The study depended on the theories of media frameworks to scientifically authenticate the thesis and decision-making theory in international relations because it is one of the theories that interpret the foreign policy of the state as a theory in international relations. In addition, it is used for the decisions made by audience as a result of watching satellite channels.

The study used survey of the context of public and private Egyptian satellite channels which address the issue of Egyptian-African relations and the field survey of various community categories to identify the role of public and private Egyptian satellites in forming the approaches of the case.
Population of the study included:

- Population of the analytical study includes public and private Egyptian satellites that handle the Egyptian-African relations in their program content and news bulletins.
- Population of field study includes experts and categories of the Egyptian society in various ages (public opinion).

The study depended on purposeful sample of 400 respondents of experts and specialists in the Egyptian-African affairs and random sample of the Egyptian public represented in university youth and staff of Ministries, Embassies, Arab and African institutes.

The researcher made context analysis form to identify the extent of Egyptian public and privates satellites treatment of the issue of Egyptian-African relations on sample of Egyptian public and private satellites.

Concerning Egyptian public satellites: (Egyptian Satellite TV, Nile News)

Concerning Egyptian private satellites: (Extra News- ON Live)

Programs presented on the channels the sample of the study for full television period of:

(July- August- September 2017) for 2017, using the comprehensive survey approach.

Key results of the study

- Public local organizations and authorities were the key source of news related to the Egyptian-African relations in Extra News Channel with percentage of 100%, in ON Live channel with percentage of 94.9%; in the Egyptian satellite TV with percentage of 91.4%; and in Nile News with percentage of 89.2%
- Coverage of political conditions gained the biggest side of coverage of Egyptian-African relations in the channels under study.
- Egypt was one of the countries that aired the four channels under study and to which those channels belong. Therefore, those channels were keen on highlighting the Egyptian side and its role in most coverage under study.
- Majority of contexts of the four channels presented single point of view with percentage of 98.3% in ON Live; 95.7% in satellite TVs, with percentage of 95.6% in Nile News TV.

- Private Satellite TVs acquired the highest percentage of viewers. While the percentage of respondents who always follow private channels was 60.2%, the percentage of permanent follow up didn’t exceed 21.7% between audience of public TVs. In addition, while there were only four respondents who don’t follow private satellite TVs, number of respondents who don’t follow public satellite TVs was 26 respondents with percentage of 6.5% of the total sample of the study.

- The first ranking was Egypt's relations with the countries of Africa which are governed by exchange of benefits and interests with relative weight of 85.2%, which means that respondents' perception of one of the essential facts in relations between countries and its implication that relations of countries are primarily based on mutual benefit between them.

- Result notes that role of Egyptian public and private satellites in addressing the Egyptian-African relations was clear for the respondents the sample of the study.

Suggestions of the study:

- To reconsider the foreign policy of Egypt towards African. Egypt shall play key role in realization of political and economic stability in the Nile Basin countries to encounter any foreign interventions that endeavor to marginalize the regional role of Egypt towards Africa.

- News of Egyptian-African relations aired through Egyptian public and private satellites shall be true, accurate and direct.

- Preparers and editors of programs related to the issue of Egyptian-African relations in all sides and dimensions of the issue and training in arts of media editing shall be provided.

- Preparers and editors of the programs that handle Egyptian and African relations and issues shall be proficient in the languages and
dialects of those countries and shall be familiar with the culture of those peoples.

- There shall be Egyptian-African media strategy to address the foreign claims in the resources of Africa.
A Hybrid Semantic Mapping Approach to the Genre Classification of Literature

The American Novel as an Example

Abstract

One of the limitations of genre classification of literature is that it is either based on philological methods or standard automatic text classification (ATC) methods. Moreover, a gap between literary analysis and computational theory further highlights this limitation. In order to bridge this gap, this study discusses a genre classification approach based on topic concepts rather than topic words where texts are represented using the Bag of Concepts (BOC) model due to the rich knowledge sources and meaningful relations concepts have. Texts were classified using a hybrid model of explicit semantic analysis (ESA) and ConceptNet. Results indicate that the proposed model led to better classification performance in relation to describing the connections between the document-members of each cluster and making generalizations about their unifying genre. 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.

