Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

Opinion Mining on Viet Thanh Nguyen's The Sympathizer Using Topic Modelling and Sentiment Analysis

Sea Yun Ying*

*Corresponding Author, School of Computer Sciences, University Sains Malaysia, 11800 Minden, Penang, Malaysia. E-mail: phslan@gmail.com

Pantea Keikhosrokiani

Corresponding author, School of Computer Sciences, University Sains Malaysia, 11800 Minden, Penang, Malaysia. E-mail: pantea@usm.my

Moussa Pourya Asl

School of Humanities, University Sains Malaysia, 11800 Minden, Penang, Malaysia. E-mail: moussa.pourya@usm.my

Abstract

In attempts to examine the mapped spaces of a literary narrative, various quantitative approaches have been deployed to extract data from texts to graphs, maps, and trees. Though the existing methods offer invaluable insights, they undertake a rather different project than that of literary scholars who seek to examine privileged or unprivileged representations of certain spaces. This study aims to propose a computerized method to examine how matters of space and spatiality are addressed in literary writings. As the primary source of data, the study will focus on Viet Thanh Nguyen's The Sympathizer (2015), which explores the lives of Vietnamese diaspora in two geographical locations, Vietnam, and America. To examine the portrayed spatial relations, that is which country is privileged over the other, and to find out the underlying opinion about the two places, this study performs topic modelling with Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) by using TextBlob. In addition, Python is used as the analytical tool for this project as it supports two LDA algorithms: Gensim and Mallet. To overcome the limitation that the performance of the model relies on the available libraries in Python, the study employs machine learning approach. Even though the results indicated that both geographical spaces are portrayed slightly positively, America achieves a higher polarity score than Vietnam and hence seems to be the favored space in the novel. This study can assist literary scholars in analyzing spatial relations more accurately in large volumes of works.

Keywords: Opinion Mining; Sentiment Analysis; Spatial Analysis; The Sympathizer.

DOI: https://orcid.org/10.22059/jitm.2022.84895 Manuscript Type: Research Paper University of Tehran, Faculty of Management

Received: January 12, 2020 Accepted: March 25, 2020

Introduction

Advancements in artificial intelligence and the emergence of big data technologies have transformed data mining research in various disciplines (Abdelrahman & Keikhosrokiani, 2020; Keikhosrokiani, 2019, 2020). The rapid developments in information technology and data sciences have revolutionized the traditional methods of text analysis. Manual analysis of textual information has always been a subjective and laborious work, and hence open to critical controversy. In other words, affixing emotional and personal values to the findings and the difficulty in producing consistent results in larger amount of information are the two major problems of manual text analyses. The advent of real life applications that combine data mining, web mining and text mining techniques has rendered the study of opinions embedded in large volumes of text much easier and more accurate (Khan et al., 2014). Automated text mining and summarization systems have helped to avoid subjective biases and overcome human limitations (Lum, 2017). In particular, text mining techniques like sentiment analysis have recently been used in literary studies to extract sentiments from texts automatically (Mohammad, 2016; Nalisnick & Baird, 2013; Roque, 2012; Schmidt et al., 2020; Ying et al., 2021). Despite the growing attempts, certain types of textual analyses such as the examination of literary geography have remained underresearched.

To examine the literary geography or the mapped spaces of a text, various quantitative approaches have been deployed to extract data from texts to graphs, maps, and trees (Pourya Asl, 2020; Van der Bergh, 2013). Some have used geographical information system (GIS) to chart a work's "character or plot trajectories along the physical topography of a given region" (Queiroz & Alves, 2015; Tally Jr, 2017). Though these methods offer invaluable insights, they undertake a rather different project than that of literary scholars examining the privileged or unprivileged representations of certain spaces.

Throughout the history, fictional writings have been used to creatively record power-ful and haunting visions of socio-political events. Realistic fictional works about political realities, conflicts, wars, and revolutions are often presented in actual geo-graphical places, thus elevating space and spatial relations to primary status. Matters of space, place, and mapping have indeed come to the forefront of critical discussions of literature and culture in the new millennium (Pourya Asl, 2019 & 2020; Tally Jr, 2017).

Globalization and the massive increase in mobile populations and border-crossings have redrawn the traditional geographical boundaries and have "helped to push space and spatiality into the foreground" (Tally Jr, 2017). Within the contemporary context of global mobility, diasporic literature as a disciplinary field of study has become more engaged with geographically based questions such as the relation of a diasporic writer or a text to its homeland and hostland (Asl, 2018 & 2019). In diasporic literary criticism, spatial representation has become a principal criterion for critical evaluation of a given text. Certain diasporic writers from the East are accused of disavowed participation in the production of favored knowledge for Western audiences by representing their country of birth as an

undesired dystopian world while depicting the West as an emancipatory and utopian one (Asl, 2020). Much of such criticism, however, not only is highly prone to the critics' biased views but also is based on manual data collection and textual analysis, hence the controversy over the accuracy of the data and credibility of the findings.

This study seeks to propose a computerized method to examine how matters of space and spatiality are addressed in literary writings. Viet Thanh Nguyen's The Sympathizer (2015) is used as the primary source of data. The text depicts the horrors and absurdity of the Vietnam War on Vietnamese people both at home and abroad. It is a layered diasporic tale that is narrated in the wry confessional voice of a "man with two minds" and two spaces, Vietnam as his country of birth and America as the host country (Hadi & Asl, 2021).

To examine the existing spatial relations, i.e. which country is privileged over the other, this study seeks to perform topic modelling and sentiment analysis. The study aims at studying how sentiment analysis can be used in textual reviews and extracting orientations in the e-book of the novel; and to demonstrate how the information can be used for trend detection and knowledge discovery. Therefore, it seeks to find out underlying opinion in the depiction of the two geo-graphical places. To achieve this goal, two models will be conducted: (1) topic modelling with LDA and (2) sentiment analysis by using TextBlob. LDA will be used to separate the whole text into two parts: one representing America and the other representing Vietnam. Sentiment analysis will be performed to find out which geographical locations are portrayed positively or negatively.

The lexicon resource needed for performing sentiment analysis used in this project is TextBlob. It stands on the giant shoulders of Natural Language Toolkit (NLTK) and pattern, and functions well with both of them. It consists of many features such as noun phrase extraction, part-of-speech tagging, sentiment analysis, classification by using Naïve Bayes or Decision Tree, Language translation and detection supported by Google Translate, Tokenization (splitting text into words and sentences), word and phrase frequencies, parsing, N-grams, Word inflection (pluralization and singularization) and lemmatization, spelling correction, add new models or language through extensions, and WordNet integration. And finally, Python will be used as the analytical tool for this project as it supports two LDA algorithms: Gensim and Mallet. It is hoped that the produced model benefits literary analysts and academics who are interested in analyzing long textual data.

Methodology

This section covers the details of the methodologies that are being used to address the following question. The question engages in a form of literary cartography by which the diasporic writer maps the real spaces of the two countries of America and Vietnam in his novel The Sympathizer. The second question is focused on determining whether topic modelling and sentiment analysis present value to text analysis–a traditional approach used by literary scholars in both its current and future form. This section will provide a detailed

explanation about every technique and tool used in this study from data collection, data pre-processing, data exploratory and also the final analysis. Besides, data science project life cycle is shown in this section.

