Indonesian Journal of Electrical Engineering and Computer Science
Vol. 19, No. 1, July 2020, pp. 353~362
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v19.i1.pp353-362

LSA & LDA topic modeling classification: comparison study on E-books

Shaymaa H. Mohammed, Salam Al-augby
Department of Computer Science, Faculty of Computer Science and Mathamatics, University of Kufa, Iraq

Article Info

Article history:

Received Aug 5, 2019
Revised Nov 20, 2019
Accepted Dec 28, 2019

Keywords:

Latent Dirichlet Allocation
Latent Semantic Analysis
Text Classification
Text Clustering
Text Mining
Topic Modeling

ABSTRACT

With the rapid growth of information technology, the amount of unstructured text data in digital libraries is rapidly increased and has become a big challenge in analyzing, organizing and how to classify text automatically in E-research repository to get the benefit from them is the cornerstone. The manual categorization of text documents requires a lot of financial, human resources for management. In order to get so, topic modeling are used to classify documents. This paper addresses a comparison study on scientific unstructured text document classification (e-books) based on the full text where applying the most popular topic modeling approach (LDA, LSA) to cluster the words into a set of topics as important keywords for classification. Our dataset consists of (100) books contain about 1 million words based on full text. In the used topic models (LSA, LDA) each word in the corpus of vocabulary is connected with one or more topics with a probability, as estimated by the model. Many (LDA, LSA) models were built with different values of coherence and pick the one that produces the highest coherence value. The result of this paper showed that LDA has better results than LSA and the best results obtained from the LDA method was (0.54846) of coherence value when the number of topics was 20 while LSA coherence value was (0.4047).

Copyright © 2020 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Salam Al-augby,
Department of Computer Science,
Faculty of Computer Science and Mathematics,
University of Kufa, Iraq.
Email: salam.alaugby@uokufa.edu.iq

1. INTRODUCTION

With the increasing amount of textual data that we face in our lives every day and more information becomes available [1] hence, it becomes difficult to get what we are looking for. So, we need tools and techniques to organize, analyze, search, discover the hidden insights in any large group of textual data [2]. These methods are called topic modeling (unsupervised learning). Topic modeling is a new powerful technique for automatic classification of document, unsupervised analysis of big document groups and understand vast quantities of information in any large group from unstructured textual data in addition to summarize large collections of textual information [3]. Topic modeling has a significant role and useful in digital libraries for creating additional metadata [4] by providing a simple way to analyze huge volumes of unlabeled text and indicate the hidden relationships between items as well as topics expressed in titles. Topic modeling is used for processing, and classifying text efficiently and perfectly. The e-sources classification is very important for both users of the digital library and the librarians. For library users, it facilitates the process of accessing the required documents by collecting documents with similar topics together, in addition to enabling access to them through a number of keywords. On the other hand, the classification of texts helps the libraries to classify the new documents within the appropriate groups with higher accuracy and less computational time by building a classification model with a training on a large number of classified documents that can deal with any new

document. As an example of large e-repository library is the Library of Congress which is one of the largest paper libraries in the world with 29 million books (New World Encyclopedia, 2018) while Google is looking forward to create the largest digital library on the Internet. In comparison with 20 million books in the Internet Archive project, the number of books scanned and archived in the Google Book project reach to 25 million books within 2015 [5]. Google was planned to scan all the books that have been published before which means approximately 129 million books [6]. The operations of searching and accessing to these textual data such as (books, documents, and research) are done through information retrieval operations, while the extraction and discovery of information from these data require more complex techniques such as text mining. For example, within academic articles text mining can provide the extraction of the interested information from big amounts of contents [7] by training how to extract information from each article. Many of the researchers used the topic modeling methods to classify the unstructured text on a pre-prepared database where they applied the topic model in many fields such as (articles, news, twitter, movies and extracting information from medical images), but a little of the researchers used text with full-content. In this paper topic model are used for a comparison study on E-books classification with their full content on a database created by the researchers in order to get the keywords that help us in determining the subject.

2. RESEARCH METHOD

Recently, many researchers have worked in the field of text classification using topic modeling methods. In [8] Deerwester, et al. introduced the LDA model. This model is a probabilistic three-level hierarchical Bayesian model for big sets of discrete data and tries to obtain short descriptions for the group to process a big group of documents and to provide beneficial inferential machinery in fields involving multiple levels of structure for essential tasks such as classification, summarization, and similarity and judgments [2] were from the first to employ a topic modeling approach in their analysis of all scientific disciplines using on 30,000 articles from the journal Science.

The study in [9] used In-house developed software for implementing the LDA model for analysis tiny document group (62 documents) of health-related issues to get an overview of the kinds of health information that are labeled by the documents in the related corpus, and also to get a list of documents concerning to the scope of mental health [10] utilized topic modeling (LDA) and SVM methods in clinical reports for analyzing the classification of CT imaging reports into binary classes which show the system ability for effectively and interpretable representation of them also the model was appropriate in reducing the dimensional. This study showed improvement for datasets with equal class distribution over baseline approaches.

Bergamaschi and L. Po implemented in [11] (Latent Semantic Allocation (LSA), Latent Dirichlet Allocation (LDA)) and combined within the recommendation system for the database consist of two hundred thousand movies for evaluation of similarity in the plot of a video that was viewed, the results showed that LSA is better than LDA in supporting the proposal of like plots for analyzing textual information, assist users in determining information on the web and finding hidden semantic relationships between web elements.

