Pertanika J. Sci. & Technol. 30 (4): 2535 - 2561 (2022)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

A Topic Modeling and Sentiment Analysis Model for Detection and Visualization of Themes in Literary Texts

Kah Em Chu[1], Pantea Keikhosrokiani[1]* and Moussa Pourya Asl[2]

[1]*School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia*
[2]*School of Humanities, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia*

ABSTRACT

Despite the growing emergence of new computer analytic software programs, the adoption and application of computer-based data mining and processing methods remain sparse in literary studies and analyses. This study proposes a text analytics lifecycle to detect and visualize the prevailing themes in a corpus of literary texts. Two objectives are to be pursued: First, the study seeks to apply a Topic Modeling approach with selected algorithms of LDA, LSI, NMF, and HDP that can effectively detect the recurring topics about the major themes developed in the dataset. Second, the project aims to apply a Sentiment Analysis model that can analyze the polarity of writers' discourse on the detected thematic topics with the algorithms of Vader and TextBlob. The implementation of Topic Modeling has detected six thematic topics of sex, family, revolution, imprisonment, intellectual, and death. The adoption of the Sentiment Analysis model also revealed that the feelings attached to all the identified themes are largely negative sentiments expressed towards socio-political issues.

Keywords: Iranian diaspora, life writing, sentiment analysis, text mining, topic modeling

ARTICLE INFO

Article history:
Received: 21 September 2021
Accepted: 04 March 2022
Published: 15 September 2022

DOI: https://doi.org/10.47836/pjst.30.4.14

E-mail addresses:
pantea@usm.my (Pantea Keikhosrokiani)
kahem@student.usm.my (Kah Em Chu)
moussa.pourya@usm.my (Moussa Pourya Asl)
* Corresponding author

INTRODUCTION

Collecting and processing textual evidence is widely recognized as the principal strategy to detect and evaluate the underlying themes of literary writings. The aim is to identify the prevailing view of the world or human nature by closely examining the story's plot, characterization, and the dominant conflicts within the text. Literary scholars traditionally use manual qualitative content analysis to perform this kind of examination

ISSN: 0128-7680
e-ISSN: 2231-8526

which poses certain methodological challenges (Ying et al., 2022). Due to the rising level of abstraction and the high degree of subjective interpretation used in the process, the credibility and authenticity of such analyses have always been a matter of critical debate and controversy (Graneheim et al., 2017; Ying et al., 2021). In recent years, advances in computer analytic software programs have made collecting and analyzing text corpus much easier by replacing manual processing with systematic and automatic procedures (Firmin et al., 2017; Misuraca et al., 2021). However, despite the emergence of new methods and analytic tools, the adoption and application of these new computer-based methods remain sparse in literary studies and analyses.

This study proposes a text analytics lifecycle for detecting and visualizing the dominant themes in a corpus of literary texts. Therefore, the main objectives of this study are as follows:

1. to apply topic modeling techniques for detecting the topics related to imprisonment from 28 Iranian diasporic life writings,
2. to apply sentiment analysis for analyzing the polarity of diasporic writers' discourse on imprisonment with the selected algorithms.

Topic Modeling (Shi et al., 2018) and Sentiment Analysis (Alaei et al., 2019) are analytical techniques in the proposed digital model. Topic modeling with selected algorithms of Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI), Non-negative Matrix Factorization (NMF), and Hierarchical Dirichlet Process (HDP) can effectively detect the recurring topics concerning the major themes developed in the texts. On the other hand, sentiment analysis can help us to analyze the polarity of writers' discourse on the major themes with the algorithms of Vader and TextBlob. However, even though both models enable researchers to produce more detailed textual analysis, the implementation of topic modeling and sentiment analysis on literary texts is scarce. This study aims to apply topic modeling and sentiment analysis to literary texts. Hence, the study would benefit literary scholars with a more accurate analytical tool and methodology and provide data scientists with a comparison of how different topic modeling and sentiment analysis algorithms perform on a text corpus.

The paper is structured as follows. After briefly reviewing the existing literature on Topic Modeling and Sentiment Analysis approaches, we elaborate on the proposed analytical strategy. An implementation on a text corpus related to different books is then presented to show how the strategy operates. Finally, the paper ends with a discussion of the results and concludes with theoretical and practical implications of the approach.

LITERATURE REVIEW

Natural Language Processing (NLP) is considered the interaction between computers and human language, which can be used for examining a text and generating insight from it. NLP

is widely used to study the opinion and sentiment of the target corpus. Sentiment analysis uses NLP and text analytics to identify, extract, analyze, and study subjective information. Topic modeling is another NLP technique for discovering the abstract "topics" that occur in a collection of documents. For instance, sentiment analysis is used to study the polarity of the opinion of Twitter users about the TV series "Game of Thrones." In contrast, topic modeling was used to visualize the weightage of the content related to the selected topic. For example, the topic can be a character or a place in the "Game of Thrones" TV Series (Scharl et al., 2016). This section includes the relevant literature related to topic modeling, sentiment analysis, and related studies.

Topic Modeling

Topic modeling refers to the weightage of a target theme in the whole corpus. For example, it can study the percentage of references to country names in a text (Costa, 2018). Topic modeling has a wide variety of visualization such as line charts, bar charts, pie charts, word clouds, and heatmaps. Topic modeling is used to group words for a set of texts. Because it automatically categories words without a specified list of labels, this is referred to as unsupervised learning (Sukhija et al., 2016). After feeding the data model, sets of words will appear from which the main topic can be assumed. However, it is challenging to understand the proper topic merely by looking at a combination of words and numbers. Therefore, topic modeling provides visualization, considered one of the most effective methods of understanding the data. Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA are popular topic modeling algorithms. LDA is famous for visualization using pyLDAvis. LSA and LDA are Vector Space Models (VSM) that represent text documents as vectors in a high-dimensional space (Rehurek & Sojka, 2010).

Table 1 is a comparison of different Topic Modeling Methods that shows a summary of the concept and advantages, and disadvantages of five topic modeling approaches (Hornick, 2017; Gabrilovich & Markovitch, 2006; Mazzola et al., 2018; Řehůřek, 2019; Rehurek & Sojka, 2010; Shi et al., 2018). As shown in Table 1, ESA and LSA are tightly connected, and in principle, ESA cannot outperform the peak performance of LSA. Therefore, ESA is not used as part of the comparison of different topic modeling algorithms for this study.

In this project, LSA, LDA, NMF, and HDP will be implemented, and the performance of each model will be compared. The four algorithms are selected because they are open-source resources and suitable for large documents.

Sentiment Analysis

Sentiment analysis is a kind of opinion mining that is usually performed on comments of users about an object or a topic (Ying et al., 2021; Malik et al., 2021; Keikhosrokiani & Asl, 2022) to determine its sentiment orientation or whether the comments of a selected

Table 1
Comparison of different topic modeling methods

Criteria / Method	Concept	Advantages	Disadvantages
Latent Semantic Analysis (LSA)	Exploits co-occurrence between terms to the project documents into a low-dimensional space. The inference is made using linear algebra Singular Value Decomposition (SVD).	Can be combined with another algorithm and can produce a relatively fast model It is fast and popular.	May not perform well when working with short documents. It is based on dimensional reduction of the original dataset, whereas the dimension factors' determination is subjective.
Latent Dirichlet Allocation (LDA)	A fully generative model is based on the bag of words paradigm, and word document counts. Documents are assumed to have been generated according to per-document topic distribution and per-topic word distribution.	Efficient when the corpus is large. Noise reduction is possible. It is popular, and there are many examples.	The topics discovered by using LDA are implicit and hard to interpret as they are defined only using their keywords, but not labels or abstract descriptions. Sometimes the keywords from different topics overlap and do not yield a proper topic name: the explanation is fuzzy.
Non-negative Matrix Factorization (NMF)	Discover topics by decomposing the document-term matrix into two low-rank factor matrices	Efficient when the corpus is large. Does not require a predefined number of topics.	May not perform well when working with short documents. Less interpretative as it lacks explicit probabilistic meaning of each factor.
Explicit Semantic Analysis (ESA)	Compute the "semantic relatedness" between the documents and humans' defined topics to improve text document categorization	It is a knowledge base; hence, it can be assigned with human-readable labels to the topic. ESA can discover relevant topics even when the topic is overlapping.	The previous ESA research used Wikipedia as a knowledge repository, more suitable for an expert system model. The conceptual motivation for ESA, recent work has observed unexpected behavior. ESA and LSA are tightly connected, and in principle, ESA cannot outperform the peak performance of LSA.
Hierarchical Dirichlet Process (HDP)	A nonparametric Bayesian approach to clustering grouped data	Does not require a predefined number of topics	The maximum number of topics can be unbounded and learned from the data rather than specified in advance. It is more complicated to implement and unnecessary when a bounded number of topics is acceptable.

topic are positive, negative, or neutral (Ding et al., 2008). According to Vinodhini and Chandrasekaran (2012), sentiment analysis can be conducted at three levels document level, sentence level, and attribute level. Two main approaches can be applied for sentiment analysis: semantic orientation and machine learning. Semantic orientation is unsupervised. Therefore, prior training in data is not needed. However, Machine learning refers to supervised learning and unsupervised learning. For supervised machine learning, prior training in data is essential (Vinodhini & Chandrasekaran, 2012). A hybrid machine learning method and semantic approach are also widely applied (Lodin & Balani, 2017).