A.Data Science Process: OSMEN Framework

A data science framework called Obtain, Scrub, Explore, Model, Interpret (OSEMN) (Kumari et al., 2020), that covers every step of the data science project life cycle from end to end is applied in this project. OSEMN process is considered as a taxonomy of tasks that can be used as a blueprint for any data science problems especially problems that can be solved by using machine learning algorithms. The pipeline of OSEMN includes: (1) O – Obtain data; (2) S – Scrub data; (3) E – Explore data to find relevant patterns and trends; (4) M – Model data; and (5) N - iNterpret data.

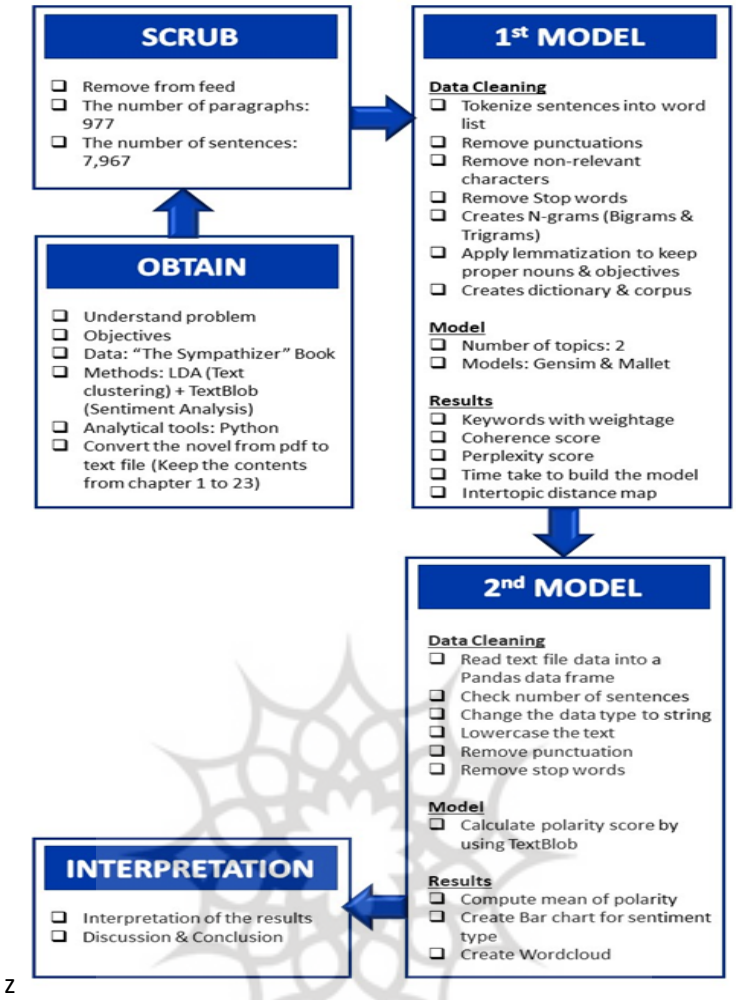
B.Data Science Project Life Cycle

This section mainly discusses the proposed data science lifecycle applied in this study. Thus, a proposed data science lifecycle with 5 main phases is designed as illustrated in Figure

Obtain Data

The first stage of a data science project is very straightforward: collect and obtain the data required for this project. The data used in this project is a novel titled The Sympathizer (2015) by Viet Thanh Nguyen. It is one of the best-selling novels that won the 2016 Pulitzer Prize for Fiction. The Sympathizer is a historical spy novel that consists of twenty-three chapters. The central theme of this novel is about the dual identity of an unnamed half-French, half-Vietnamese narrator, as a mole and immigrant, and the Americanization of the Vietnam War in international literature. It is interesting to understand how the writer represents the narrator's country of birth (Vietnam) and the adopted homeland (America). The data, i.e. the novel in PDF format, used here is an example of unstructured data.

It is necessary to know how to automatically obtain the data rather than the manual processes of data collection. The example of manual processes for this task is pointing and clicking with a mouse and then copy and pasting the whole text from the document. In this work, scripting using Python is suggested as scripting languages like Python can make data retrieval a lot easier. To read the data directly into the data science programs by using Python, specific packages and coding are used.



z

Figure 1. The proposed Data Science Lifecycle

Scrub Data

The main task in the second stage is to clean and filter the data from errors, miss-ing values, irrelevant records, and so on. However, cleaning the data means to throw away, replace, and/or fill missing values/errors if necessary. The skills required in this stage is the scripting language of Python. After checking the text file, the symbol of '\x0c' or '\f' which is a symbol of form feed are detected. The function of the symbol is forcing a printer to move to the next sheet of paper. This might be a problem when separating the whole text either into paragraphs or sentences. The total number of the paragraph in the novel is 977 paragraphs. The whole text is then split by sentences and extra spaces before every sentence stored into a data frame are removed. The total number of sentences in this novel is 7967.

Justification on Text Clustering by using LDA

The faster way to find how America and Vietnam are represented in the book is to separate the whole data into two parts: America and Vietnam and then perform sentiment analysis for both parts of America and Vietnam. The first task focuses on discussing how to split the whole text into each part. Supervised learning, semi-supervised learning and unsupervised learning are

the main categories of machine learning. However, this study will select one of the unsupervised learning algorithms to perform the text clustering, details of which were explained earlier. Topic modelling is unsupervised learning that helps to analyze large volumes of text data by clustering the documents into groups. It is a type of learning that makes extracting previously unknown patterns or information in data set without any target labels. Latent Dirichlet Allocation or LDA is chosen as it is one of the most popular probabilistic topic modelling techniques in machine learning started in 2003 (Keikhosrokiani, 2020; Wei & Croft, 2006). However, they also mentioned the feasibility and effectiveness of LDA in information retrieval is mainly unsure. In short, unsupervised learning will never be the best choice if a huge training data is provided. The accuracy and effectiveness of the model cannot be known unless the whole text is being manually labeled and then compared with the label that the model provided. The first requirement is to pick the number of topics before building the LDA model. In this case, the number of topics will be set at 2 as the data source is mainly related to Vietnam and America. Each sentence in the novel is represented as a distribution over topics. The last assumption of LDA is each topic is represented as a distribution over a group of words (keywords).

Explore Data of LDA

In the exploration phase, the goal is to understand the patterns and values in the data. It is useful to build an intuition for the form of data to get ideas for data transforms and predictive models to use in the model data step. Before creating the classification model, the details of the data are being checked. After checking, there is no missing value and then proceed to the model data step. The sentences are dropped when it only consists of one word. However, there are no sentences that consist of only one word in this dataset after checking the number of rows of the data frame.

Before conducting LDA in Python, it is necessary to install some packages in command prompt. The stopwords from NLTK and spacy's en_core_web_sm model for text pre-processing. The spacy model will be used for lemmatization – a step that converts a word to its root word. It is necessary to import packages for LDA before start to build the model. The purpose to import the packages is mainly for data handling, visualization and record the time to build a model.