In [12] there is a utilization of the text mining with a probabilistic topic model Latent Dirichlet allocation of two (Wikipedia articles and users' tweets) in order to solve, find, extract, and recommending articles in Wikipedia and analysis the Twitter users' interest. This paper used sample relatively small and ignore the pictures users posted but it was good tool for social and business research. In [13] the Latent Dirichlet Allocation (LDA) model was applied and Softmax Regression with topic vector for news text for classification a real news text. The results of text classification were suitable and good to minimize features dimension but there are some weaknesses in that model such as parameters used for topic model and the size of news text in addition there are some deficiencies in the proposed model such as the selection of the parameters of the topic model and size of news text.

Rajasundari et al employed three different machine learning methods (Naive Bayes, K-NN and K-means) and topic modeling techniques Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) for BBC news dataset, where topic modeling demonstrated its ability to detect hidden topics and bond between words and documents in addition to work better with many probability distributions [14]. LDA (Latent Dirichlet Allocation) was used by Mouhoub & Al Helal in [15] and bigram for obtaining the core topics of Bangla language news corpus and classifying news, where it proposed the first ever topic modeling tool for Bangla and this core is a challenge because the research in Bangla is not repeated due to insufficient datasets, unorganized grammar rules.

In [16] Kurata et al used (LDA) for examining topics in a library by making analysis of 1.648 full-text articles from 2000 to 2002 and (1.087) articles from (2015 to 2017). The articles was from five journals where specified 30 identified topics based on the highest 10 highly weighted terms for each subject, title, and body of articles.

3. THEORETICAL BACKGROUND

Dealing with unstructured data need to use Text Mining, which also known as text data mining [17]. Text mining is defined as "the process of deriving high-quality information from the text" [18], or it represents the extraction of hidden, valuable, and non-trivial patterns from unstructured text documents [19]. From other point of view it can be considered as an extension of data mining) [20, 21]). The text exploration process includes many functions such as (cleaning up unstructured data to be available for text analytics, text classification, text clustering, keyword extraction, document summarization, and entity relationship model in [22]. For the purpose of classifying texts, Text Classification is used which is a supervised machine learning technique, and considered as one of the basic tasks in Natural Language Processing (NLP) and used for a broad category of tasks such as (Sentiment Analysis, Topic Detection).

Text Classification is a significant part of text mining [23] utilized in a large number of applications in different domains like (image processing, document organization, medical analysis). The goal of Text Classification is to assign predefined categories or to classify sentences or textual documents in one or more defined categories. For example, new articles can be classified by topics, text classification can be performed in two various ways: manual and automatic classification. Manual classification depended on a human to classify text (provide quality results but it is time-consuming and costly), while the automatic classification uses (machine learning, natural language processing, and another technique) to automatically classify text in a quicker and higher cost-effective way. There are many Text Classification Algorithms such as (K nearest-neighbor algorithm, Rocchio algorithm, Naïve Bayes classifier, Support Vector Machines, decision tree, and rule learning). The work in the Text Classification of the unstructured text includes three stages after data preparation they are text preprocessing, topic modeling, and evaluation.

3.1. Text preprocessing

Preprocessing is the first step in text mining. In text mining techniques pre-processing acts a significant role [14] for transferring text from human language to machine-readable format. The preprocessing stage is important for structure the unstructured text and keep the keywords which are useful to represent the category of text topics [24]. Natural language text can contains many of words with no specific meaning, such as prepositions, pronouns, etc. So, after a text is obtained the preprocessing process consists of two steps.

3.1.1. Text cleaning:

In order to simplify the text data, clean data and reduce noise. Text cleaning includes:

a) Text normalization includes (converting all letters to lower case, removing all numbers, removing signs, removing symbols, removing non-English letters, removing particular words or letters which are not useful [18] and removing punctuation they didn't add to the meaning of the text).
b) Tokenization (is the process of splitting the text into sentences and the sentences into smaller pieces called (tokens).
c) Removing Words that have fewer than 3 characters which do not give an important sense in a sentence.
d) Removing stop words or frequent words such as" the"," is", etc. that do not have specific semantic
e) Stemming: Is the process of minimizing the number of words by retrieving its root and deleting inflection through dropping unnecessary characters, usually a suffix.

3.1.2. Re-configuration

This step is necessary to convert text data to an appropriate format for automated processing.One method to do this is a bag of words matrix representation or called (corpus) where each document represent a vector of tokens and the entries in this matrix represent the number of times a word found in a document [14].

3.2. Topic modeling

Can be defined as a type of statistical model [25] for discovering hidden topical patterns that occur in a set of documents through machine learning. In NLP, topic models are generative models which provide a probabilistic framework can be described as a method for finding a group of words ([26, 12]),where topic modelling is assumed that the document can be interpreted as a mix of subjects and that each subject consists of a set of frequently occurring words and can obtain the topics by linking words with similar meanings and distinguish between the use of words with multiple meanings through the discovery of words that help to determine the boundaries between subjects or find patterns of data that can be used to achieve the conclusion and the final decision. Many techniques can be used to obtain topic models such as (Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM) etc.…).