Key words: genre classification- concept mapping/representation- ConceptNet- Explicit Semantic Analysis (ESA) - Vector Space classification (VSC) - semantic mapping

1. Introduction

Different approaches have been proposed for the genre classification of literary texts based mainly on textual content and/or biographical considerations using what is referred to as the ‘philological method’, that is, by individual researcher’s reading of printed materials and the intuitive abstraction of generalisations from that reading. With the advent of electronic text, however, a large number of literary works have become available, and this electronic format has permitted computational data analysis concepts and procedures to be applied to them. In bridging the gap between literary studies and the computational theory, different classification systems including vector space clustering and naïve Bayes have been used in processing literary texts for genre classification,
authorship attribution, and thematic analysis purposes. However, questions of reliability and meaningfulness are often 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 should be considered too. To put it into effect, standard (also referred to as classical) classification systems suffer two main problems. First, they do not consider the semantic relationships among words; therefore, the meanings of documents are not accurately represented. Second, there is a gap between extracting the most distinctive features within a given corpus and assigning appropriate description for the generated clusters. In the face of 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. For classification purposes, texts are classified using a hybrid model of Explicit Semantic Analysis (ESA) and ConceptNet.

By way of illustration, this study is based on the automatic classification of 346 American novels written by 86 novelists. The American novel is rich with numerous genres which include 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. Works of the period are traditionally classified on the historic period and/or the background of the author. The result is that many texts were differently classified due to the 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 (Douglas, Biber, 1986; Douglas Biber, 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 meaningfulness, usefulness and reliability of these classifications are still unresolved. In relation to the classification of literature, standard classification methods are not effective in representing document meanings and assigning appropriate descriptions for the generated clusters which have their negative implications 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 reconciliation between literary analysis and
the computational theory makes it possible to overcome many of the inherent problems within literary studies including genre classification and analysis. In this context, this study is concerned with exploring alternative methods for traditional classification methods and systems for suggesting more meaningful and reliable genre classification of literature.

2. Literature Review: genre classification

Genre classification of literary texts— the process of classifying texts according to what they have in common, either in their formal structures 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 an old discipline in critical studies, it is often argued that genre classification is significant in critical studies in a number of ways. First, identifying the genre of a text gives us a potential opportunity to have a better idea of its intended overall subject. Second, grouping similar texts together can deepen our sense of the value of any single text, by allowing us to view it comparatively, alongside many other texts of its type (Fowler, 1982; Saricks, 2009). Third, with so many texts available today, sometimes our information need is not for something specific. Rather, we seek answers for general questions which are typically answered by techniques which look at the entire document, or set of documents such as clustering, categorization, or genre classification.

Although the idea of genre classification is as old as Aristotle’ Poetics, so far there is no systematic classification approach to literary works (Utas, 2006). There are no fixed or universal formulas that can be adopted in objective and replicable manner in defining literary genres or classifying literary texts. With the lack of standards, genre classification has traditionally been based partly on textual content and partly on biographical considerations. These have been generated by what can be defined as ‘philological method’, that is, by individual researcher’s reading of printed materials and the 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 opened new avenues for more objective methods. As an example, the 2015 Edition of The Best American Poetry included a poem by a 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 the Chinese pseudonym in order to get his poems published. According to Hudson, his poems were rejected over 40 times only for 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 the ethical considerations of the issue, this problem brought into consciousness 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 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 descents. Works of the kind are usually concerned with concepts such as national consciousness, time, space, and belonging (Baym, 2007; Franco, 2006; Gomaa, 2016; Grice, 2001). Nevertheless, classifications of such works are usually based on the ethnic background and biographical considerations of authors even if they do not 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, 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, literature came to be considered as a science. The formalists developed what is called a scientific method for studying the poetic language (Erlich, 1981). In spite of severe criticisms, the school initiated debates about the importance of adopting objective grounds in evaluating poetry and literature. Today, there is a tendency to rebuild literary studies based on scientific grounds.
where objectivity and empirical determination are given the highest priority. The recent years have witnessed the development of what can be translated into English as literary science. This is originally a German term ‘Literaturwissenschaft’ that refers to all disciplines of literature that adopt the scientific method (Baasner & Zens, 2005).