Supervised Machine Learning methods can achieve 80–84% accuracy for Sentiment Analysis (Lodin & Balani, 2017). The popular Machine Learning methods in Sentiment Analysis are Support Vector Machine (SVM), Naïve Bayes, and Maximum Entropy (Abdelrahman & Keikhosrokiani, 2020; Teoh & Keikhosrokiani, 2020). SVM is one of the most popular machine learning algorithms as it is a high-performing algorithm with a little tuning. When there is a clear margin of distinction between classes, SVM performs well. In high-dimensional spaces, SVM is more effective. When the number of dimensions is more than the number of samples, SVM is more effective. SVM uses a small amount of memory. Machine Learning models have higher accuracy for sentiment analysis, but they also require more time to train the model. Thus, semantic analysis is more suitable for real-time applications (Vinodhini & Chandrasekaran, 2012). An unsupervised machine learning method applied for Sentiment Analysis is Clustering (Alaei et al., 2019).

In a study, Ding et al. (2008) proposed a lexicon-based method to use the opinion-bearing words to determine whether the comments are positive or negative. Opinion-bearing words are often used to express positive or negative opinions, for example, amazing and ugly. The algorithm counts the number of positive and negative opinions bearing words near the study's subject or topic. If there are more positive than negative sentiments, then the opinion on the topic is positive, or vice versa. However, this method is ineffective in context-dependent opinion-bearing words (Ding et al., 2008). Lexicon-based techniques are based on word dictionaries, where each word relates to a certain sentiment, and the overall sentiment is calculated. The lexicon-based methods are restricted by their lexicon, and the sentiment values assigned to the words in the dictionary neglect the context. In different contexts, the same adjective could have a different sentiment. Lexicon-based approaches are limited by their lexicons, specifically the static prior sentiment values of words or concepts in all circumstances. In order to overcome this limitation, various techniques have been proposed to study the semantics (Lodin & Balani, 2017). One of the solutions to overcome Lexicon-based techniques is to assign an updated sentiment strength to words in the lexicon. However, it still needs to be trained from manually annotated corpora. Another issue with lexicon-based techniques is that they rely entirely on the presence of phrases that express sentiment overtly, but the sentiment of a term is often implicitly reflected by the semantics of its context (Lodin & Balani, 2017).

The semantic analysis was introduced to improve the lexicon-based approach. Compared to the lexicon-based approach, the semantic analysis uses a dictionary of domain-specific terms, and the polarity of the terms is required (Alaei et al., 2019). There are two methods for semantic analysis: contextual semantic methods and conceptual semantic methods. Contextual semantic methods, also known as statistical semantics or corpus-based methods, determine the semantics based on the co-occurrence patterns of terms. Conceptual semantic methods, sometimes called dictionary-based approaches, use an external semantic knowledge base or a dictionary with the domain specified terms and sentiments attached to the words with NLP techniques to capture the sentiment. Examples include SenticNet, SentiStrenght, and WordNet (Ding et al., 2008; Lodin & Balani, 2017; Alaei et al., 2019). In some studies, a hybrid approach uses both machine learning and lexicon-based approaches in the same model. As a result, the accuracy is higher than the Naïve Bayes model (Appel et al., 2016; Lodin & Balani, 2017). Another study by Mumtaz and Ahuja (2018) proposed a model in which the accuracy of the hybrid system is found to be around 93%, which is higher than the pure lexical and SVM algorithm.

In this project, a lexicon-based or rule-based algorithm will be used for sentiment analysis as the data are in text form with no train and test set available to create a Machine Learning model. Preparing a labeled dataset from pure text data is time-consuming. Furthermore, the machine learning classifier is domain-based and needs to be retrained if future works use a dataset from a different domain. Hence, TextBlob and VADER libraries written in Python for rule-based sentiment analysis will be used. For sentiment analysis, three approaches are available, as shown in Table 2 (Devika et al., 2016; Ding et al., 2008; Lodin & Balani, 2017).

Table 3 summarizes the comparison of the related studies based on the data source, methods and tools, and their visualization.

MATERIALS AND METHODS

In order to achieve the main goal of this study, Cross-Industry Standard Process for Data Mining (CRISP-DM) Methodology is improvised as a new proposed text analytics lifecycle in this project. There are six stages in CRISP-DM Methodology: (1) Business Understanding, (2) Data Understanding and Preparation, (4) Modeling, (5) Evaluation, and (6) Deployment. For this project, data understanding and data preparation steps are combined. Then, two new steps are added: text processing and text exploration, while the deployment phase is not included in this project. Figure 1 shows the proposed digital model developed and adopted to detect and visualize themes in literary works.

The main steps of the proposed digital model are business understanding, data understanding and preparation, text processing, text exploration, modeling, and evaluation which are explained in detail. Business understanding emphasizes the specification of the

Table 2
Comparison of different sentiment analysis methods

Criteria \ Method	Machine Learning Approach	Semantic-Based Approach	Lexicon/Rule-Based Approach	Hybrid Approach
Classification	1. Supervised Learning 2. Unsupervised Learning	Unsupervised Learning	Unsupervised Learning	Combination of Machine learning and a lexicon-based approach
Advantages	1. Not necessary to have a dictionary 2. High accuracy of classification demonstrated 3. Can train with or without Semantic concept	1. Performance of sentiment classification at the sentence level is better than word level 2. Consider the context of the words 3. Does not need labeled data	1. Labeled data and the procedure of learning are not required 2. Fast	Higher accuracy
Disadvantages	1. Classifier trained are domain-based 2. Required significant time to train	1. Accuracy and efficiency depend on defining rules 2. Domain Specified	1. Requires powerful linguistic resources, which is a scarce resource 2. Context of the words is not considered 3. Relying on POS Tagger	1. Domain specified 2. Require time
Algorithm	1. Naïve Bayes 2. Support Vector Machine 3. K Nearest Neighbour	1. SenticNet 2. SentiStrenght	1. TextBlob 2. VADER	Any method from the previous approach

Table 3
Comparison of related studies

No	Author (Year)	Topic	Data Source	Method / Tools	Visualization
1	(Scharl et al., 2016)	Analyzing the public discourse on works of fiction – Detection and visualization of emotion in online overage about HBO's Game of Thrones	Website: Anglo-American News Media social media: Twitter, Facebook, Google+, YouTube	NOVEL Developed Westeros Sentinel, utilizes the webLyzard: web intelligence Sentiment analysis, topic modeling	Interactive dashboard shows the weightage of the occurrence of the events and the characters in the data, and the sentiment orientation.
2	(Costa, 2018)	A method for content analysis applied to newspaper coverage of Japanese personalities in Brazil and Portugal	Portuguese and Brazilian newspaper CHAVE corpus	Topic Modeling log-likelihood ratio to rank words according to their relative frequency differences in two corpora	Percentage of texts referring to other countries Percentage of texts by section Percentage of texts mentioning Japanese personalities Percentage of texts referring to Kurosawa, Oshima, and the Emperor Akihito Distribution by a week of texts referring to Hosokawa in the first semester of 1994
3	(Ding et al., 2008)	A Holistic Lexicon-Based Approach to Opinion Mining	Customer reviews of 8 products: two digital cameras, one DVD player, one MP3 player, two cellular phones, one router, and one anti-virus software	Opinion Observer (Proposed model) Sentiment Analysis	Accuracy of the model
4	(Paroubek & Pak, 2010)	Twitter as a Corpus for Sentiment Analysis and Opinion Mining	Twitter	Sentiment Classification (unigrams, bigrams, and trigrams)	The distribution of the word frequencies Impact of different parameters on the accuracy of the Sentiment Classifier