3.2.1. Latent semantic analysis (LSA)

Is one of the foundational techniques in topic modeling and NLP of analyzing relationships between a set of documents and the terms they contain or it can be defined as a model for extracting and representing the contextual-usage meaning of words [27] to compute the similarity between words, sentences, or whole documents [28] that can be achieved by producing a set of concepts related to the documents and terms, where LSA assumes that words will occur in similar pieces of text if they have a similar meaning and the objective is reducing dimension for classification or noise reducing technique [8]. LSA attempts to leverage the context around the words to capture the hidden concepts, also known as topics using Singular Value Decomposition (SVD).

3.2.2 Latent dirichlet allocation (LDA)

Is an unsupervised machine learning technique used to recognize the latent topic structure of textual documents [29] or used in the information retrieval field, document modeling and classification. LDA is one of the most popular probabilistic text modeling techniques in machine learning [12]. It is like probabilistic latent semantic analysis (pLSA), but LDA uses Bayes estimation instead of maximum likelihood estimation [30]. LDA overcomes all the drawbacks of LSA and PLSA model. LDA assumes each document in a corpus is a random mixture over latent topics, and each latent topic is characterized by a distribution over words [25]. And these latent topics can be generated from a collection of documents but the proportion of each topic in each document is different.

3.3. Evaluation

Topic Coherence measure is a metric generally used to evaluate topic models by measuring the degree of semantic similarity scores of the words in a topic. There are two measures in topic coherence (intrinsic measure (UMass), extrinsic measure (UCI)). Both measure calculates the sum of pairwise scores on the words. The high value of topic coherence score model will represent a good topic model [30].

$$coherence(V) = \sum_{(w_i,w_j)\in V} score(w_i, w_j, \varepsilon) \quad (1)$$

where (V is a collection of the word describing the topic, ϵ refers to a smoothing factor (original = 1.))

Intrinsic Measure is represented as UMass. It measures to compare a word alone to the preceding and succeeding words respectively.

$$score_{UMass}(w_i, w_j) = \log \frac{D(w_i,w_j)+1}{D(w_i)} \quad (2)$$

Extrinsic Measure is represented as UCI. In UCI measure, each single word is paired with every other single word. it uses point wise mutual information (PM) [30].

$$score_{UCI}(w_i, w_j) = \log \frac{p(w_i,w_j)+1}{p(w_i)p(w_j)} \quad (3)$$

For evaluation, in this work we will use Topic Coherence measure using (1), (2) and (3) to evaluate the topic modeling methods.

4. METHODOLOGY

In this section, we tried to illustrate the models used in our experiment and their methods, which are illustrated below:

4.1. Data preprocessing:

The data are collected from the library called (bookboon), 100 text files such as (books and articles) are chosen randomly with its full content, after reading and extraction the unstructured text from text files, the Data Pre-processing was applied. Data preprocessing consists of six tasks.

a) Normalization In this step, the text was transformed into a single basic format or a more uniform sequence by converting the characters to lowercase, deleting all numbers, symbols, removing particular words or letters which are not useful and punctuation such as (commas, quotes, question marks, and apostrophes).This step is important in order to shrink the size of the vocabulary.
b) Tokenization In this step, the given text splitting into smaller parts called (sentences), and the sentences into smaller pieces called tokens. Tokens are separated by whitespace or line breaks.

c) Removing Words that have fewer than 3 characters which do not carry important meaning in a sentence such as hm,at,ab,cc,er,ww,zc,nm,and ect.
d) Remove stop words in this step, we remove all English words or most common words in the English language which does not add much meaning to the sentence such as the, he, have, a, on, is, that,the,it and etc.
e) Stemming In this step the words of a sentence are converted to its non-changing parts or reduction of words into their stem for reducing the text data and improving the system performance, for example, two or more words have a common root such as (amusing, amusement, and amused) the stem would be amus.
f) Building Corpus: To speed up the processing, a dictionary was created to assign an integer ID to each unique word kind in the sets. After that, a corpus was created to train the topic model. In the corpus, each document is represented by a sequence of number of pairs. The first digit in the pair expresses the integer ID refers to a word and the second digit in the pair denotes how often that word occurs. For example [(1, 1)] where "1" refers to the word "Friendship" (for example) and "1" refers to the number of times the word occurs in the document. This step is depended on applying of the two mention methods from the topic modeling methods (LSA, LDA) on the corpus (Building Corpus) produced from preprocessed data to train the model.

In Figure 1 all the steps are performing for the swame sentence in Example. the steps 1 illustrates the text normalization step (converting all letters to lower case) on the text, while step (2) refers to text normalization steps (Numbers deleting), step (3) represents removing particular words or letters which are not useful , then step (4) is punctuation and symbols removal, step (5) refers to perform (Tokenization step), step (6) remove stop words removing words that have fewer than 3 characters, the last but not the least step (7) and (8) refers to Stemming step for the same sentence of the example.