Over the last two decades, the developments of computational approaches have encouraged scholars to think about novel approaches that can address many inherent problems within literary studies. Works of the kind are classified under the broad headings of ‘digital humanities’ and/or ‘literary computing’ studies, where researchers use computational methods either to answer existing research questions or to challenge existing theoretical paradigms, generating new questions and pioneering new approaches (D. M. 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 the keywords within these texts and finally labels or genres are suggested for each group (Chakraborty & Pagolu, 2014; Chakraborty, Pagolu, & Garla, 2014; Manning, Raghavan, & Schütze, 2008; Riesen & Bunke, 2010).

VSC ignores the word order and the context where 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 independence assumption. It estimates the probabilities of each attribute based on their frequencies over 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 given attribute depend on each other or upon the existence of other features, the classifier considers these attributes to be belonging to a given class. Both VSC and naïve Bayes are based on what is known as the Bag of Words (BOW) representation model and they are described today as classical classification systems. 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 vocabulary with no regard of syntax, semantic relationships among words, or the context where 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 considered. In other words, standard classification systems based on word similarity may not be feasible for literary data. In traditional VSC methods, based on keyword indexing, there are serious shortcomings that need to be considered in literary studies. In such methods, classification is based on word similarity where users can form impressions of what texts are about. In genre classification studies, however, critics are rarely interested in lexical databases. Rather, they need concepts that are often difficult to be abstracted, represented, and processed. Olmos, León, Jorge-Botana, and Escudero (2013) argue that in spite of the reasonable success of recent classification systems, their effectiveness in dealing with literary texts is still very limited. They explain that literary studies including genre classification, authorship attribution, and thematic analysis have tended to use what is called BOW representation which yielded controversial and even confusing results. BOW is a model of representation that is adopted in different classification methods including VSC and Naïve Bayes classification where similar texts are grouped together based on the salient terms or features within the corpus. One main problem with this model is that meanings of documents are not represented in an appropriate manner.

Although the BOW model proved to be effective in different classification tasks including news and scientific articles (M. W. Berry, 2004), the claim here is that it is not appropriate for the applications of literary studies. The general assumption in ATC is that extraction of the most distinctive or key features within a given corpus leads to successful grouping of similar texts together. This can be illustrated by the following example. Given a corpus of news reports is classified into two main clusters or groups where the most distinctive features of the first group 1 are the terms Obama, Pentagon, Congress, policy, politics, issue;
administration, peace, terrorism, China, Arab, Middle, East, Israel, United, Nations, Security, Council, and those of group 2 are Messi, Ronaldo, World, Cup, FIFA, champions, league, premier, and football; it will be easy to assign appropriate and meaningful to each cluster or group. It can be meaningfully suggested that texts in Group 1 are classified together because they can be categorized under the broad heading “World Politics”, while those in Group 2 are 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 for generating successful and reliable classifications since the lexical relations of these terms are not considered. In his classification of Shakespeare’s plays, Ramsay (2005) used the VSC based on BOG model for objective genre classification of the plays. The plays were grouped in 4 distinct clusters. These are 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 cannot be quite convincing. However, he stressed the importance of thinking about objective criteria for genre classification. Similarly, Jockers (2009) attempted a genre classification of Shakespearean plays where he used Naïve Bayes method and BOG model for generating a classification of 37 plays of Shakespeare. The plays were classified under three distinct groups which Jockers labelled as comedies, histories, and tragedies. Again, there were some problems with Jopckers’ analysis. For some texts, there was no justification for classifying some texts under one category or group. Although some classifications of literary texts seemed reasonable and were supported by internal and external evidence, the distinctive lexical features 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 4 distinct groups. The most distinctive lexical features of Group 2, for instance, are the words husband, captain, squire, France, Casterbridge, sergeant, curate, dance, countess, knight, duke, shepherd, architect, and horse, which cannot by themselves express or indicate the themes or genre they represent. In this way, literary texts
need to be addressed differently. More reliable methods need to be developed in relation to the classification of literary texts.