Data pre-processing step is a necessary step before building any models. It is a process that transforms the data to the language that a machine or computer can understand. The main task of the pre-processing step in text mining is to remove irrelevant data. In natural language processing, irrelevant data can be referred to as stop words. Stop words are the words that are the most commonly exist in a language but it does not bring important significance to Search Queries. The existing of those words in the model might run up the memory space of the machine or take up the processing time. The new line characters, extra space and single quotes in the dataset are being removed. The stop words will be also removed for Model 1.

The sentences are tokenized into list of words, dropping punctuations and non-relevant characters by using Gensim's simple_preprocess. The command of deacc=True is applied to remove the punctuation in the sentences. N-grams are n number of words frequently occurring together in the text. It is possible to build and implement the bigrams, trigrams, and n-grams in Gensim's Phrases model by setting suitable min_count and threshold arguments. The lower the values of these parameters, the easier it is for words to be combined to N-grams. After building bigrams and trigrams models, it is time to remove stop words, make n-grams and lemmatization to keep only proper nouns and adjectives. The main purpose to apply LDA in this project is to separate the whole text in the novel into two parts: America and Vietnam. Thus, the keywords generated by the model must be easier to determine whether the part is belonging to America or Vietnam. For example, the word 'American' might be one of the keywords that is used to represent the part of America in the novel. 'American' can be labelled either as a proper noun or adjective in the English language and this is the main purpose to include only proper noun and adjective in this model.

The last step before building an LDA model is to create the dictionary and corpus. Gensim will generate a unique id for each word in the text. For example, the word 'able' occurs once in the first sentence. Then, the produced corpus is displayed as (0, 1) where the word id of 'able' is 0 and the number 1 shows that it occurs only one time in the first sentence.

Model Data and Interpret Results of LDA

In addition to the corpus and dictionary, it is necessary to set the number of topics. In this case, the number of topics is set to 2. Topic models with LDA using Gensim is built. Clustering methods do not need any training data to group or cluster the observations that have similar characteristics because they let algorithm to define the output based on its theory. The time taken used to build the model, keywords and its weightage, perplexity score, coherence score, and an intertopic distance map will be the methods used to interpret the results of LDA with Gensim. The results of this model are interpreted.

Gensim provides a wrapper to enable Mallet's LDA to be implemented within Gensim itself. Thus, the coherence score of LDA Mallet model can be computed and then compare with the coherence score of LDA with Gensim. The algorithm with the best coherence score will be selected. After choosing the best model, it is able to assign and tag each statement with its most relevant topic number. 0 is label for America part while 1 is label for Vietnam part. Before performing the second task in this project, it is necessary to export the America and Vietnam part into text file respectively. America and Vietnam part are exported into a text file respectively.

Justification on Sentiment Analysis by using Textblob

The second task is mainly about how to perform sentiment analysis for both America and Vietnam part. Before performing the sentiment analysis, it is necessary to choose a suitable

approach for this dataset. Sentiment analysis can be grouped into three common approaches: Lexicon-based approach, Learn-based approach, and Hybrid approach. In this study, Lexicon-based approach is selected as it is an unsupervised learning, and it does not require any labelled data. Lexicon-based approach can be divided into three common types: dictionary-based approach, corpus-based approach, and manual approach. Manual approach is not logical to apply in this study as it is a very time-consuming method. Manual approach cannot be applied alone because it is usually combined with dictionary-based and corpus-based approach in order to prevent the mistakes that appear when performing sentiment analysis by using only dictionary-based or corpus-based approach alone. In this study, dictionary-based approach is used because it is the most effective methods among the three.

Dictionary-based method is a method that depends on searching opinion seed words before looks for the dictionary of their synonyms and antonyms that required for polarity detection in sentiment analysis tasks. This method can use existing dictionaries such as SentiWordNet or Textblob. However, different dictionaries contain different words and have their own rules to set the polarity of the words. In this project, the Textblob dictionary will be selected because the accuracy is normally higher than the other dictionaries. The performance of VADER is also quite high but it is a lexicon and rule-based sentiment analysis tool special design to deal with social media texts, movie reviews, and product reviews. VADER always works well on social media type text. Besides, it also supports emoji for sentiment classification. However, the text in this study is not social media texts or domain specific/ noisy text but a well-written text. Therefore, TextBlob is used as it performs well across domains such as movie, healthcare, and political.

Explore and Model Data of SA

The text file for the America part is imported. The number of sentences for the America part and Vietnam part are 6 343 and 1 625 respectively. The data is then stored in a data frame. Some packages were required to install in Command Prompt and Python notebook 3 before starting to perform sentiment analysis by using TextBlob.

In the data pre-processing step, the text type is changed to string before dropping any non-relevant data. After that, the text is changed to lowercase. Punctuation and stop words also removed before calculating the polarity for the sentences. Lemmatization and Stemming is not considered in this case because it is not necessary to be used especially for the well written text as suggested by Ganesan (2019). The polarity score is being calculated and recorded in the same data frame.

Interpret Results of SA

The mean of the polarity score will be computed in order to identify whether the spatial representation of America and Vietnam in the novel is positive or negative. A bar chart will be used to show the number of sentiment types for each space. Word cloud will be used to depict

keyword for the sentences with polarity score either equals to 1.0 or -1.0. Then the result for sentiment analysis is interpreted.

Interpret Final Results

Interpreting models and data is the final and most crucial step of a data science project. Interpreting results normally refers to presentation of data and delivering the results in order to answer the client's questions, together with the actionable insights gained from data science process. The key skills to have in this process is beyond technical skills, data scientist should know the way to present findings in a way that can answer clients' questions. Some basic technical skills needed might be included visualization tools like Matplotlib for Python. Soft skills like presenting and communication skills, paired with a flair for reporting and writing skills are also the skills that cannot be ignored in this stage.

Findings

A.Introduction

In this section, the results of the study are presented and discussed with reference to the aim of the study, which was to find how America and Vietnam are represented in Nguyen's The Sympathizer. The sub-aim of this project is to show how opinion mining presents value to text analysis – a traditional approach used by literary scholars in current form and its limitations. The results of the first objective will be presented first, followed by the limitations of the selected technique.

B.Results of Topic Modelling by LDA

As described earlier, there are two LDA algorithms that are used to build the model and the ones with the best coherence score will be chosen. The results of the algorithms are presented in Table 1.

Table 1. Results for LDA Model: Gensim Vs. Mallet

Performance	**LDA with Gensim**	**LDA with Mallet**
Time taken to build model	11.78 seconds	34.78 seconds
Perplexity Score	-7.215	-
Coherence Score	0.6601	0.7027
Keywords	Vietnam: general, white, last, little, many, dead, first, bad, great, black, right, long, **Vietnamese**, open, true, high, major, red, close, dark America: good, **American**, much, least, young, old, next, enough, well, poor, small, important, free, new, revolutionary, real, human, ready, hot, entire	Vietnam: white, dead, black, great, long, young, **Vietnamese**, major, open, small America: general, good, **American**, bad, poor, high, important, free, Chinese, innocent

The time taken to build the model of LDA with Gensim and Mallet is 11.78 seconds and 34.78 seconds respectively. However, the time taken is not the most important criteria needed to be considered here. The model with the higher coherence score and the higher weightage of the particular keywords that can be used to represent Vietnam and America will be selected. Perplexity score is not considered in this study because it may not correlate to human judgment. In addition, the perplexity function is not implemented for the Mallet wrapper so the perplexity score for LDA with Mallet is not given in Table I.