Example: 1-Friendship is like a cloud that showers friends with a lot of goodness, love and loyalty, \t and ;@#,!" #hm{ **1** 1-friendship is like a cloud that showers friends with a lot of goodness, love and loyalty, \t and ;@#,!" #hm{
2 -friendship is like a cloud that showers friends with a lot of goodness, love and loyalty,\t and ;@#,!" #hm{
3 -friendship is like a cloud that showers friends with a lot of goodness, love and loyalty, and ;@#,!" #hm{
4 friendship is like a cloud that showers friends with a lot of goodness love and loyalty hm
5 "friendship","is"," like","a ","cloud"," that"," showers"," friends"," with"," a"," lot"," of"," goodness"," love"," and ","loyalty ","hm,"
6 "friendship"," like","cloud "," showers"," friends"," lot","goodness"," love ","loyalty ","hm"
7 "friendship"," like","cloud "," showers"," friends"," lot","goodness"," love ","loyalty "
8 "friendship"," like","cloud","shower"," friend "," lot","good"," love","loyalti"

Figure 1. The basic steps of text preprocessing

4.2. Topic modeling

As illustrated before there many topic modelling techniques, in this paper the researchers compare two techniques as followings:

a) Latent Semantic Analysis (LSA) is a model for finding hidden concepts, selecting and describing the contextual-usage meaning of words applied to a large corpus of documents data. LSA uses a bag of the word (Y), where rows represent terms and columns represent documents and the value of a cell represents (occurrence of terms in a document). LSA learns topics by making a matrix decomposition on a matrix of the document term using Singular Value Decomposition (SVD).
b) Latent Dirichlet Allocation (LDA) is a generative probabilistic topic model for a presenting corpus assumes that documents are arandom mixture of latent topics, where each topic is distinguished by a distribution of words. LDA depends on important parameters that must be considered when applying it:
   1. Number of topics to optimize results by knowing an optimum amount of topics.
   2. Beta: represents a topic word density (Topic concentration), it assumes that the topic is made of up most of the words and result in a more specific word distribution per topic. A high beta value means each topic is more likely to contain a specific word mix and in practice, that leads to topics being more alike in terms of what words they include and the lower value of beta, means they are composed of few words.
   3. Alpha represents a document topic density (Document concentration). A high alpha value point to that every document is tend to contain a mixture from the most of the topics, and not any single topic especially. The lower the value of alpha, means that the documents contain fewer topics.

4.3. Evaluation

Many measurements are utilized for evaluating performance process of the topic model methods such as Topic Coherence measure. In this work Topic Coherence measure are used using (1), (2) and (3) (as shown in above) to evaluate the topic modeling methods.

5. RESULTS AND DISCUSSION

The experiment was performed on the sample a special database consisting of 100 randomly selected books. The pre-processing operations were performed, which included cleaning the text from the characters and numbers, removing the stop signs and special symbols, removing the insignificant words, then reconfiguration data for representation unstructured data in a bag of words (corpus). The second step is training the data by using two methods from topic modeling (LSA, LDA) method with different number of topics (10, 15, 20) in order to see which one will give the best performance from these two methods with our database. The keywords for each topic are used as features to classify the books. The next step is using topic coherence measures (Umaas, UCI) to evaluate the topic modelling methods.

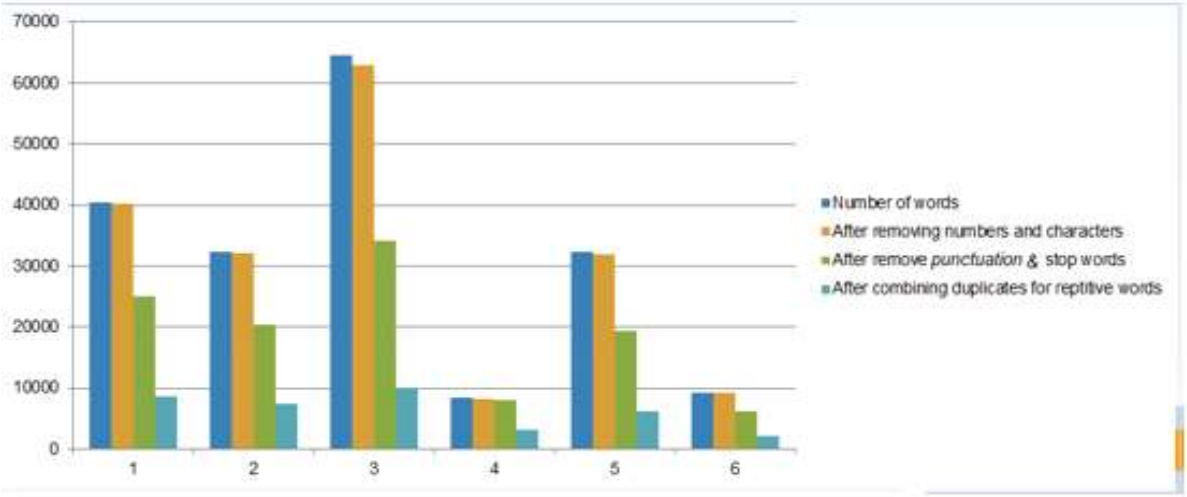
During our work, we got the following results. The results shows that the number of words in the documents have greatly reduced by up to approximately 80% in the pre-processing and retaining the desired words. As shown below in Table 1 and Figure 2. .