3. Statement of the problem

The recent years have witnessed great developments in the quality and effectiveness of automatic text classification (ATC) performance. ATC is now widely used in different applications including information retrieval (IR) and text mining. In the field of IR, for instance, research engines including Google and Yahoo have made considerable achievements in developing classification methods and approaches with the purpose of successfully grouping similar texts together. These developments have their positive implications on the IR performance of these research engines and reliability on ATC in general (Liao, Chu, & Hsiao, 2012; Wei, Luc, & Changb, 2015). The success of automatic classification methods has encouraged researchers in different disciplines including medicine, engineering, and even linguistics to adopt such classification systems in dealing with their data. In literature, unfortunately, applications are still beyond the expectations. This can be due to the idea that literary analysis and computational theory used to be seen as two different cultures. Hammond and Brooke (2014) assume that there are cultural barriers that make it difficult for literary scholars and researchers to make use of the computational theory. In this way, Kessler et al. (1997) assert that genre classification has long been ignored in computational linguistics. Furthermore, the ineffectiveness of the computational methods in dealing with literary texts has raised many doubts concerning their usefulness. This can be attributed in part to the different problems within the BOW model and standard classification methods that used to had their negative implications on the reliability of classification performance. These included the misrepresentation of document meanings as well as the failure to assign meaningful description for generated clusters. In the light of the limitations of ATC systems in relation to the literary studies, this study proposes the integration of semantic relations among words using Bag of concepts (BOC) representation where the text is a vector in the space of concepts. The rationale is that concepts have rich knowledge and semantic sources which will have positive implications on generating meaningful categories that best describe the document collections. 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 into the concepts they represent. It is therefore referred to as a 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 collections and predict the associations between these features with for grouping 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 thought to be useful in supplementing text representation with the rich knowledge and information available on world internet-based encyclopaedias. Linking concepts with web-based wikis has the advantage of supplementing text representation with related and big amounts of world language.

In technical terms, 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 text can be improved by explicitly representing the meaning of any text in terms of encyclopaedia or knowledge-based concepts (Gabrilovich & Markovitch, 2006, 2007). ESA is essentially based on the use of Wikipedia due to the idea that it is the largest and most diverse online knowledge base so far and it has the biggest amounts of highly organized human knowledge so that concepts
can be related to articles in an easy and useful manner (Gabrilovich & Markovitch, 2006, 2007).

ESA is different from conventional clustering methods (basically vector space clustering) in the sense that it represents the meaning of a text, not just the vocabulary meaning. The text here is seen as a combination of concepts which are found in the text not only a sequence of words. In this way, ESA lowers the dependence on the key words (known as the distinctive features in BOW) and thus improves the clustering performance. ConceptNet is used for the purposes of this study. In other words, this study adopts ConceptNet rather than WordNet since the former has more semantic relationships and larger vocabulary. Furthermore, due to the limitedness of WordNet, many relations and concepts are missing; therefore, it is not appropriate for literary data where more concepts and relations are needed to have a general idea of what texts are about.