Topic coherence gives a convenient measure in order to judge how good a given topic model is. It will be very problematic to set the optimal number of topics with-out going into the content. Many LDA models, with different numbers of topics, should be built and the ones with the highest coherence score will be chosen. Choosing too much value in the number of topics always leads to more detailed sub-themes, however, some keywords might be repeated again and again. In this task, the number of the topic is fixed to 2 as the writer mainly describe two countries: Ameri-ca and Vietnam in the novel. Therefore, the algorithm of LDA with Mallet is select-ed as it achieves a higher coherence score and the weightage of the keywords 'Vietnamese' and 'America' for Mallet are also higher than Gensim. Each topic built by the LDA model is a combination of keywords and each keyword contributes a certain weightage to the topic. The weightage represents the importance of the keyword to that particular topic. The weightage of the keywords is provided in Table 2.

Table 1. Keywords With Weightage: Gensim Vs. Mallet

Model	**Keywords with weightage**	
	Vietnam	**America**
LDA with Gensim	general: 0.025; white: 0.020; last: 0.018; little: 0.017; many: 0.017; dead: 0.016; first: 0.014; bad: 0.013; great: 0.013; black: 0.012, right: 0.012; long: 0.010, **Vietnamese: 0.009**; open: 0.008; true: 0.008; high: 0.008; major: 0.007; red: 0.007; close: 0.006; dark: 0.006	good: 0.031; **American: 0.018**; much: 0.015; least: 0.014; young: 0.011; old: 0.011; next: 0.010; enough: 0.010; well: 0.009; poor: 0.008; small: 0.008; important: 0.008; free: 0.007; new: 0007; revolutionary: 0.007; real: 0.006; human: 0.006; ready: 0.006; hot: 0.006; entire: 0.006
LDA with Mallet	white: 0.032; dead: 0.022; black: 0.018; great: 0.018; long: 0.017; young: 0.016; **Vietnamese: 0.016**; major: 0.013; open: 0.013; small: 0.012	general: 0.045; good: 0.044; **American: 0.027**; bad: 0.020; poor: 0.013; high: 0.011; important: 0.011; free: 0.010; Chinese: 0.009; innocent: 0.008

After choosing the model with a higher coherence score and higher weightage for particular keywords, the next step is to examine the produced topics and the associated keywords by using pyLDAvis library. The package of pyLDAvis is designed to help users interpret topics found by the algorithm easily. It extracts information from the resulting LDA model – LDA Mallet model to provide interactive visualization. It is the best way to illustrate the distribution of topics – keywords in jupyter notebook. The pyLDAvis's output of the LDA Mallet model is given in Figure 2.

Each bubble displayed on the left-hand side plot represents a topic. The bigger the bubble, the more widespread is the topic in the document. Thus, a good topic model will have fairly big and non-overlapping bubbles decentralized throughout the chart instead of being clustered in one region of the chart. A model with too many topics will have a higher possibility to have many overlaps, small-sized bubbles that clustered in one quadrant of the chart. With prior knowledge, the number of topics is fixed to 2 so the best model can be found easily. The keywords 'American' and 'Vietnamese' are the forth and eleventh most salient terms in the document.

The most salient terms for the different topics can be viewed by clicking the particular bubbles. The words and bars on the right-hand side will update after any of the bubbles in the diagram is selected. The most relevant words for Topic 1 (Ameri-ca) and Topic 2 (Vietnam) is given in Figure 3 and 4 respectively. In Topic 1 (America), the keyword of 'American' is the third most relevant word in the topic. The estimated term frequency of 'American' within the selected topic also quite high. In Topic 2 (Vietnam), the keyword of 'Vietnamese' is the seventh most relevant word in the topic. The estimated term frequency of 'Vietnamese' within Vietnam's topic also very high. Therefore, the model built by using LDA Mallet model can be considered as a good model as it successfully separates the terms of 'American' and 'Vietnamese' in the document. The Intertopic Distance Map of LDA Genism model and the details of LDA Gensim model are shown in Figure 2.