In Table 1 for example, the number of words was (64504) in the document (No. 3) and after applying the step of removing the numbers and symbols, the words were reduced by about 1% and a number of words became (62986). After the step of removing (punctuation & stop words), the number of words was reduced by about (40%) where the number of words in the document became (34124) and after combining duplicates for repetitive words in one document, the number of words became about (10043) the words reduced up to (80%) where the time of execution is (1588.0066668987274 seconds). This is very necessary to reduce the number of dimensions in dataset which used for training the data where the number of features became about 67903 for all documents.

Table 1. Results of pre-processing for (5) files (books) from (100)files

Docs	Number of words	After removing numbers & characters	After remove Punctuation & stop words	After combining duplicates for reptitive words
1	40455	40256	24968	8672
2	32388	32127	20352	7394
3	64504	62986	34124	10043
4	8378	8152	8046	3259
5	32401	31902	19295	6108

Source: Our own evaluation.



Source: Our own evaluation.

Figure 2. Graph of pre-processing step for sample of the results

(Corpus, dictionary, and a number of topics) are needed to train the (LDA and LSA) model, where each word in the corpus of vocabulary is then connected with one or more topics with a probability, as estimated by the model. (LDA, LSA) model is built with (10, 15, 20) various topics where each topic is a mixture of keywords and each keyword contributes a certain weight to the topic.

The input to topic model (LDA, LSA) is the number of topics that have to be discovered (k) and the (document-word matrix) which has the histograms of words (word count) presented in each document. The dimensions of this matrix are (D, W) i.e. a number of documents * a number of terms in vocabulary. The output of topic modeling (LDA, LSA) algorithms is two matrices (document - topic matrix) and a (topic -word matrix). Document-Topic matrix is of (D, K) dimensions where D is a number of documents and K is the number of topics in the vocabulary present the probability distribution of the topics being in the documents are shown below in Table 2. Similarly, Topic-Words matrix is of (K, W) where W is the number of words in the vocabulary present the probability distribution of words that they have been generated from those topics are shown below in Table 3, and Table 4, respectively. Topic models learn topics typically and represented as collections of important words automatically from unlabeled documents in an unsupervised way for getting the mixture of alike words together, thus molding the topic.

The main goal of a topic modelling is to provide interpretable document representations that can be used to explore the topics in a group of unlabeled documents. Table 2 shows an example of an interpretable document representation (Each document is made up of some topic distribution) for example document (0) is (0.0160618) of topic (6), (0.976266) of topic (8), and (0) to other topics and in Table 2 it is noted also results in documents vectors consist of a lot of zeros, that means there are only a limited number of topics occur per document and that corresponds with the concept that documents typically only speak around a limited number of topics. These results help in improving the human interpretability for documents vectors. For example document (11) is (0.999947) of the topic (2) and (zeros) of all other topics from that we can conclude that topic (2) is assigned for document (11).

Tables 3, and 4 show sample of calculation of 10 topics by using LSA and LDA models as well as the words related to each topic where top ten terms are listed for each topic. Each topic is connected with one or more documents in the group with a given mixing proportion based on the occurrences of words per document. The words in one topic tend to be similar for example in Table 4 in topic 9 it is noted that the words (drug, agent, rate, concentrate, use, reaction, and dose) tend to be similar and the label by each topic may be given by the analyst via evaluating the words allotted to the particular topic for example in Table 4 topic 9 can be assigned to the medicine based on evaluating the words.

In topic model (LDA, LSA) models, each document is a collection of multiple topics. But, typically only one of the topics is dominant. Table 5 and Table 6: shows assigning the document to the topic that has the highest weight in that document. In Table 5 there is a number of topics dominants on more than one document for example topic 7 was the dominant topic on document number (1, 6, 7), while in Table 6 it is noted that the existence of a single topic dominant in more than one document approximate (14 doc) for example topic 0 was dominant on document number (1, 2, 3, 4, 5, 6, 7, 9,…).

Table 2. Topics Distribution over each document (10 from100) docs, number of topic =10 using LDA

Doc index	Topic0	Topic1	Topic2	Topic3	Topic 4	Topic 5	Topic6	Topic 7	Topic 8	Topic9
0	0	0	0	0	0	0	0.0160618	0	0.976273	0
1	0	0	0.201245	0	0	0	0	0.794209	0	0
2	0	0	0.12514	0.610722	0	0.0217977	0.0214596	0	0.191833	0
3	0	0	0	0	0.999925	0	0	0	0	0
4	0.0176452	0	0.024141	0.796033	0	0.111023	0	0	0.0510887	0
5	0	0	0.98521	0	0	0	0.0147375	0	0	0
6	0	0	0	0	0.0153877	0	0	0.995476	0	0
7	0.0207129	0	0.0924715	0	0	0	0	0.871369	0	0
8	0	0	0.999945	0	0.664881	0	0	0	0	0
9	0	0	0.0180457	0	0	0	0	0.290758	0.0187322	0

Source: Our result using spyder (python 3.6)

Table 3. Topic terms with a probability (With LDA training) of a 10 topic run on document set (100 documents)

Topic 0	0.017*"use" + 0.009*"image" + 0.009*"oper" + 0.008*"system" + 0.008*"example" + 0.007*"one" + 0.007*"signal" + 0.007*"value" + 0.007*"number" + 0.006*"relat"
Topic 1	0.009*"use" + 0.006*"search" + 0.006*"number" + 0.006*"agent" + 0.006*"example" + 0.005*"data" + 0.005*"research" + 0.005*"knowledg" + 0.005*"set" + 0.005*"value"
Topic 9	0.014*"use" + 0.006*"sentenc" + 0.006*"word" + 0.006*"english" + 0.005*"one" + 0.005*"see" + 0.005*"time" + 0.005*"mean" + 0.004*"follow" + 0.004*"articl"

Source: Our own evaluation.