Table 3 *(continue)*

No	Author (Year)	Topic	Data Source	Method / Tools	Visualization
5	(Rehurek & Sojka, 2010)	Software Framework for Topic Modeling with Large Corpora	Mathematical papers from the Czech Digital Mathematics Library DML-CZ, from the NUMDAM repository, and the math part of arXiv	Latent Semantic Analysis Latent Dirichlet Allocation	None
6	(Vinodhini & Chandrasekaran, 2012)	Sentiment Analysis and Opinion Mining: A Survey	Movie review dataset Product review from amazon.com	Sentiment analysis Naïve Bayes Classifier	Performance of the sentiment classification model and sentiment analysis model.
7	(Grayson et al., 2017)	Exploring the Role of Gender in 19th-century Fiction Through the Lens of Word Embeddings	Forty-eight novels from twenty-nine 19th-century novelists sourced from Project Gutenberg,	word2vec t-Distributed Stochastic Neighbor Embedding (t-SNE)	Word frequencies for our initial list of gender-encoded words. The cosine similarity scores between female and male-authored words in our gender-encoded list.
8	(Grayson et al., 2016)	Novel2Vec: Characterizing 19th-Century Fiction via Word Embeddings	Twelve popular 19th-century novels were written by Jane Austen, Charles Dickens, and Arthur Conan Doyle.	Two variants of word2vec, a continuous bag-of-words strategy and a skip-gram strategy,	Word embedding visualization Context window sensitivity comparison
9	(Grayson et al., 2016)	The Sense and Sensibility of Different Sliding Windows in Constructing Co-occurrence Networks from Literature	A collection of nine novels from two 19th century British novelists— six by Jane Austen and three by Charles Dickens— sourced from Project Gutenberg	Applying different sliding window methodologies to capture character co-occurrences within the literature to build social networks	Social network of the characters in the novel
10	(Leavy et al., 2020)	Mitigating Gender Bias in Machine Learning Data Sets	A set of over 16,000 volumes of 19th-century fiction from the British Library Digital corpus	Sentiment analysis	Terms denoting emotion associated with men and women were extracted, and the levels of association

Table 3 *(continue)*

No	Author (Year)	Topic	Data Source	Method / Tools	Visualization
11	(Leavy et al., 2020)	Curatr: A Platform for Exploring and Curating Historical Text Corpora	35,918 English language fiction and non-fiction books dating from 1700 to 1899.	Curatr	Semantic network
12	(Leavy, 2019)	Curatr: A Platform for Semantic Analysis and Curation of Historical Literary Texts	16,426 works of fiction in the corpus	Curatr	None
13	(Suhendra et al., 2022)	Opinion Mining and Text Analytics of Literary Reader Responses	Goodreads review for KL Noir Volumes	Sentiment analysis & topic modeling	Book's rating, most salient terms
14	(Jafery et al., 2022)	Text Analytics Model to Identify the Connection Between Theme and Sentiment in Literary Works	Diasporic women from Iraq wrote six life writings were	Latent Dirichlet Allocation (LDA), sentiment analysis	Most salient terms, sentiments related to themes
15	(Al Mamun et al., 2022)	Sentiment Analysis of the Harry Potter Series Using a Lexicon-Based Approach	Harry Potter book series	Lexicon-based sentiment analysis	Sentiment frequency, sentiment analysis, AFINN sentiment dictionary, a sentiment of hero characters, a sentiment of houses
16	(Sofian et al., 2022)	Opinion Mining and Text Analytics of Reader Reviews of Yoko Ogawa's The Housekeeper and the Professor in Goodreads	Reader reviews of Yoko Ogawa's The Housekeeper and the Professor on Goodreads	Sentiment analysis & topic modeling	Book's rating, most salient terms, bigrams, trigrams

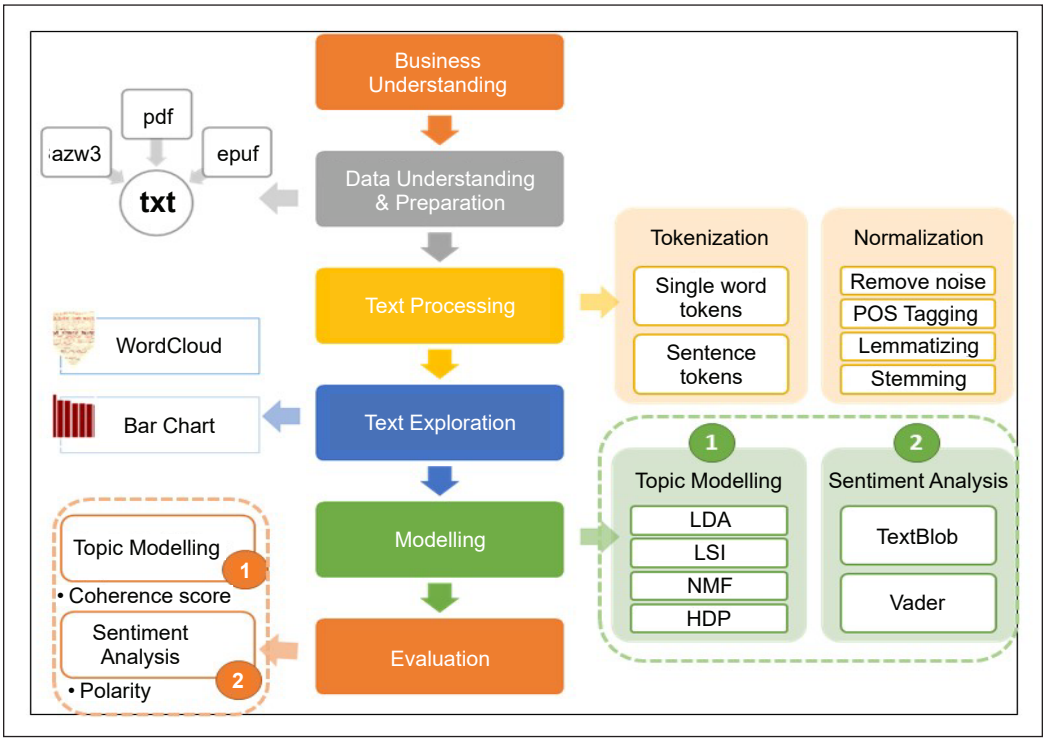


Figure 1. Proposed text analytics lifecycle

problem and methods of evaluating the achievement of the goal. The data understanding and preparation step is used to discover and reformat the secondary data, which are 28 literary books selected for this project. In the text processing step, tokenization and normalization are applied. The modeling step consists of two main models: (1) topic modeling and (2) sentiment analysis. First, topic modeling is used to develop a model with selected algorithms of LDA, LSI, NMF, and HDP that can detect the topics related to imprisonment from the 28 Iranian diasporic life writings. Then, the results of different topic modeling algorithms are compared based on coherence scores. The second part of the modeling step focuses on sentiment analysis to create a model that can analyze the polarity of diasporic writers' discourse on imprisonment with the selected algorithms of Vader and TextBlob. Finally, the sentiment analysis models are evaluated by comparing the polarity scores. Each step of the proposed digital model is explained in detail in the following sections.

Data Understanding and Preparation

The data used in this study are secondary data from a corpus of 28 books that cover contemporary life writings by the Iranian diaspora, as shown in Table 4. All the works were published after the 1979 Islamic revolution and revolved around similar subject matters. The works are temporally, spatially, and thematically related. Together, they serve as a text

corpus for the aims of this study. Since the dataset is not labeled, unsupervised learning, particularly the lexicon/rule-based approach, is the best method (Ding et al., 2008; Lodin & Balani, 2017).