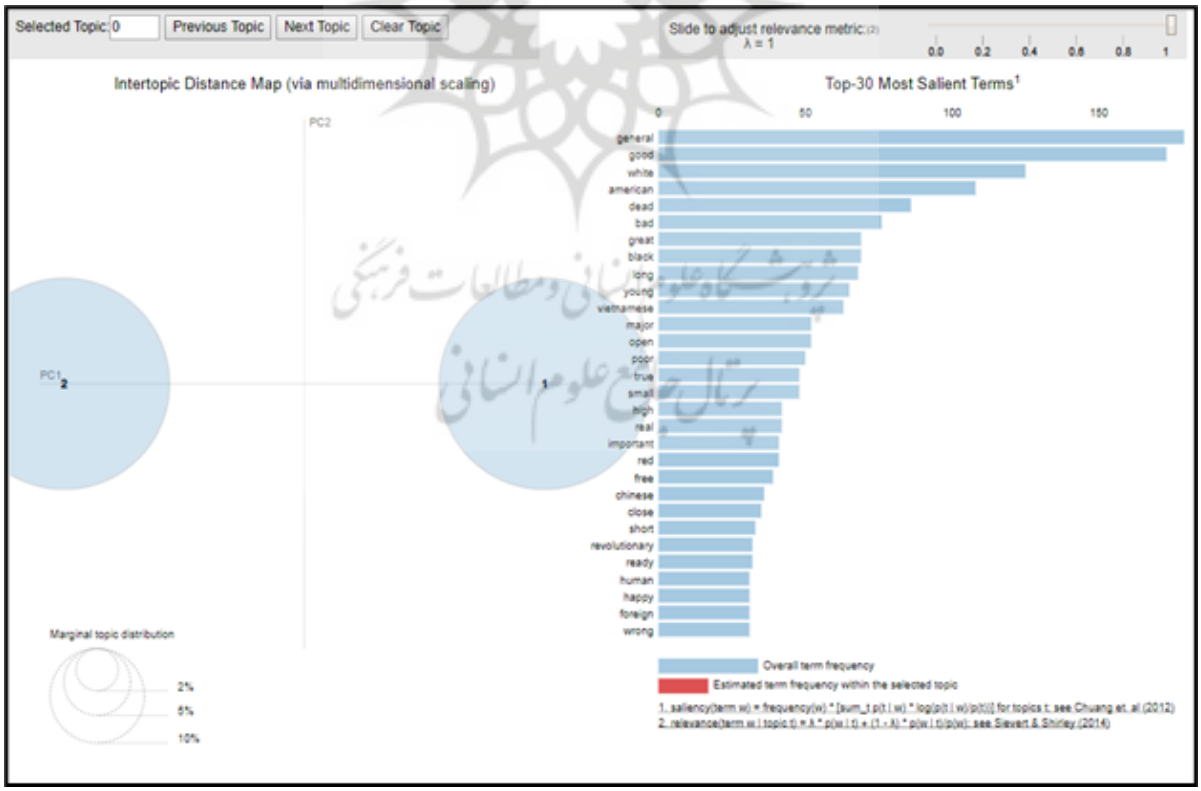


Figure 1. Intertopic Distance Map of LDA Mallet model

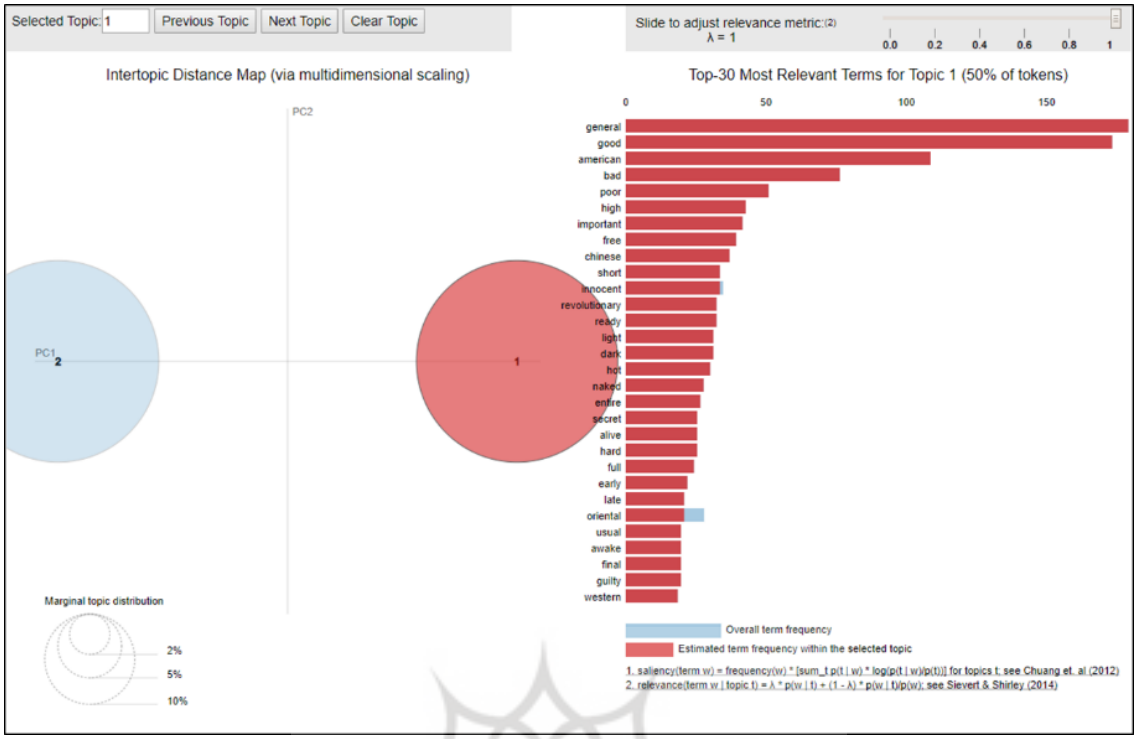


Figure 2. Top 30 most relevant terms for Topic 1 (America) in LDA Mallet model

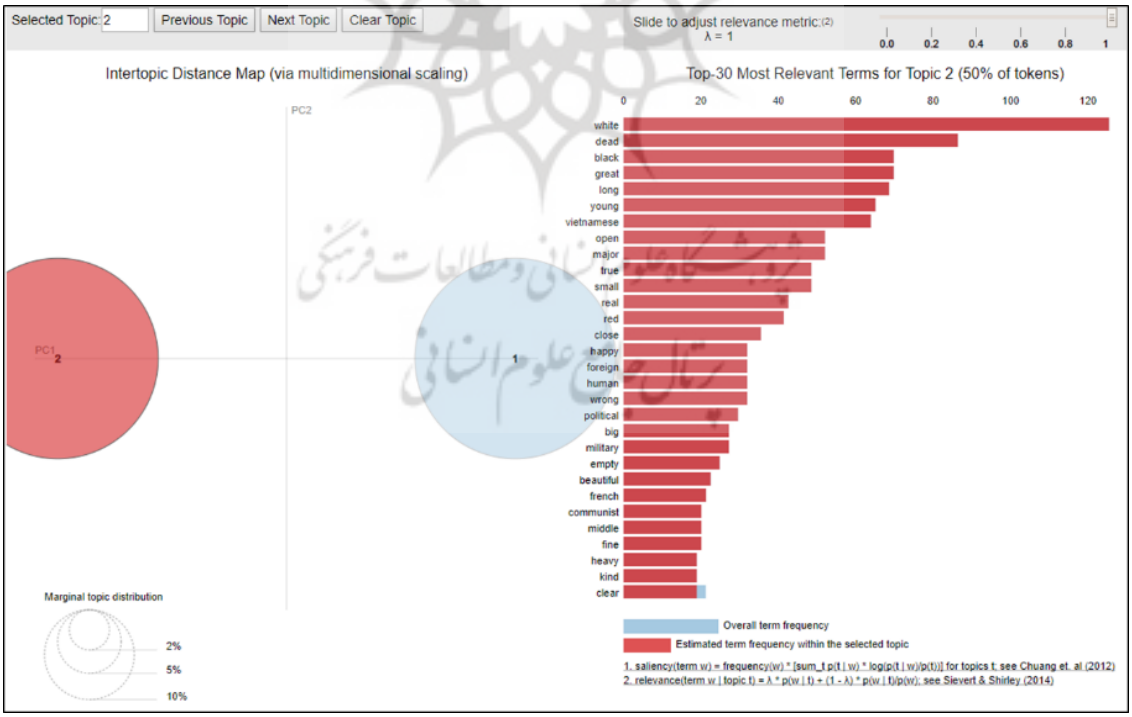


Figure 3. Top 30 most relevant terms for Topic 2 (Vietnam) in LDA Mallet model

C. Limitation of LDA

LDA is an unsupervised learning model able to find underlying topics in unlabeled data. It is able to perform text clustering when there is no training data given. However, the "parts" separated by the unsupervised learning model may not match the true parts in the data. Topic coherence score is a good approach to compare the different topic models based on their human-interpretability. Table 3 shows some relevant spatial keywords to evaluate the LDA Mallet model. LDA Mallet model success to separate the part with most of the keywords that represent for America from the whole text. However, LDA Mallet model did not perform well when separate the Vietnam part with the keywords of Vietnam.