Table 4. Topic terms with a probability (With LSA training) of a 10 topic run on document set (100 documents)

Topic 0	0.316*"use" + 0.200*"one" + 0.144*"market" + 0.137*"time" + 0.126*"product" + 0.119*"example" + 0.119*"also" + 0.112*"two" + 0.102*"function" + 0.098*"system"
Topic 1	-0.504*"eng" + -0.476*"chem" + -0.285*"die" + 0.239*"market" + -0.210*"der" + 0.151*"bank" + -0.148*"da" + -0.118*"acronym" + -0.107*"pharm" + -0.092*"referr"
Topic 9	0.319*"drug" + -0.239*"agent" + 0.216*"rate" + 0.207*" concentrate " + 0.196*"use" + -0.191*"contract" + -0.181*"search" + -0.140*"knowledge" + 0.129*"reaction" + 0.126*"dose"

Source: Our own evaluation.

Table 5. Dominant topic that has the highest percentage contribution in that document (LDA)

Doc index	Dominant_topic	Topic_prec_contril	Keywords
0	8	0.9762	one, brain,function,speci, theori,time,studi,area,also,human
1	7	0.7942	eng , chem,use,die ,laccas,forc,der,object, process,da
2	3	0.6107	contract, cell , parti ,cancer ,compound ,term ,water ,answer,question,exampl
3	4	0.9999	nerv, muscl ,arteri ,fig ,ligament ,joint ,posterior, superior, anterior, later
4	3	0.7960	contract, cell , parti ,cancer ,compound ,term ,water ,answer,question,exampl
5	2	0.9852	equat, use,soiut ,two,reaction,rate, caicul,one,concenter,time
6	7	0.9954	eng, chem,use,die ,laccas,forc,der,object, process,da
7	7	0.8713	eng, chem,use,die ,laccas,forc,der,object, process,da
8	2	0.9999	equat, use,soiut ,two,reaction,rate, caicul,one,concenter,time
9	4	0.6648	nerv, muscl ,arteri ,fig ,ligament ,joint ,posterior, superior, anterior, later

Source: Our own evaluation.

Table 6. Dominant topic that has the highest percentage contribution in that document (LSA)

Doc index	Dominant _topic	Topic_prec	Keywords
0	0	447.273	use, one, market, time, product, example ,also, two, function, system
1	0	318.686	use, one, market, time, product, example ,also, two, function, system
2	0	414.565	use, one, market, time, product, example ,also, two, function, system
3	0	99.7946	use, one, market, time, product, example ,also, two, function, system
4	0	134.506	use, one, market, time, product, example ,also, two, function, system
5	0	575.053	use, one, market, time, product, example ,also, two, function, system
6	0	453.955	use, one, market, time, product, example ,also, two, function, system
7	0	218.581	use, one, market, time, product, example ,also, two, function, system
8	0	623.6	drug, agent, rate, concentr, use, contract, search, knowledge, reaction, dose
9	0	271.168	use, one, market, time, product, example, also, two, function, system

Source: Our result using spyder (python 3.6)

Topic models deliver no guaranty on the interpretability of their output, therefor topic coherence measures was used to the evaluation of (LDA, LSA). The Umass and UCL topic coherences capture the optimal number of topics by giving the interpretability of these topics a number called coherence score. In this work, many (LDA, LSA) models were built with different values of the number of topics (k) and pick the one that gives the highest coherence values as shown below in Table 7.