Table 4
List of Literary works used in this project

No	Book Name
1	Fatemeh Keshavarz–Jasmine and Stars_Reading More Than Lolita in Tehran -The University of North Carolina Press (2007).pdf
2	Nemat, Marina–Prisoner of Tehran_One Woman's Story of Survival Inside an Iranian Prison.pdf
3	Andalibian, Rahimeh–The Rose Hotel_A Memoir of Secrets, Loss, and Love from Iran to America.pdf
4	Anita Amirrezvani–The Blood of Flowers-Back Bay Books (2008).pdf
5	Ansary, Nina–Jewels of Allah _the untold story of women in Iran-Revela Press (2015) .pdf
6	Azadeh Moaveni–Guest House for Young Widows_Among the Women of ISIS-Random House (2019).pdf
7	Azar Nafisi–Reading Lolita in Tehran_A Memoir in Books-Random House (2003).pdf
8	Azar Nafisi–Things I've Been Silent About_Memories of a Prodigal Daughter-Random House (2008).pdf
9	Basmenji, Kaveh–Afsaneh_Short Stories by Iranian Women.pdf
10	Bijan, Donia–Maman's Homesick Pie A Persian Heart in an American Kitchen.pdf
11	Dalia Sofer–The Septembers of Shiraz (2007).pdf
12	Dina Nayeri –The Ungrateful Refugee-Canongate Books (30 May 2019).pdf
13	Ebadi, Shirin–Until we are free _ my fight for human rights in Iran-Random House (2016).pdf
14	Entekhabifard, Camelia–Camelia_ Save Yourself by Telling the Truth_ A Memoir of Iran.pdf
15	Esfandiari, Haleh–My Prison, My Home_One Woman's Story of Captivity in Iran.pdf
16	Firoozeh Dumas–Funny in Farsi_ A Memoir of Growing Up Iranian in America-Villard (2003).pdf
17	Firoozeh Dumas–Laughing without an accent_adventures of an Iranian American, at home and abroad -Villard (2008).pdf
18	Gohar Homayounpour–Doing Psychoanalysis in Tehran-The MIT Press (2012).pdf
19	Goldin, Farideh–Wedding Song_ Memoirs of an Iranian Jewish Woman.pdf
20	Nafisi, Azar–The Republic of Imagination_ America in Three Books.pdf
21	Nouraie-Simone, Fereshteh_ Farrokh, Faridoun_ Khalili, Sara–The Shipwrecked_ Contemporary Stories by Women from Iran.pdf
22	Rachlin Nahid–Persian-Girls_-A-Memoir.pdf
23	Rostampour, Maryam–Captive in Iran.pdf
24	Roxana Saberi–Between Two Worlds_ My Life and Captivity in Iran -Harper Perennial (2011).pdf
25	Shahla Talebi–Ghosts of Revolution_ Rekindled Memories of Imprisonment in Iran -Stanford University Press (2011).pdf
26	Shirin Ebadi_Rich, Nathaniel–The golden cage _ three brothers, three choices, one destiny-Kales Press (2011).pdf
27	Zanjani, Sohila_ Brewster, David–Scattered Pearls_ Three generations of Iranian women and their search for freedom.pdf
28	Zarah_Ghahramani–My-Life-as-a-Traitor_-An-Iranian-Memoir.pdf

Text Processing

Text processing consists of two main phases: tokenization and normalization (Mayo, 2017). Tokenization is splitting longer text strings into smaller pieces or tokens (Subramanian, 2019). Normalization refers to converting numbers to their word equivalents, removing punctuation, converting all text to the same case, and lemmatization, which means returning the word to its base form. Tagging is the second step after tokenization in the typical NLP pipeline (Bird et al., 2009). Part-of-speech (POS) tagging is the process of classifying and labeling words into their parts-of-speech. Then, *tagset* refers to the collection of tags used for a task. After POS Tagging, only nouns and adjectives remain for the following process as other word tags like pronouns and determiners do not carry any insight.

Text Exploration

For an expository text, text exploration is when the text is examined critically and tries to find and prioritize the main ideas or facts, or the essential event and characters, in the case of a narrative composition. As the original text contains many topics, only the sentences which contained any keyword from the list of 324 related keywords to the topic of imprisonment are extracted. About 19 667 sentences are extracted out of the 137 319 sentences in the corpus. After extracting the sentences, the data cleaning processes of tokenization, removing stop words and irrelevant values, POS Tagging, and lemmatization are performed on the extracted text for a second time. Bar charts of top frequent words, Bigrams and Trigrams WordCloud, are created based on the cleaned text using python. These visualizations are used as descriptive data to give a general idea about the dataset. Figure 2 shows 10 frequent words, among which *woman* is the most frequent term while the *law* is the least frequent one. Within the extracted text, nouns such as woman and time

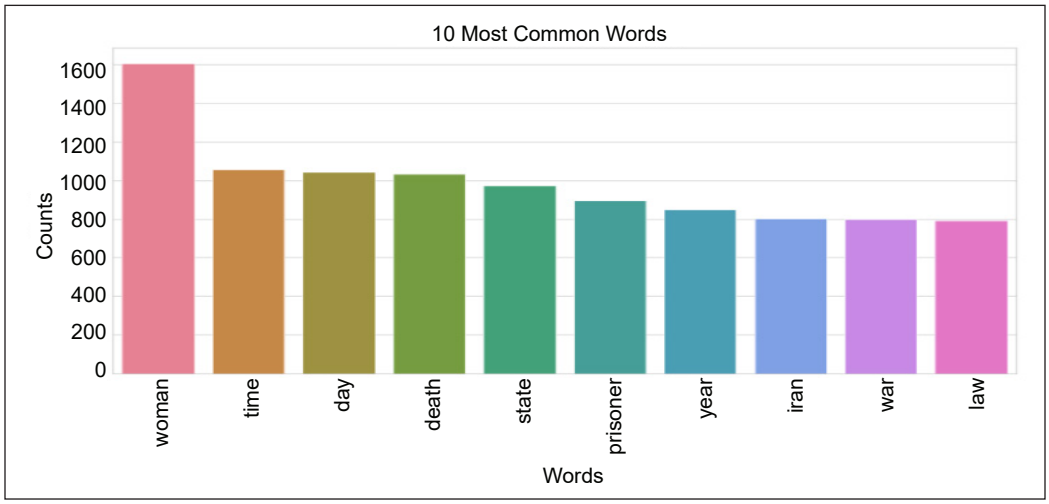


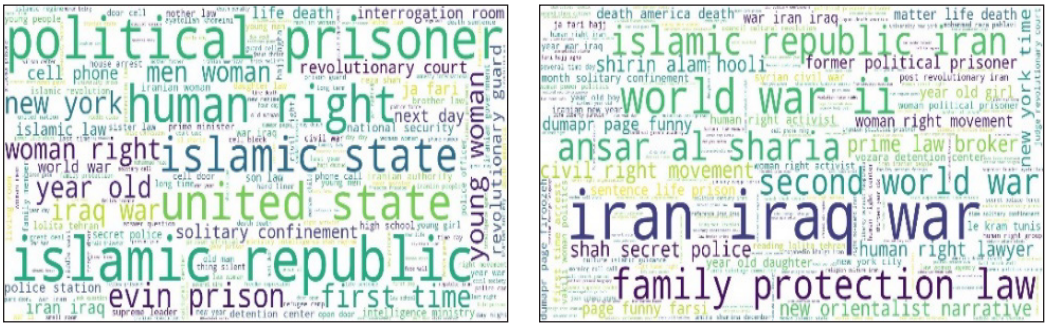
Figure 2. Bar chart of top 10 frequent terms

are still the main discussion subject, accompanied by other terms such as state, prisoner, and death. The most frequent adjectives also appear semantically neutral, leaning slightly towards positivity.

The WordClouds in Figure 3 shows the most frequently occurring bigrams and trigrams. An expert interprets the WordClouds in Iranian diasporic literary and cultural studies. Based on the WordClouds and experts' opinions, the following assumption can be made about the topics:

1. The prisoners are criminals, political prisoners, or war prisoners in Evin Prison.
2. Women demand their rights and family protection laws related to the hypothesis made with WordCloud of Unigram that women seek a new and better life.
3. The prisoners demand their rights, such as human rights and civil rights.

Similar to topic modeling, text exploration is done before sentiment analysis. In this phase, the text is first tokenized, then nouns and adjectives are extracted to create WordClouds [Figures 4(a) and 4(b)]. The WordCloud of adjectives shows that the dominant sentiment is rather positive because of the high frequency of terms like *new* and *good* and the neutrality of other adjectives like *Iranian* that indicate nationality. Here, a hypothesis is suggested that Iranian women are looking forward to having a new and better life.



(a) (b)

Figure 3. WordClouds: (a) Most frequent bigrams; (b) Most frequent trigrams

(a) (b)

Figure 4. WordClouds*:* (a) WordCloud of nouns; (b) WordCloud of adjectives

Topic Modeling

Two models are built in this project for sentiment analysis and topic modeling. The process of topic modeling is shown in Figure 5.

The first step of developing the model is to generate a dictionary and a corpus from the text preprocessed in the previous sub-section. Here, the dictionary refers to id2word, which means giving an ID to every unique word in the text. Corpus refers to the term-document frequency, where the frequency of the term in the document is calculated. The corpus only contains the ID of the word and its frequency. By passing the dictionary, the ID can be changed to the word. The example of the corpus is shown in Figure 6.

In this project, four algorithms, namely, LDA, LSI, NMF, and HDP, are used, and their performance is compared in topic modeling. LDA, LSI, and NMF require the number of topics defined to build the model. The number of topics representing the number of keyword lists is generated from the model. The performance of topic modeling models is evaluated based on coherence score. The performance is better when the coherence score is higher. In order to obtain the number of topics that yield the optimal coherence score, the model is iterated from 10 topics to 30 topics. The reason to start from 10 topics rather than from one is to ensure the sufficiency of the generated topics to cover all discussions in the corpus.