Table 2. Human Judgment Vs. LDA Mallet Model

Keywords		**Topic (number of keywords)**	
		Vietnam	**America**
Vietnam	Vietnam	11	15*
	Vietnamese	**59***	18
	Saigon	28	53*
America	USA	2	**6***
	United States	0	**4***
	States	0	**11***
	America	17	**55***
	American	35	**147***
	Americans	26	**58***
	Washington	0	**3***

POS or Part-of-speech tagging is the process of marking up a word in a corpus to a corresponding part of a speech tag, based on its context and definition. In this case, the spaCy POS Tagging did not tag the word 'Vietnam' and 'Saigon' as proper nouns in singular form (NNS) but tag them as noun (NN). However, the spaCy library is chosen because it is the fastest NLP framework and easy to learn and use. The library of spaCy uses the latest and best algorithms so it performs bet-ter than NLTK in word tokenization and POS-tagging task. The POS tagging problem might be one of the reasons that the LDA Mallet model do not separate the Vietnam part well. Another reason is some of the sentences in the text contains both America and Vietnam keywords. The example is given in the next page:

*"The young **Vietnamese** who are enamored of **America** hold the key to South **Vietnam**'s freedom."*

D. Sizing of Graphics

As described earlier, the library of TextBlob is used to perform the sentiment analysis. TextBlob is a python library that offers a simple API to access its methods in order to perform some basic NLP tasks. It goes along finding words and phrases before assigning polarity to them. It will average the polarity of them all together for longer text. The polarity of TextBlob is a float value within the range from -1 to 1 where 0 indicates neutral, 1 indicates a very positive sentiment and -1 indicates a very negative sentiment. The result of sentiment analysis

for both America and Vietnam part are concluded in Table IV. Types of sentiment in America and Vietnam part are given in Figure 5 and 6.

Table 3. SA Results for America and Vietnam Part

Topic	Polarity score	Types of sentiment (number of sentences)			Total number of sentences
		Positive	**Neutral**	**Negative**	
America	0.0414	1 834 (28.9%)	3 519 (55.5%)	990 (15.6%)	6 343
Vietnam	0.0304	665 (40.9%)	441 (27.2%)	519 (31.9%)	1 625

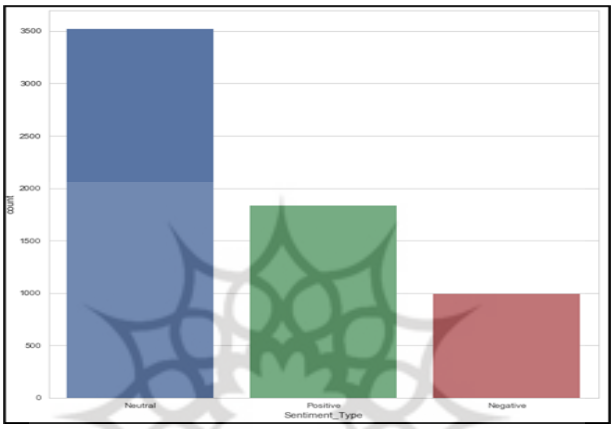


Figure 4. Types of sentiment in America part

The total number of sentences for the part of America and Vietnam are 6 343 and 1 625 respectively. It looks logical as the number of the keywords for America is higher than the number of keywords for Vietnam. The polarity score for America and Vietnam part are 0.0414 and 0.0304 respectively. Both of them show slightly positively in the novel but America is more positive than Vietnam. Based on the result given, most of the sentiments in America part are neutral at 55.5%. The rest of sentiments are positive and negative sentiments at 28.9% and 15.3% respectively. In part of Vietnam, most of the sentiments are positive at 40.9%, followed by negative sentiments at 31.9% and neutral sentiments at 27.2 %.

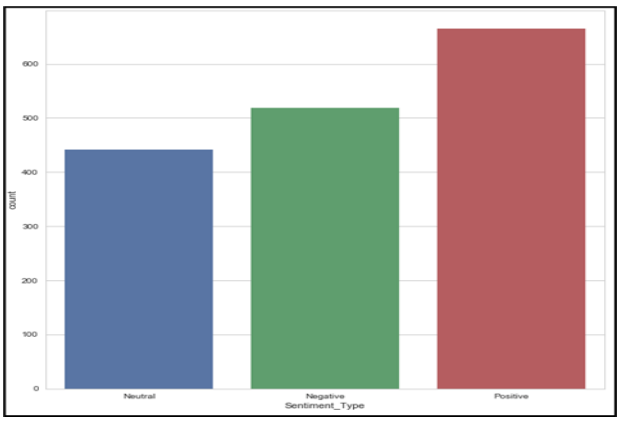


Figure 5. Types of sentiment in Vietnam part

In order to know the positive and negative words used to describe both places, word cloud is used to depict keyword for the sentences with polarity score equal to 1 and -1. In the word cloud, only the adjectives and adverbs will be displayed as Text-Blob only assign polarity to adjectives and adverbs. The wordcloud of positive sentiment and negative sentiment for America part is shown in Figure 7 and Figure 8. Figure 9 and Figure 10 show the wordcloud of positive sentiment and negative sentiment for Vietnam part.

Figure 6. Keywords of positive sentiment in America part

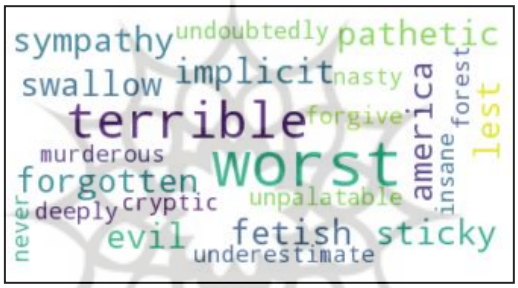


Figure 7. Keywords of negative sentiment in America part

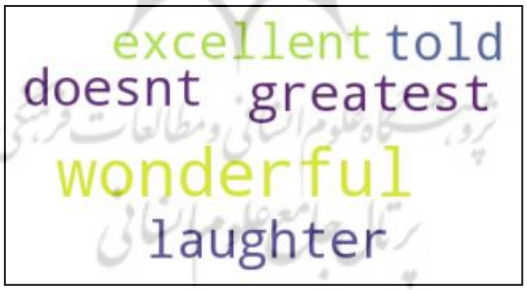


Figure 8. Keywords of positive sentiment in Vietnam part

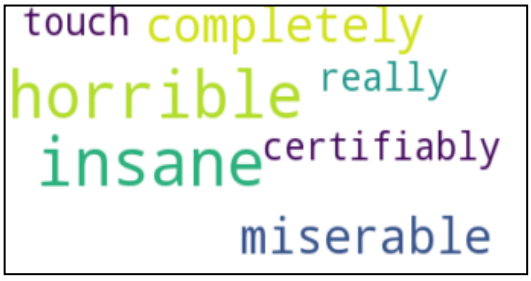


Figure 9. Keywords of negative sentiment in Vietnam part

In America part, the top 5 positive words are 'delicious', 'best', 'perfect', 'perfect-ly', and 'greatest'. The top 5 negative words in America part are 'worst', 'terrible', 'insane', 'evil' and 'pathetic'. In Vietnam part, the positive words in the text are 'wonderful', 'excellent', and 'greatest'. The negative words in Vietnam part are 'horrible', 'insane', and 'miserable'. The other words are ignored as some of them do not have any polarity score after checking by using TextBlob library.