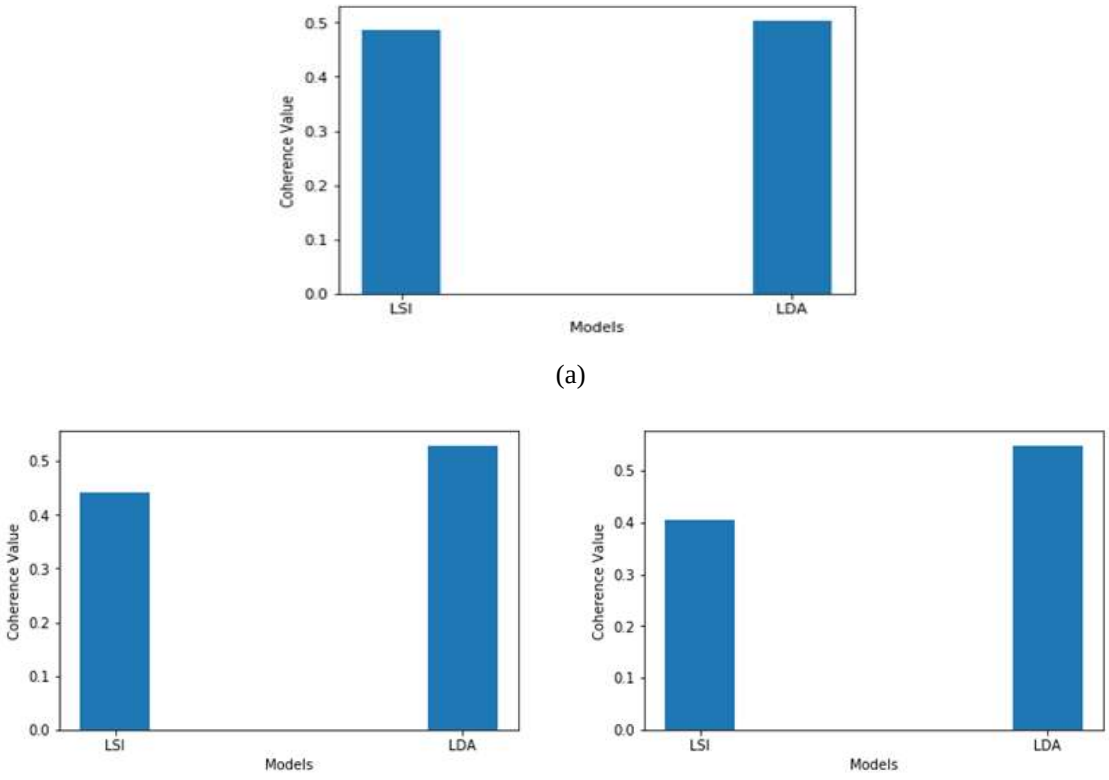
Table 7 shows that the coherence score (UCL) of LDA is increased with the increasing of the topic numbers from (0.5040) to (0.54846), with a decline of the value of Coherence UMass Score of (LDA) from (-0.5331) to (-0.5756) as compared to LSA the number of topics increases, while the value of a scale Coherence UCI Score will decrease from (0.4806) to (0.4047) with a decline of the value of scale Coherence UMass Score of LSA from (-0.5339) to (-0.7778). It is noticed that the best performance for LDA according to the Coherence UCI scale was when using a number of topics was 20 and the value reached to (0.54846) while the best performance to LSA according to the Coherence UCI scale was when the number of topics was 10 and the value reached to (0.4806). On the other hand, the best performance of LDA and LSA based on the Coherence Umass scale was when the number of topics was 20 the value was (-0.5756) and (-0.7778) to LDA and LSA respectively. As a conclusion, the coherence score (UCL) of LDA is higher than of LSA when increasing the number of topics while Coherence UMass Score (LSA) is declining more than Coherence UMass Score (LDA).Topic coherence gives a convenient measure to estimate how good a given topic model is. Based on the above results, topic coherence score, in particular, has been more helpful. The best coherence score was obtained from LDA.

Figure 3 (a) illustrates The coherence UCI scale value when number of topic 10 for (LDA, LSA), Figure 3 (b) refers to The coherence UCI scale value when number of topic 15 for (LDA, LSA), while Figure 3 (c) refers The coherence UCI scale value when number of topic 20 for (LDA, LSA).

Table 7. Topic coherence measures score

Number of Topic	Coherence UCI Score (LDA)	Coherence UCI Score (LSA)	Coherence Umass Score (LDA)	Coherence Umass Score (LSA)
10	0.5040833343374764	0.48064575748816346	-0.533159838315441	-0.5339049814458069
15	0.5290307935251595	0.43548343590104444	-0.5615390274802826	-0.6958054656610132
20	0.5484628906072918	0.4047114818513041	-0.5756349066981308	-0.7778111104144966

Source: Our own evaluation.



(a)

(b) (c)

Figure 3. The coherence UCI scale value (LDA, LSA)

6. CONCLUSION

In this paper the researchers tried to use two topic modeling techniques (LDA and LSA) for classifying the collected data according to its dominant topics and making a comparison between the results. Based on the results that got one can conclude the followings: (a) Classification plays an important role for both users of the library and the librarians by obtaining the required document and classify the new documents easily. Among many topic modeling techniques, LDA & LSA techniques was used to classify a large number of unstructured text documents. This work is a comparative study between two methods of topic modeling to classify e-books and to do so first keywords were identified because they play a major role in determining the topics for each subject. This work started with a number of pre-processing operations after that training the model by using (LSA, LDA) and finally evaluation the results was done by using Coherence value. These results show that the LDA technique gave better results than the LSA technique depending on the scale Coherence UCI with our dataset. (b) For both used techniques the pre-processing stage is essential stage because it provides good dimensionality reduction and remove unnecessary words from the unstructured textual data. The eliminated words does not significant effect and they may increased the dimensionality. (c) Choosing the number of topics still field dependent because, for example, the topic has good coherence scores but may have repeated keywords in the topic. (d) Based on the results LDA has better results than LSA in this work. (e) Topic Coherence measure can be considered a useful way to compare different topic modeling techniques according to their human-interpretability that leads to provide a clear view and hence take a good decisions. Whatever, the experiment results showed that both techniques (LDA & LSA) have limitations in their performance according to the used dataset and for future, there is a need to increase the size of dataset for better performance.

REFERENCES

[1] D. (Pew) Putthividhya, H. T. Attias, and S. Nagarajan, "Independent factor topic models," *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 833–840, June 2009.

[2] D. Blei and J. Lafferty, "Correlated topic models," *Advances in Neural Information Processinf Systems*, vol. 18, p. 147, January 2005.