Although ESA has proven efficacy for semantic analysis of textual data and especially short texts, experimental results indicate that ESA has some problems. Weiping et al. (2008) argue that ESA is not effective in dealing with long documents. Shalaby and Zadrozny (2017) add that ESA has some problems that include vectors are sparse causing 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 which cannot be generated by relying on Wikipedia or different knowledge bases as the main and only source. In the face of 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 an automated categorization of documents based on extracting concepts embedded in documents. It is a workflow that is used to discover implicit and explicit relationships, useful associations and groupings in a set of documents or data collection with the purpose of detecting similar documents in a large corpora and classifying them by topic. It can provide thus powerful insights into the meaning, provenance and similarity of documents (Fang, Mehlitz, Li, & Sheng, 2006; Han & Kamber, 2001; Looks et al., 2007). The assumption is that each word in a given document relates to several
possible concepts which makes it possible to cluster documents based on their content. The underlying principle of ConceptNet is that texts are processed and general concepts are extracted. This is done in two subsequent steps. First, documents are reduced into a sequence of words that describes the content. Second, these words are mapped into concepts. In this way, given that we have a number of documents on generative grammar; ConceptNet is conducted by means of identifying relationships and generating facts based on the data within collection and the dimensions of the subject. That is, a concept may carry different importance in different sentences and/or documents. These can be something like Chomsky and generative grammar; theoretical linguistics and generative grammar; Phrase Structure Rules (PSR) and Generative grammar; deep and surface structures in generative grammar; etc. Documents can also be classified by topic as WH-movement; linguistic competence; etc. In our case, ConceptNet is 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 in the sense that it effectively captures a wide range of commonsense concepts and relations which 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 focus on concepts rather than words. ConceptNet, however, is based on more diverse relational ontologies that makes it, along with ESA, more useful in making practical, context-oriented, commonsense inferences over real-world texts. The implication here is that the integration of ESA and ConceptNet will lead to a better classification performance which will be useful in classifying texts of the same genre together and assuming appropriate categories for these texts (Havasi, Speer, & Alonso, 2007; Speer & Havasi, 2013).

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 the 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 because of availability and trustworthiness. A list of the selected texts is illustrated 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 in order to abstract all and only the comprehensible concepts within the documents. In this method, each concept is depicted in the form of a node and has a semantic interpretation and is associated with Wikipedia and ConceptNet. The rationale is that ConceptNet is a large semantic network consisting of large number of common-sense concepts (Havasi et al., 2007; Liu & Singh, 2004). One major problem with the corpus, however, is the high dimensionality of data which definitely has negative implications on the interpretability of the results generated and the reliability of classification performance.

The problem with high dimensionality of data is that there is lots and lots of irrelevant information which makes clustering performance poor and unsuccessful. In ATC applications, the larger the data dimensionality, the more difficult it becomes to define the manifold sufficiently well to achieve reliable analytical results. Therefore, classification applications or tasks should be based only on the most important or distinctive features because irrelevant and redundant information has a fundamental bearing on the reliability and accuracy of classification performance. In order to address this problem, only distinctive concepts were selected by means of weighing vector concepts using TF-IDF (Term Frequency Inverse Document Frequency) 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 characterizes the set of documents. In TF-IDF, the most discriminant terms are the highest TF-IDF variables. The implication to document clustering is that if the highest TF-IDF variables, which are taken to be the most discriminant
terms, are identified, then unimportant variables can be deleted and data
dimensionality is reduced. In our case, TF-IDF was used to identify the
most statistically significant concepts, not words since this study takes
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 are not significant for the classification task and retaining
only the most distinctive concepts. To put it into context, 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 good advantage about TF-
IDF is that it helps 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 is more appropriate for the purposes
of the study. It is a rich digital knowledge platform that provides highly
relevant and authentic information. It is also constantly updated in order
to give people the timely information they need and trust. As a second
step, ConceptNet was used in order to improve and support the clustering
performance.

5. Results
Having 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 is generated based on the idea that each
group has a number of distinctive features where members of the same
group are semantically and conceptually similar making it possible to
divide the fiction section by genre. There are features which suggest that
members of each group share a number of genre features that make them different from those of other groups or clusters. Texts in Group 1, for instance, are best described as immigrant or ethnic American novels where the most distinctive features are 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 (Boelhower, 1981; Walkowitz, 2010). The words like return home, a wish to return, yearning to return therefore tell cross border tales as stark and dark stories of wrath and oppression, humiliation and identity crisis 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 gradually constitute the genesis of a genre, that is, a new kind of genre is born from the depiction of the sufferings of the immigrants. The immigrant is no more a historical protagonist but an alien with multicultural visitations to a dystopian land of a 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 include 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 can be suggested then that these novels can be described as children’s novels. Mickenberg and Vallone (2012) argue that children’s novel has not only kept the
tradition of tales of simple fantasy, or folklore and folk tales as should be seen in words 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 more thrillers, more mysteries, and more adventures.