The word 'delicious' gives a meaning highly pleasant to the taste. Thus, it appears in America part which means that there are so many delicious foods in America. People live in America can have a good meal there. The word 'best' brings a meaning of the most excellent or desirable type or quality. It can be used to describe things, places or people. Figure 11 shows the Top 10 rows in America part that consists of the word 'best'. Democracy, things, and people in America are described by the word 'best'.

	Full Text	Text	senti_score	polarity	Sentiment_Type
964	But\tyour\tnumbers\tare\tgrowing,\tand\tdemocr...	numbers growing democracy gives best chance fi...	Sentiment(polarity=1.0, subjectivity=0.3)	1.0	Positive
1549	That\t was\t one\t of\t the\t best\r\nundergr...	one best undergraduate theses ive ever read	Sentiment(polarity=1.0, subjectivity=0.3)	1.0	Positive
1581	We\tteach\tyou\tthe\tbest\tof\twhat\twas\r\nth...	teach best thought said explain america world ...	Sentiment(polarity=1.0, subjectivity=0.3)	1.0	Positive
1584	What\tabout\tthose\two\r\nhave\tnot\tlearned\...	learned best thought said	Sentiment(polarity=1.0, subjectivity=0.3)	1.0	Positive
1873	He??s\tthe\tbest\tthing\tthat\tcould\thave\tha...	hes best thing could happened us said	Sentiment(polarity=1.0, subjectivity=0.3)	1.0	Positive
2159	That??s\t the\t best\t newspaper\r\npolicy.	thats best newspaper policy	Sentiment(polarity=1.0, subjectivity=0.3)	1.0	Positive
2730	They\t owned\t the\t means\t of\t production,...	owned means production therefore means represe...	Sentiment(polarity=1.0, subjectivity=0.3)	1.0	Positive
2786	The\tbest\tmoney\tcan\tbuy.	best money buy	Sentiment(polarity=1.0, subjectivity=0.3)	1.0	Positive
3516	\r\nEven\tas\tI\twas\tplanning\tfor\tthe\tposs...	even planning possibility returning also best ...	Sentiment(polarity=1.0, subjectivity=0.3)	1.0	Positive
3722	But\t the\t best-\r\ndressed\tpeople\tin\tthe...	best dressed people hotel appeared fellow coun...	Sentiment(polarity=1.0, subjectivity=0.3)	1.0	Positive

Figure 10. Top 10 rows of America part that contains the word 'best'

The words 'perfect' and 'perfectly' have the meaning of having all the required or desirable elements, qualities, or characteristics. It might be used to describe people, things or places. According to Figure 12, the words 'perfect' and 'perfectly' are used to describe the people and things in America part.

	Full Text	Text	senti_score	polarity	Sentiment_Type
1280	Your\ttiming\tcould\tnot\tbe\tmore\tperfect.	timing could perfect	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive
2458	Meanwhile,\t a\r\nblond-wigged\tband\tfrom\tM...	meanwhile blondwigged band manila pounded perf...	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive
2861	\r\nPerfect,\tClaude\tsaid.	perfect claude said	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive
2921	One\t night\t in\t my\t quarters,\t after\t m...	one night quarters rage cooled hardened struck...	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive
5301	The\tOriental\r\nis\ta\tperfect\tstudent,\tthe...	oriental perfect student department chair rema...	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive
5447	Do\t you\t remember\t our\t exams,\t when\t y...	remember exams would always score perfectly wo...	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive
5720	Perfectly\tharmless,\tsaid\tthe\tdoctor.	perfectly harmless said doctor	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive

Figure 11. Text of America part that contains the word 'perfect'

Figure 13 shows the text in the America part that consists of the word ‘greatest’. The word ‘greatest’ is used to describe the talent of people in America. The people there are able to build the world’s greatest weapon arsenal. Some of them might become the greatest revolutionary poets.

	Full Text	Text	senti_score	polarity	Sentiment_Type
4013	Men\t like\t you\t have\t built\t the\t world...	men like built worlds greatest weapon arsenal ...	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive
5026	But\r\nnot\tto\tmention\tTo\tHuu,\tour\tgreate...	mention huu greatest revolutionary poet	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive

Figure 12. Text of America part that contains the word ‘greatest’

Figure 14 displays the text in America part that consists of Top 5 negative words listed above. The word ‘worst’ used to describe corruption, war and enemy while ‘terrible’ is used to describe foods and loss. The word ‘insane’ is used to describe the fireflies. ‘Evil’ is used for describing people: evil communists. ‘Pathetic’ is the word for resignation.

	Full Text	Text	senti_score	polarity	Sentiment_Type
911	A\r\nboring\tjob.	boring job	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
1892	The\tworst\tthing\tabout\tliving\tin\tAmerica\...	worst thing living america corruption	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
1978	The\tperformance\r\nwas\tso\tinsulting\tit\tev...	performance insulting even deflated fetish aud...	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
2228	We\thave\tno\tchoice\tbut\tto\tfight,\tto\tres...	choice fight resist evil resist forgotten	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
3014	He\t spoke\t with\t such\t pathetic\t resigna...	spoke pathetic resignation felt renewed sympathy	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
3109	To\ttell\tyou\tthe\ttruth,\tI\tsaid,\tproceedi...	tell truth said proceeding lie feel terrible	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
3157	Of\tthe\tthree\ttypes\tof\tforgetting,\tthis\t...	three types forgetting worst	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
3630	The\tonly\tflaw\tin\tthis\tmethod\r\nwas\tthat...	flaw method auntie terrible cook sticky rice b...	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
4055	You\tare\ta\tman\r\nwho\thas\tundoubtedly\tsee...	man undoubtedly seen worst war forgive speak u...	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
4545	Very\tevil\tcommuniststhere.	evil communiststhere	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
4946	Dozens\tof\tinsane,\tmurderous\tfireflies\tfli...	dozens insane murderous fireflies flickered fo...	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
5414	I\twouldn??t\tdo\tthis\tto\tmy\tworst\tenemy.	wouldnt worst enemy	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
5416	Never\tunderestimate\twhat\tyou\tcan\tdo\tto\t...	never underestimate worst enemy	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
5627	I\tfelt\tdeeply\tfor\r\nyou,\tthe\tterrible\tl...	felt deeply terrible loss hinted cryptic messages	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
5825	Claude\t told\t me\t that\t this\r\nwas\t nas...	claude told nasty business see	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative

Figure 13. Text of America part that contains the Top 5 negative words

Figure 15 gives the text in the Vietnam part consists of the positive words: ‘wonderful’, ‘excellent’, and ‘greatest’. The word ‘wonderful’ is used to describe idea while the word ‘excellent’ is specific for joke. The last word ‘greatest’ is used for describing the things that are built by a person.