[3] P. Anupriya and S. Karpagavalli, "LDA based topic modeling of journal abstracts," *2015 International Conference on Advanced Computing and Communication Systems*, Coimbatore, pp. 1-5, 2015.

[4] K. Hagedorn, D. Newman, and Y. Noh, "How Topic Modeling is Useful in Digital Libraries," 2010. [Online]. Avaible : https://www.lib.umich.edu/files/grants/topic/imls_topic_model_presentation.ppt.pdf.

[5] S. Heyman, "Google books: A complex and controversial experiment," *The New York Times*, 2015.

[6] J. Jackson, "Google- 129 Million Different Books Have Been Published," PC World, 2010. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2013GL058951.

[7] A. Kaur and D. Chopra, "Comparison of text mining tools," *2016 5th Int. Conf. Reliab. Infocom Technol. Optim.* (Trends Futur. Dir., pp. 186–192, 2016.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, September 1990.

[9] J. W. Uys, N. D. Du Preez, and E. W. Uys, "Leveraging unstructured information using topic modelling," in *PICMET '08 - 2008 Portland International Conference on Management of Engineering & Technology*, Cape Town, pp. 955-961, 2008.

[10] E. Sarioglu, K. Yadav, and H.-A. Choi, "Topic modeling based classification of clinical reports," in *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pp. 67–73, 2013.

[11] S. Bergamaschi and L. Po, "Comparing LDA and LSA topic models for content-based movie recommendation systems," *International Conference on Web Information Systems and Technologies*, vol. 226, pp. 247–263, 2014.

[12] Z. Tong and H. Zhang, "A Text Mining Research Based on LDA Topic Modelling," *The Sixth International Conference on Computer Science, Engineering and Information Technology*, pp. 201–210, 2016.

[13] Z. Li, W. Shang, and M. Yan, "News text classification model based on topic model," *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pp. 1–5, 2016.

[14] T. Rajasundari, P. Subathra, and P. Kumar, "Performance analysis of topic modeling algorithms for news articles," in *Journal of Advanced Research in Dynamical and Control Systems*, vol. 2017, no. 11, pp. 175-183, July 2017

[15] M. Mouhoub and M. Al Helal, "Topic Modelling in Bangla Language: An LDA Approach to Optimize Topics and News Classification," *Computer and Information Science*, vol. 11, no. 4, pp. 77–83, 2018.

[16] K. Kurata, Y. Miyata, E. Ishita, M. Yamamoto, F. Yang, and A. Iwase, "Analyzing library and information science full-text articles using a topic modeling approach," *Proc. Assoc. Inf. Sci. Technol.*, vol. 55, no. 1, pp. 847–848, 2018.

[17] M. A. Hearst, "Text data mining: Issues, techniques, and the relationship to information access," in *Presentation notes for UW/MS workshop on data mining*, vol. 1, p. 997, 1997.

[18] K. R. Bindu, L. Parameswaran, and K. V Soumya, "Performance evaluation of topic modelling algorithms with an application of Q & A dataset," *Int. Journal Appl. Engineering Res.*, vol. 10, pp. 23–27, 2015.

[19] Z. Zainol, M. T. H. Jaymes, and P. N. E. Nohuddin, "VisualUrText: A Text Analytics Tool for Unstructured Textual Data," *Journal of Physics Conference Series*, vol. 1018, no. 1, p. 12011, May 2018.

[20] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in knowledge discovery and data mining," *American Association for Artificial Intelligence445 Burgess Drive Menlo Park, CAUnited States*, 1996.

[21] E. Simoudis, "Reality check for data mining," in *IEEE Expert*, vol. 11, no. 5, pp. 26-33, Oct. 1996.

[22] S. V. Gaikwad, A. Chaugule, and P. Patil, "Text mining methods and techniques," *International Journal of Computer Application*, vol. 85, no. 17, 2014.

[23] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in *Mining text data*, pp. 163–222, January 2012.

[24] T. Gonçalves and P. Quaresma, "Evaluating preprocessing techniques in a text classification problem," *São Leopoldo, RS, Bras. SBC-Sociedade Brasilleira De Computacao*, pp. 841-850, 2005.

[25] C.-K. Yau, A. Porter, N. Newman, and A. Suominen, "Clustering scientific documents with topic modeling," *Scientometrics*, vol. 100, no. 3, pp. 767–786, May 2014.

[26] K. Hornik and B. Grün, "topicmodels: An R package for fitting topic models," *Journal of Statistical Software*, vol. 40, no. 13, pp. 1–30, 2011.

[27] K. K. Mino George, P. Beaulah Soundarabai, "Impact Of Topic Modelling Methods And Text Classification Techniques In Text Mining: A Survey," *International Journal of Advances in Electronics and Computer Science*, vol. 4, no. 3, pp. 72–77, March 2017.

[28] T. Cvitanic, B. Lee, H. I. Song, K. Fu, and D. Rosen, "Lda v. lsa: A comparison of two computational text analysis tools for the functional categorization of patents," *Int. Conf. Case-Based Reason.*, pp. 42-50, 2016.

[29] J. C. Campbell, A. Hindle, and E. Stroulia, "Latent Dirichlet allocation: extracting topics from software engineering data," *The Art and Science of Analyzing Software Data*, pp. 139–159, 2015.

[30] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 952–961, 2012.