On the other hand, to prevent generating too many topics, the iteration is stopped at 30. After the iteration is completed, the number of topics with the highest coherence score is retrieved, and the generated topics are saved for analysis. Figures 7(a) to 7(c) show the chart of coherence score of the models from 10 topics to 30 topics. An Iranian diasporic literary and cultural studies expert manually evaluates the polarity accuracy. Based on the line chart of the coherence score of each model, each algorithm has different changes when the number of topics is increased. LDA models' coherence score showed an increasing trend when the number of topics increased. For LSI and NMF models, the performance decreased when

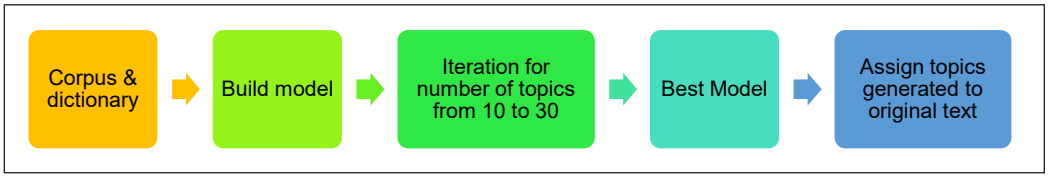
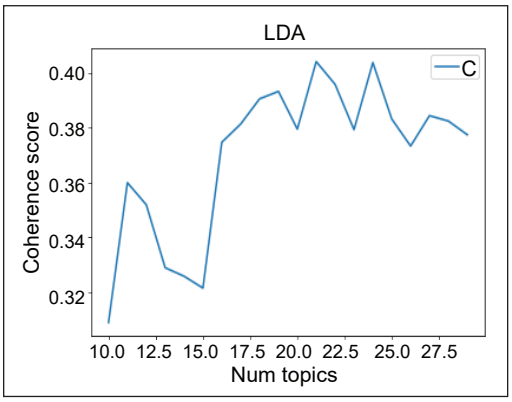


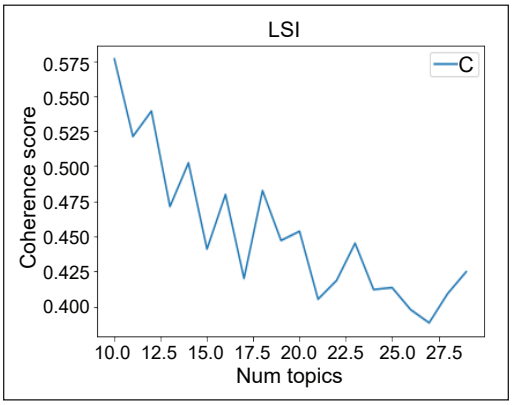
Figure 5. Topic modeling process

```
Corpus:  [[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1),
(13, 1), (14, 1), (15, 1), (16, 1), (17, 1), (18, 1), (19, 1), (20, 1), (21, 2), (22, 1), (23, 1), (24, 2), (25, 1),
(26, 2)]]
Corpus with original text:  [[('asmine', 1), ('ava', 1), ('bruce', 1), ('bryant', 1), ('carolina', 1), ('chapel', 1),
('civilization', 1), ('editor', 1), ('ernst', 1), ('galliard', 1), ('hill', 1), ('inc', 1), ('indd', 1), ('islamic',
1), ('jasmine', 1), ('keshavarz', 1), ('keystone', 1), ('lawrence', 1), ('lolita', 1), ('network', 1), ('north', 1),
('press', 2), ('right', 1), ('samp', 1), ('star', 2), ('tehran', 1), ('university', 2)]]
```

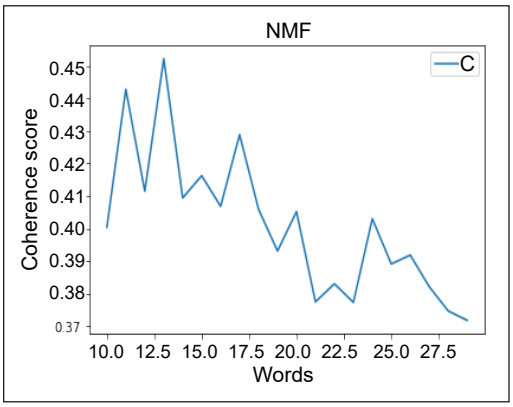
Figure 6. Example of corpus



(a)



(b)



(c)

Figure 7. Line Chart of Coherence Score: (a) Line Chart of Coherence Score for LDA Model, Best Coherence Score: 0.404, Number of Topics: 21; (b) Line Chart of Coherence Score for LSI Model, Best Coherence Score: 0.576, Number of Topics: 10; (c) Line Chart of Coherence Score for NMF Model, Best Coherence Score: 0.452, Number of Topics: 13.

the number of topics increased. HDP model does not require several topics; hence, the model is built directly. After generating the list of keywords, the keywords can be used to assign the most relevant topic for each sentence in the text.

Sentiment Analysis

This project uses two python libraries, Vader and TextBlob, to calculate the corpus' sentiment based on a rule-based approach. The development of Sentiment Analysis is more straightforward than Topic Modeling. The only step needed is to pass the sentences to the analyzer, which will return with the polarity of the sentences, as shown in Figure 8.

In Vader libraries, the analyzer calculates the score of negativity, neutrality, and positivity of the sentences and calculates the compound score based on the earlier scores. While in Text blob, the analyzer calculates the polarity and subjectivity of the sentences. Both compound score and polarity range from -1.0 to +1.0, while subjectivity ranges from 0 to 1, where 0 means the sentence is very objective, and one indicates that the sentence is very subjective. Like Topic Modeling, only the extracted text is used in Sentiment Analysis. The text data used in this project is extracted purely from the discourse; hence there are no labels on the polarity for each sentence. Instead, a sample is extracted from the model's output to evaluate the accuracy of the polarity calculated by the models. The polarity of each piece of text in the sample is manually identified to verify if the model performs as anticipated.

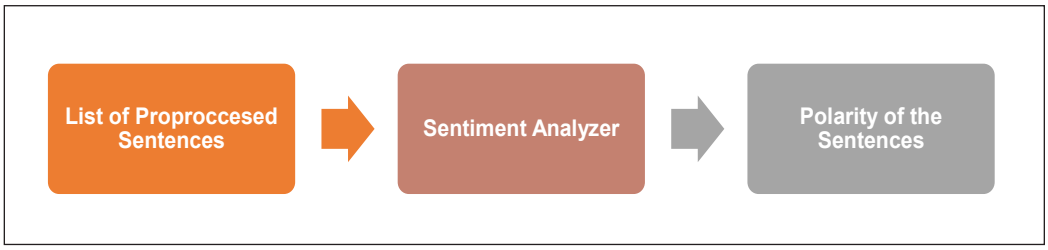


Figure 8. Process of sentiment analysis

Evaluation

The model's performance is measured with a coherence score for Topic Modeling. The coherence score of each model is calculated and compared. The higher the coherence score, the better the model. For Sentiment Analysis, the model's performance is analyzed by comparing the polarity (TextBlob Model) and Compound Score (Vader) with the manually assigned polarity to the text sample. The results from the model are also compared with the related literature.

RESULTS

Objective 1: Topic Modeling

As shown in Table 5, the best model for topic modeling is LSI, followed by HDP, NMF, and LDA. In LSI, the keywords of the 10 topics are repeated, but they can still be identified as different from each other by retrieving part of the text that falls on a specific topic. Six themes are derived from the 10 groups of keywords and their related texts: (1) sexism, (2) family, (3) revolution, (4) imprisonment, (5) intellectual, and (6) death. Based on the line chart of the coherence score of each model, each algorithm shows different changes when the number of topics varies. LDA models' coherence score is higher when the number of topics grows. For LSI and NMF models, the performance declines when the number of topics drops. As HDP does not need to determine the number of topics, this model is formed directly. After generating the list of keywords, the keywords are used to assign the most relevant topic for each sentence in the text.