In Group 3, the most distinctive concepts include *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 does make sense thus to assign the genre science fiction to this group of novels. The genre of science fiction tells tales about science and technology, worlds in space and set in the future. These tales though are fictitious but closely resemble the theories of science. Closely resembling the genre fantasy, it often describes hence words like *against the stars*, *unknown spacecraft*, *alien spacecraft*, *climbed into the stars*, take the reader to distant regions in the space; at the same time, such descriptions like *biologically destroyed*, *center of gravity*, *virus*, *hydrogen bombs*, *atomic bombs* etc. speak of the abuses of science power by man. Grim it may be, but such expressions make this genre powerful and distinct from its closely resembling genres of dystopian fiction or fantasy novel.

In Group 4, the most distinctive variables are *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 assumed to be described as political novels. Political fiction has been recognized as a distinct genre in the 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 the world literature today, definition of the genre remains vague and unclear. Apart from the complexity and vagueness of the term, the political novel is usually seen as a political instrument that reflects the national character (Blotner, 1966, 1977) and the political milieu is the dominant setting (Howe, 1987). It is a tool of political propaganda for a particular ideology of parties, leaders, presidents, or even citizens. The political novelist is also seen as a political historian who reflects on the political history of a particular period or an event such as treaties, world struggles, and presidential elections.

In Group 5, the most distinctive words and phrases are 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 are best described thus as social realism novels or realistic fiction. Texts under this category usually highlight social events and issues that depict contemporary life, such as falling in love, marriage, finding a job, divorce, alcoholism, etc. Social realism novels have usually romance elements (Claybaugh, 2018; Kaplan, 1992). That is why, the most distinctive words in this group are words related to love, marriage, and affairs.

In Group 6, the most distinctive concepts are concentration camps, barbed wire or battleships, missiles and 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 can be war novels. The war novel genre has given the author 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 do speak of this genre as penetrating to people’s consciousness much like the dates of world wars when words like gas chambers and Nazi atrocities hounded the minds of people.

In Group 7, the most distinctive variables are 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 can be classified under the genre of fantasy fiction. Fantasy novels are always confused with science fiction novels. This may be attributed to the fact that so far there is no definition that is perfect or 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 (Armitt, 2005; Chester, 2016). Heroes usually have superpowers that enable them to destroy villain creatures and save the world and the human race from their evils.

It is evident then that the categorization of the texts was not based in anyway on any biographical considerations. It is only the semantics of
texts that is used for the classification of texts in our case. 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 the American life such as unemployment, crimes, prostitution, and poverty. These works are not limited to the culture and problems of the black Americans in the United States. Similarly, Anita Diamant’s novel *The Boston Girl* was grouped under the realistic novels although her works are traditionally classified under the Jewish literature. The idea that Diamant is concerned with the Jewish issues in different books including *Choosing a Jewish Life* (1998) and *Living a Jewish Life* (2007) made critics classify her novels under the heading of Jewish novels. In *The Boston Girl*, however, Diamant is concerned with reflecting on the complicated life women has to lead America in the twentieth century. Likewise, Jin’s *War Trash* (2005) and David Anthony Durham’s *The Sacred Band* (2011) which are usually classified under the ethnic American literature, is classified in this study as a war novel. In this way, race and ethnicity cannot be indicators of genre analysis. This supports the claim that there is always a need for internal evidence in the literary studies of genre classification and analysis.

6. Conclusion
The proposed system proved to be useful for improving the classification performance of literary texts in relation to assigning appropriate and meaningful attributes to each group or category. It can be then 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 literature indicating that the continuing development of ATC algorithms offers new avenues for the classification of literature. Within the unprecedented increase in the number of literary texts and the emergence of literary subgenres, it can be suggested that conventional methods are no longer appropriate in genre classification studies. The findings can also 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.

References
Erich Schmidt.