	Full Text	Text	senti_score	polarity	Sentiment_Type
783	It??s\tthe\tgreatest\tthing\tyou\tbuilt.	greatest thing built	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive
845	A\t wonderful\t idea,\t isn??t\t it?	wonderful idea isnt	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive
977	This??he\t tapped\t the\t newspaper?? sounds\t...	thishe tapped newspapersounds wonderful doesnt	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive
1118	\r\nOur\thosts\troared\twith\tlaughter\tas\tif...	hosts roared laughter told excellent joke	Sentiment(polarity=1.0, subjectivity=1.0)	1.0	Positive

Figure 14. Text of Vietnam part that contains the Top 3 positive words

Figure 16 displays the text in the Vietnam part consists of the negative words: 'horrible, 'insane, and 'miserable. The words 'horrible', 'insane', and 'miserable' are used to describe people here.

	Full Text	Text	senti_score	polarity	Sentiment_Type
434	\r\nSo\tyou\tgot\tme\tthere,\tthe\tman\tsaid,\...	got man said miserable	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
513	\r\nShe??s\t completely\t insane,\t Madame\t d...	shes completely insane madame declared	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
941	Like\tBon,\tthey\twere\tcertifiably\r\ninsane\...	like bon certifiably insane men volunteered re...	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
1406	Am\t I\t really\t so\t horrible\t that\t you\...	really horrible recognize friend	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative
1445	But\tcan\tyou\tknow\twhat\tit\tis\tlike\tto\tb...	know like horrible children cry see wife flinc...	Sentiment(polarity=-1.0, subjectivity=1.0)	-1.0	Negative

Figure 15. Text of Vietnam part that contains the Top 3 negative words

Conclusion and Discussion

The main objective of this study was to offer a computerized method of text analysis to objectively examine the literary spatiality of the two countries of America and Vietnam as depicted in Viet Thanh Nguyen's The Sympathizer. To this end, LDA Mallet model and sentiment analysis by using TextBlob library were employed. Throughout the novel, the term America is repeated more than Vietnam. Even though both places are given a slightly positive polarity score in the novel, America as the diasporic character's hostland gains a higher positive score than Vietnam. While the latter is portrayed as a dystopian locality under the Communist regime, America is generally depicted as a utopian one that offers emancipation and a promising life. More specifically, the spatial analysis reveals the ways in which memories of loss, trauma, corruption war, and evil communism are attributed to the geographical location of the motherland Vietnam. The techniques used in this project allow the researcher to achieve the results faster with a higher level of accuracy and quality. The Lexicon-based approach also accomplishes a more robust performance than could be achieved by other approaches such as learn-based. However, the limitations listed in the previous section might lower the accuracy of the results. That said, a big and good quality of labelled data can also be employed to maximize the accuracy of the lexicon-based approach. Indeed, the lack of such data might lead to the formation of a poor model. As many of the Machine learning algorithms need enormous amounts of data in the process of designing a model, the

existence of good data guarantees the model to capture relevant and sufficient data from the training data set, and hence the better performance of the algorithms. The main limitation of the designed model in this project is that its performance relies on the performance of the library available in Python. For example, NLTK and spaCy libraries are used in Model 1 while NLTK and TextBlob library are used in Model 2. The POS tagging function of spaCy and the words with polarity scores available in TextBlob library are two of the main problems that affect the performance of the model.

Acknowledgements

The authors are thankful to School of Computer Sciences, Universiti Sains Malaysia and the lecturers of CDS590 for unlimited supports to finish this study. In addition, the authors are grateful to School of Humanities for financial support from Short Term Grant (304/PHUMANITI/6315300) granted to Dr Moussa Pourya Asl.

Conflict of interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

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/0.1109/ACCESS.2020.3029826

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.org/10.17576/3L-2018-2402-15

Asl, M. P. (2019). Leisure as a space of political practice in Middle East women life writings. *GEMA Online Journal of Language Studies*, *19*(3), 43-56. https://doi.org/10.17576/gema-2019-1903-03

Asl, M. P. (2020). The politics of space: Vietnam as a communist heterotopia in Viet Thanh Nguyen's The Refugees. *3L: Language, Linguistics, Literature*, *26*(1), 156-170. https://doi.org/10.17576/3L-2020-2601-11

Hadi, N. H. A., & Asl, M. P. (2021). The objectifying gaze: A Lacanian reading of Viet Thanh Nguyen's *The Refugees*. *GEMA Online® Journal of Language Studies*, 21(1), 62-75. https://doi.org/10.17576/gema-2021-2101-04

Keikhosrokiani, P. (2019). *Perspectives in the development of mobile medical information systems: Life cycle, management, methodological approach and application* (1st ed.). Academic Press. https://doi.org/10.1016/C2018-0-02485-8

Keikhosrokiani, P. (2020). Chapter 1 - Introduction to mobile medical information system (mMIS) development. In A. Press (Ed.), *Perspectives in the development of mobile medical information systems* (pp. 1-22). https://doi.org/10.1016/B978-0-12-817657-3.00001-8

Khan, K., Baharudin, B., Khan, A., & Ullah, A. (2014). Mining opinion components from unstructured reviews: A review. *Journal of King Saud University-Computer and Information Sciences*, *26*(3), 258-275. https://doi.org/10.1016/j.jksuci.2014.03.009

Kumari, K., Bhardwaj, M., & Sharma, S. (2020). OSEMN approach for real time data analysis. *International Journal of Engineering and Management Research*, *10*(2). https://doi.org/10.31033/ijemr.10.2.11

Lum, K. (2017). Limitations of mitigating judicial bias with machine learning. *Nature Human Behaviour*, *1*(7), 1-1. https://doi.org/10.1038/s41562-017-0141

Mohammad, S. M. (2016). 9 - Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion measurement* (1st ed., pp. 201-237). Elsevier. https://doi.org/10.1016/B978-0-08-100508-8.00009-6

Nalisnick, E. T., & Baird, H. S. (2013). Character-to-character sentiment analysis in Shakespeare's plays. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria.

Pourya Asl, M. (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.org/10.1386/jciaw_00010_1

Pourya Asl, M. (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.org/10.33806/ijaes2000.20.2.12

Queiroz, A. I., & Alves, D. (2015). Walking through the Revolution: A spatial reading of literary echoes. *JSSE - Journal of Social Science Education*, *14*. https://doi.org/10.4119/jsse-741

Roque, A. (2012). Towards a computational approach to literary text analysis. Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature, Montreal, Canada.

Schmidt, T., Kaindl, F., & Wolff, C. (2020). Distant reading of religious online communities: A case study for three religious forums on reddit. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, Riga, Latvia.

Tally Jr, R. T. (2017). *The Routledge handbook of literature and space* (1st ed.). Taylor & Francis. https://doi.org/10.4324/9781315745978

Van der Bergh, R. H. (2013). The contrasting structure of Acts 12: 5-17: A spatial reading. *HTS Teologiese Studies/Theological Studies*, *69*(1), 1-5. https://doi.org/10.4102/hts.v69i1.1313

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA.

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* (pp. 523-535). Springer International Publishing. https://doi.org/10.1007/978-3-030-70713-2_49

Bibliographic information of this paper for citing:

Yun Ying, Sea, Keikhosrokiani, Pantea, & Pourya Asl, Moussa (2022). Opinion Mining on Viet Thanh Nguyen's The Sympathizer Using Topic Modelling and Sentiment Analysis. *Journal of Information Technology Management*, Special Issue, 163-183.