Table 5
Results of topic modeling

Algorithm	Number of Topics	Coherence Score
LDA	21	0.404
LSI	10	0.576
NMF	13	0.452
HDP	-	0.496

Objective 2: Sentiment Analysis

Two python libraries have been used to generate the sentiment of the corpus. Comparing a sample from the results obtained from each library reveals that Vader generates more meaningful results than others, and the classification of the sentiments is more accurate than others. The top 10 positive sentences from the corpus are shown in Tables 6 and 7 using Vader and TextBlob. The positivity comes from appreciation, freedom, intellectual companion, power, wealth, and glory based on the sentences. Even though the extracted sentences are related to the notion of imprisonment, the detected positive sentiments are

Table 6
Top 10 positive sentences generated by Vader

Top Positive Vader		
Score Compound	**Sentence**	**Justification (Theme)**
0.9903	but i have grown wiser and more appreciative not only of the material comforts i unthinkingly enjoy every day a leisurely cup of coffee a moment in the sunlight the reassuring touch of shauls hand on mine but of the freedom with which i am blessed	Appreciate to have freedom
0.985	many years ago my mother asked me if i knew why it was a great deal more important to be happy than to be rich famous and beautiful because as schopenhauer informs us if we are telling a friend about a very rich famous and attractive person the first question we have to answer is but is he or she happy	Happiness is more important than rich, famous and beauty
0.9847	to the memory of our dear friend shirin alam hooli whose courage kindness and love live on in the hearts of all who knew her to the precious women who were with us in evin during our imprisonment some of whom have since been released and to all the women in evin today still waiting for the justice that only a free nation can give them	Memory in prison
0.9795	one can see why babbitt would be both attracted to the joys of freedom and frightened by its perils for freedom does have many perils and the best way to confront them is not to avoid being free but to cultivate independence of thought the kind of freedom that incidentally has been the great engine of american creativity and vitality in all fields from engineering to literature	Freedom is the source of creativity and vitality
0.9765	but of course there are all different kinds of freedom and the kind that is most precious you will not hear much talked about in the great outside world of wanting and achieving and displaying	Freedom
0.9719	but my heart was so filled with love and our shared experience that i felt him inside and tried to convince myself he was still alive	Love
0.9689	each one of us was honored with a special role in farahs life and everyone was more than his or her assigned role mahnaz was more than a sister neda more than a daughter nema more than a son hamid more than a brother jaleh more than a former comrade and best friend roshanak more than a former sisterinlaw bahram more than an intellectual companion and within that exclusive list i was left with the role of more than a childhood friend	Intellectual companion

Table 6 *(continue)*

Top Positive Vader		
Score Compound	**Sentence**	**Justification (Theme)**
0.9676	but it is also about wealth its great attraction as well as its destructive power the carelessness that comes with it and yes it is about the american dream a dream of power and wealth the beguiling light of daisys house and the port of entry to america	Power and Wealth
0.9666	cent warriors gained ready to give her rudabeh who is how shabby all when devoid of the love for which one of the many important who perform feats is women all norms of his own shahnameh her son are the war and win glory is for their their role as moth a different kind of courage as men they love	Glory
0.9635	before he is fully awake that alarm clock is described in great detail early in the novel we are invited to recognize that this all american businessman a defender of individualism and free trade is best defined not by any peculiarity of temperament or cherished keepsake but by his ownership of the best of the nationally advertised and quantitatively produced alarm clocks with all modern attachments making its owner proud of being awakened by such a rich device	Proud to be awake because of individualism and free trade

Table 7
Top 10 positive sentences generated by TextBlob

Top Positive TextBlob		
Polarity	**Subjectivity**	**Sentence**
1	1	she has always seemed perfect serving others having nothing but goodness in her heart
1	1	he read them with authority as if they were his as if he had composed every word like a perfect melody
1	0.3	i gave the best speech of my life there in that courtroom
1	0.3	maryam lived in one of the best an octagonal building known as eight heavens
1	0.3	60 farzanehs exceedingly unbiased viewpoint may best be exemplified in the philosophy of abbasgholizadeh we know that secular women do not share our convictions but this does not give us any problems since we are all working to promote the status of women
1	1	walid had listened to their arguments as a teenager on cassette tapes and videos impressed by their plans to bring justice to society
1	0.3	after school we were taken minas house where layla who we were entertained did their best to divert us
1	1	being i ate to hear her talk it would seem my greatest fault has been not at deaths of ourselves start door
1	1	the prosecu of my father with his greatest detrac seyyed mehdi pirasteh the minister of the tor on bail
1	1	azarmis voice could be heard God bless this wonderful motherinlaw

related to certain themes such as freedom, love, and wealth that carry positive connotations and sentiments. The results show that book authors celebrate their freedom after moving to the west. Freedom has brought happiness, and in the eyes of some of the authors, this happiness is more important than wealth, fame, and beauty. The results reveal that freedom is the source of creativity and vitality.

Based on the top 10 negative sentences shown in Tables 8 and 9 obtained from Vader and textBlob, respectively, the dominant feeling of the Iranians is against torture, which they have experienced during the war, in prison, in the house, or in being raped. It is because women in Iran do not have equal social and political rights (Asl, 2019, 2020, 2021). In addition, the experiences of mental and physical torture have traumatized the victims.

In summarizing the Tables 8 and 9 above, it can be said that Iranian diasporic writers express negative opinions about life in prison. Here, imprisonment not only refers to

Table 8
Top 10 negative sentences generated by Vader

Top Negative Vader		
Score Compound	**Sentence**	**Justification (Theme)**
-0.9869	but then others who witness my death or hear about it will know that i died because i refused to give in to hatred and violence and theyll remember and maybe someday theyll find a peaceful way of defeating evil	Last wish of a man who dying soon
-0.9833	she was scared of the snow and disease and loneliness and closed doors and the sulking of her son in law but she wasnt scared of death provided that she wouldnt suffer any pain that is provided that she wouldnt notice that she was dying provided that she died in her sleep	Torture by son in law
-0.9826	emerging from prison opening the enormous metal gate the guard suddenly took away my blindfold and asked me tauntingly if i would recognize my parents	Torture in the prison
-0.9816	as they navigate observing the lonesomeness of the river they are constantly threatened by the danger and violence that emanate like poisonous fumes from the land and its smothery houses the feuding between the seemingly civilized and church going grangerfords and shepherdsons the coldblooded and open killing of a helpless drunk seething mob anger the tarring and feathering of the duke and the dauphin	Murder
-0.9801	did he starve to death or die from lack of water as had imam hussein the third imam of shii muslims who in 681 was killed by yazid the caliph along with seventytwo of his companions because the enemy had deprived them of water for whose thirst and death yousuf used to cry so sincerely	Torture by enemy
-0.9738	the isis chroniclers remained obsessed with religion vividly portraying every atrocity the group committed as some facet of islam rape that longtime tactic of war used to humiliate the enemy became a theology of rape sickness rather than a war crime	Rape was used to humiliate the enemy

Table 8 *(continue)*

Top Negative Vader		
Score Compound	**Sentence**	**Justification (Theme)**
-0.9738	the torturers were taking turns, but he was being interminably beaten either hung from the ceiling so that his feet could not touch the floor and his arms were stretching to the point of being torn from the joints or on the torture bed or on the floor where all of them would attack him as if he were a dangerous animal	Torture
-0.9733	the argument in response often went like this such brutality was certainly not desirable but the west had left the militants no choice there was no other way left to resist nonviolent protest would not sway the dictator assad whose military was torturing and killing scores in detention centers nor would it sway the united states which had invaded and occupied iraq killed countless civilians and sustained and protected arab tyrants	Torture
-0.9727	there was no denying prerevolution corruption or abuse of power by the former regime but she knew that the mullahs brutality in the name of islam far surpassed that of the shahs secret police that they would suppress women gag them under their wretched veils deny them equal rights and put children their softest target in the line of fire in an ideological war	Women got no equal right
-0.9724	this is a problem not just for the uneducated but for trauma victims who have gaps in memory and rape victims who are ashamed	Trauma of victims

Table 9
Top 10 negative sentences generated by TextBlob

Top Negative TextBlob		
Polarity	**Subjectivity**	**Sentence**
-1	1	in this case ignorance is not the worst problem either
-1	1	farrukhlaqa asks why if this woman is insane has she not been taken to an asylum
-1	1	we are in a war against evil
-1	1	the records of that terrible war are not classied information
-1	1	i drove people insane with questions i am told
-1	1	something terrible is happening in this country i can smell it in the air and it smells of blood and disaster
-1	1	we have to protect islam gods law and gods people from the evil forces that are at work against them
-1	1	i had tried to accept my situation and to understand him but i couldnt pretend i didnt know about the horrible things he had done
-1	1	youve been up for two days this is insane
-1	1	at the ceremony the scent of jasmine casablanca lilies dahlias and gardenias at the sofreh mingled with the odor of esfand seeds burned to ward o the evil eye

jails and incarceration centers but also the constraining conditions of life within family, community, and religion. For men, the feeling of imprisonment mostly emanates from their experiences in war prisons and punitive facilities, but for women, the feeling mainly originates from a bad family milieu or community injustice where the law does not protect them nor are they granted equal rights (Asl, 2019; Hadi & Asl, 2022). War prisoners suffer from similar feelings of trauma as the physical torture by the enemy has increased the psychological feelings of hopelessness in the captives. Overall, most of the sentences are negative from the extracted text, followed by neutral and positive ones, as shown in Figure 9.

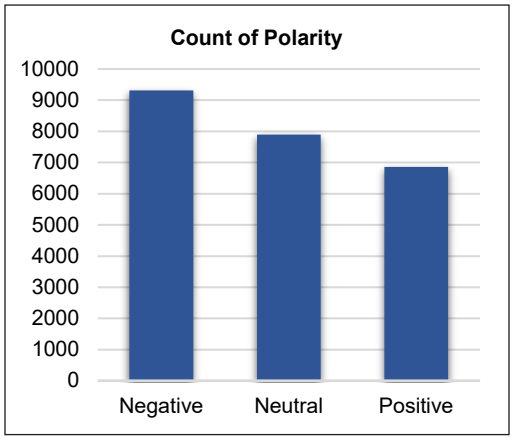


Figure 9. Sentiment bar chart

DISCUSSION

This project analyzed the underlying emotion toward the theme of imprisonment in a corpus of 28 life writings by Iranian diasporic writers. The sentiment attached to the theme has been obtained using sentiment analysis. Two sentiment analyzers from Python libraries have been used: VADER and TextBlob. The output of VADER is more insightful and interpretable. The main contribution to positive sentiment is the feeling of freedom. After the diasporic subjects leave their country of origin, which is struck by a devastating war and an increasing rate of crime and social discrimination, they experience a true sense of freedom that brings joy and happiness. On the other hand, the negative sentiments come from the destructions of the revolution and the eight-year war during which many were captured and tortured as prisoners, as well as the social-political inequalities that deny female citizens of equal rights and protection against patriarchal injustice.

From Topic Modeling, six interrelated themes have been derived from the 10 groups of keywords: sexism**,** family**,** revolution**,** imprisonment, intellectual, and death. A Sentiment Analysis revealed the underlying sentiment of the writers toward each of the themes. Since the 1979 revolution in Iran, the policies and practices of moral purification have generated a new sexual economy that has denied women certain social and political rights. The new sexism, for instance, has restricted women's presence and mobility in public places (Nekai, 2013; Shahrokni, 2020). The exclusionary policies have similarly reinforced the patriarchal power within the family structure to control women. Most women show their discontent with a family structure in which men are in charge of women by both marriage laws and what religious principles indoctrinate (Afary, 2009; Asl, 2018; Asl, 2021). The demand for

women's obedience, modesty, and self-sacrifice concerning their male guardians makes the family structure and milieu restricting and prison-like. The third theme revolves around the revolution and the subsequent individual and collective feelings of trauma. As a result of the revolution, the diasporic Iranian population has suffered painful losses: "the loss of family and friends, the loss of economic and social status, and the loss of their home country" (Naghibi, 2016, p. 4). While the revolution has been formative for people, it has been destructive for some others, as the losses brought about by the revolution have been completely disruptive and upsetting (Vasapollo, 2020). The role of silence and censorship as imposed practices by the state similarly denies intellectuals their right and will to voice out against injustice. Such policies have generated various forms of discontent, hatred, miseries, and despair about the incarcerating socio-political realities (Ranucci, 2019). This strong sense of entrapment and imprisonment is closely tied with the formation and expression of negative sentiments. Finally, the imminent threat of death and annihilation exerts a strong grip on the bodies and minds of Iranians. For a repressive society that seeks collective silencing of its nation, corporeal elimination of individuals is a common practice (Chiaramonte, 2013). Comparing the results of this project with related ethnographical literature makes it clear that the six themes identified in the selected 28 literary writings of diasporic Iranians evoke undesired negative sentiments.

CONCLUSION

This study aimed to propose a text analytics lifecycle for detecting and visualizing the prevailing themes in a corpus of literary texts. As shown in Figure 1, the model was developed in two stages. First, topic modeling techniques with selected algorithms of LDA, LSI, NMF, and HDP are applied to detect the main topics related to the theme of imprisonment in a corpus of 28 Iranian diasporic life writings. LSI is the best algorithm for this study's proposed Topic Modeling model based on the coherence score. By implementing Topic Modeling, six major sub-themes have been derived: sexism, family, revolution, imprisonment, intellectual, and death. Second, sentiment analysis is applied to analyzing the polarity of the diasporic writer's discourse on imprisonment with the selected algorithms. The feeling attached to each of the six topical themes has been visualized by adopting a Sentiment Analysis model. Specifically, two sentiment analyzers from Python libraries, VADER and TextBlob, were adopted (Figure 1). The main contribution to positive sentiment is the feeling of liberation and freedom. On the other hand, the negative sentiments are expressed and directed toward the 1979 revolution, the Iran-Iraq war, and the existing socio-political discrimination and injustice for the disadvantage of women. It is further concluded that the output of VADER is more insightful and interpretable.

The four algorithms compared in this paper can be utilized with different datasets to determine the best model for future works. Besides, this project focused only on published

writings from diasporic Iranian writers. In future studies, text data from social media related to the thematic topics discovered in this project can be the object of analysis. Furthermore, machine learning and deep learning techniques can be utilized to classify author opinions in these books further.

ACKNOWLEDGEMENT

The authors are thankful to School of Computer Sciences and Division of Research & Innovation, Universiti Sains Malaysia for the support from Short Term Grant (304/PKOMP/6315435), which is granted to Ts.Dr. Pantea Keikhosrokiani.

REFERENCES

Abdelrahman, O., & Keikhosrokiani, P. (2020). Assembly line anomaly detection and root cause analysis using machine learning. *IEEE Access, 8*, 189661-189672. https://doi.org/10.1109/ACCESS.2020.3029826

Afary, J. (2009). *Sexual politics in modern Iran*. Cambridge University Press.

Al Mamun, M. H., Keikhosrokiani, P., Asl, M. P., Anuar, N. A. N., Hadi, N. H. A., & Humida, T. (2022). Sentiment analysis of the Harry Potter Series using a lexicon-based approach. In P. Keikhosrokiani & M. Pourya Asl (Eds.), *Handbook of Research on Opinion Mining and Text Analytics on Literary Works and Social Media* (pp. 263-291). IGI Global. https://doi.org/10.4018/978-1-7998-9594-7.ch011

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: Capitalizing on big data. *Journal of Travel Research, 58*(2), 175-191. https://doi.org/10.1177/0047287517747753

Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems, 108*, 110-124. https://doi.org/https://doi.org/10.1016/j.knosys.2016.05.040

Asl, M. P. (2018). Practices of counter-conduct as a mode of resistance in Middle East women's life writings. *3L: Language, Linguistics, Literature®, 24*(2), 195-205. https://doi.10.17576/3L-2018-2402-15

Asl, M. P. (2019). Foucauldian rituals of justice and conduct in Zainab Salbi's between two worlds. *Journal of Contemporary Iraq & the Arab World, 13*(2-3), 227-242. https://doi.10.1386/jciaw_00010_1

Asl, M. P. (2020). Micro-Physics of discipline: Spaces of the self in Middle Eastern women life writings. *International Journal of Arabic-English Studies, 20*(2), 223-240. https://doi.10.33806/ijaes2000.20.2.12

Asl, M. P. (2021). Gender, space and counter-conduct: Iranian women's heterotopic imaginations in Ramita Navai's City of Lies. *Gender, Place & Culture*, 1-21. https://doi:10.1080/0966369X.2021.1975100

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python.* O'Reilly Media, Inc.

Chiaramonte, P. (2013, January 12). *Hell on earth: Inside Iran's brutal Evin prison*. Fox News. https://www.foxnews.com/world/hell-on-earth-inside-irans-brutal-evin-prison

Costa, L. F. (2018). A method for content analysis applied to newspaper coverage of Japanese personalities in Brazil and Portugal. *Digital Scholarship in the Humanities, 33*(2), 231-247. https://doi.org/10.1093/llc/fqx050

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: A comparative study on different approaches. *Procedia Computer Science, 87*, 44-49. https://doi.org/https://doi.org/10.1016/j.procs.2016.05.124

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 231-240). ACM Publishing. https://doi.org/10.1145/1341531.1341561

Firmin, R. L., Bonfils, K. A., Luther, L., Minor, K. S., & Salyers, M. P. (2017). Using text-analysis computer software and thematic analysis on the same qualitative data: A case example. *Qualitative Psychology, 4*(3), 201-210. https://doi.org/10.1037/qup0000050

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *AAAI* (Vol. 6, pp. 1301-1306). American Association for Artificial Intelligence.

Graneheim, U. H., Lindgren, B. M., & Lundman, B. (2017). Methodological challenges in qualitative content analysis: A discussion paper. *Nurse Education Today, 56*, 29-34. https://doi.org/10.1016/j.nedt.2017.06.002

Grayson, S., Mulvany, M., Wade, K., Meaney, G., & Greene, D. (2016, September 20-21). Novel2vec: Characterising 19th century fiction via word embeddings. In *24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16)*. Dublin, Ireland.

Grayson, S., Mulvany, M., Wade, K., Meaney, G., & Greene, D. (2017). Exploring the role of gender in 19th century fiction through the lens of word embeddings. In J. Gracia, F. Bond, J. McCrae, P. Buitelaar, C. Chiarcos & S. Hellmann (Eds.), *Language, Data and Knowledge*, (pp. 358-364). Springer. https://doi.org/10.1007/978-3-319-59888-8_30

Grayson, S., Wade, K., Meaney, G., & Greene, D. (2016). The sense and sensibility of different sliding windows in constructing co-occurrence networks from literature. In B. Bozic, G. Mendel-Gleason, C. Debruyne & D. O'Sullivan (Eds.), *Computational History and Data-Driven Humanities*, (pp. 65-77). Springer.

Hadi, N. H. A., & Asl, M. P. (2022). The real, the imaginary, and the symbolic: A Lacanian reading of Ramita Navai's City of Lies. *GEMA Online Journal of Language Studies*, *22*(1), 145-158. https://doi.org/10.17576/gema-2022-2201-08

Hornick, M. (2017, November 17). Explicit semantic analysis (ESA) for text analytics. *Oracle Machine Learning.* https://blogs.oracle.com/r/explicit-semantic-analysis-esa-for-text-analytics

Jafery, N. N., Keikhosrokiani, P., & Asl, M. P. (2022). Text analytics model to identify the connection between theme and sentiment in literary works: A case study of Iraqi life writings. In P. Keikhosrokiani & M. P. Asl (Eds.), *Handbook of research on opinion mining and text analytics on literary works and social media* (pp. 173-190). IGI Global. https://doi.org/10.4018/978-1-7998-9594-7.ch008

Keikhosrokiani, P., & Asl, M. P. (Eds.). (2022). *Handbook of Research on Opinion Mining and Text Analytics on Literary Works and Social Media*. IGI Global. https://doi.org/10.4018/978-1-7998-9594-7.

Leavy, S., Meaney, G., Wade, K., & & Greene, D. (2019). Curatr: A platform for semantic analysis and curation of historical literary texts. In E. Garoufallou, F. Fallucchi & E. W. De Luca (Eds.), *Metadata and Semantics Research*, (pp. 354-366). Springer. https://doi.org/10.1007/978-3-030-36599-8_31

Leavy, S., Meaney, G., Wade, K., & & Greene, D. (2020). Mitigating gender bias in machine learning data sets. In L. Boratto, S. Faralli, M. Marras & G. Stilo (Eds.), *Bias and Social Aspects in Search and Recommendation* (pp. 12-26). Springer. https://doi.org/10.1007/978-3-030-52485-2_2

Lodin, H., & Balani, P. (2017). Rich semantic sentiment analysis using lexicon based approach. *ICTACT Journal on Soft Computing, 7*(04), 1486-1491. https://doi.org/10.21917/ijsc.2017.0206

Malik, E. F., Keikhosrokiani, P., & Asl, M. P. (2021, July 4-5). Text mining life cycle for a spatial reading of Viet Thanh Nguyen's The Refugees (2017). In *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*. Taiz, Yaman. https://doi.org/10.1109/ICOTEN52080.2021.9493520

Mayo, M. (2017). *A general approach to preprocessing text data*. KDnuggets. https://www.kdnuggets.com/2017/12/general-approach-preprocessing-text-data.html

Mazzola, L., Siegfried, P., Waldis, A., Kaufmann, M., & Denzler, A. (2018, September 25-27). A domain specific ESA inspired approach for document semantic description. In *2018 International Conference on Intelligent Systems (IS)*. Funchal, Portugal. https://doi.org/10.1109/IS.2018.8710507

Misuraca, M., Scepi, G., & Spano, M. (2021). Using opinion mining as an educational analytic: An integrated strategy for the analysis of students' feedback. *Studies in Educational Evaluation*, 68, Article No. 100979. https://doi.org/10.1016/j.stueduc.2021.100979

Mumtaz, D., & Ahuja, B. (2018). A lexical and machine learning-based hybrid system for sentiment analysis. In B. Panda, S. Sharma & U. Batra (Eds*.), Innovations in Computational Intelligence: Best Selected Papers of the Third International Conference on REDSET 2016* (pp. 165-175). Springer. https://doi.org/10.1007/978-981-10-4555-4_11

Naghibi, N. (2016). *Women Write Iran: Nostalgia and Human Rights from the Diaspora*. University of Minnesota Press.

Nekai, P. (2013). *From education to segregation:Iran's sexist policy*. PROSPECT https://prospect-journal.org/2013/01/18/womens-access-to-higher-education-in-iran/

Paroubek, P., & Pak, A. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

Ranucci, D. (2019). *We never forgot Princeton's Xiyue Wang during his imprisonment in Iran, student says*. https://www.nj.com/opinion/2019/12/we-must-keep-fighting-for-princeton-grad-student-xiyue-wangs-release-from-an-iranian-prison-opinion.html

Řehůřek, R. (2019). *Models.hdpmodel–Hierarchical Dirichlet Process*. Gensim. https://radimrehurek.com/gensim/models/hdpmodel.html

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45-50). CiteSeer. https://doi.org/10.13140/2.1.2393.1847

Scharl, A., Hubmann-Haidvogel, A., Jones, A., Fischl, D., Kamolov, R., Weichselbraun, A., & Rafelsberger, W. (2016). Analyzing the public discourse on works of fiction–Detection and visualization of emotion in

online coverage about HBO's Game of Thrones. *Information Processing and Management, 52*(1), 129-138. https://doi.org/10.1016/j.ipm.2015.02.003

Shahrokni, N. (2020). *Women in Place: The Politics of Gender Segregation in Iran*. University of California Press.

Shi, T., Kang, K., Choo, J., & Reddy, C. K. (2018). Short-text topic modeling via non-negative matrix factorization enriched with localword-context correlations. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1105-1114). ACM Publishing. https://doi.org/10.1145/3178876.3186009

Sofian, N. B., Keikhosrokiani, P., & Asl, M. P. (2022). Opinion mining and text analytics of reader reviews of Yoko Ogawa's The Housekeeper and the Professor in Goodreads. In P. Keikhosrokiani & M. P. Asl (Eds.), *Handbook of Research on Opinion Mining and Text Analytics on Literary Works and Social Media* (pp. 240-262). IGI Global. https://doi.org/10.4018/978-1-7998-9594-7.ch010

Subramanian, D. (2019). *Text Mining in Python: Steps and Examples*. Medium. https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b

Suhendra, N. H. B., Keikhosrokiani, P., Asl, M. P., & Zhao, X. (2022). Opinion mining and text analytics of literary reader responses: A case study of reader responses to KL Noir volumes in Goodreads using sentiment analysis and topic. In P. Keikhosrokiani & M. P. Asl (Eds.), *Handbook of Research on Opinion Mining and Text Analytics on Literary Works and SocialMedia* (pp. 191-239). IGI Global. https://doi.org/10.4018/978-1-7998-9594-7.ch009

Sukhija, N., Tatineni, M., Brown, N., Van Moer, M., Rodriguez, P., & Callicott, S. (2016). Topic modeling and visualization for big data in social sciences. In *2016 International IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)* (pp. 1198-1205). IEEE Publishing. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0183

Teoh, Y. Z. I., & Keikhosrokiani, P. (2020). Knowledge workers mental workload prediction using optimised ELANFIS. *Applied Intelligence, 51,* 2406-2430. https://doi.org/10.1007/s10489-020-01928-5

Vasapollo, S. (2020). *Causes of the Iranian 1979 Revolution: Historical and Political Aspects.* ASERI.

Vinodhini, G., & Chandrasekaran, R. (2012). Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering, 2*(6), 282-292.

Ying, S. Y., Keikhosrokiani, P., & Asl, M. P. (2021). Comparison of data analytic techniques for a spatial opinion mining in literary works: A review paper. In F. Saeed, F. Mohammed & A. Al-Nahari (Eds.), *Innovative Systems for Intelligent Health Informatics. IRICT 2020. Lecture Notes on Data Engineering and Communications Technologies* (pp. 523-535). Springer. https://doi.org/10.1007/978-3-030-70713-2_49

Ying, S. Y., Keikhosrokiani, P., & Asl, M. P. (2022). Opinion mining on Viet Thanh Nguyen's the sympathizer using topic modelling and sentiment analysis. *Journal of Information Technology Management, 14*(Special Issue), 163-183. https://doi.org/10.22059/jitm.2022.84895