Analysis of Domain-Independent Unsupervised Text Segmentation Using LDA Topic Modeling over Social Media Contents

Zeinab Shahbazi, Yung-Cheol Byun

Jeju National Uinversity
z.shahbazi72@gmail.com, yungcheolbyun@gmail.com

Abstract

Topic models generate the probability of words to analyze segregated data in available contents. In recent researches, that used conventional topic modeling on twitter posts, updated posts from same user combined into one single document and the reason of combination is texts information which is very short to conclude the topic relations properly. Similarly, this procedure can reduce the issue of sparsity and can't obtain the differences between topics which one user is posting. In this paper, the proposed system is a new topic modeling used to analyze short text contents which contain user information e.g., News, tweeter and Face book. Similarly, this model tokenized the above issue by applying user content clustering and text segmentation based on the topic assignments of words and each cluster in document have general topic relationship. Using huge data in this process by combining texts, various topic relations can generate from similar user. In this system, collection of texts which collected from same user combine into multiple topics. By applying combination of Latent Dirichlet Allocation and word2vec, number of clusters can automatically generated. Developed system is based on collapsed Gibbs sampling which can capture the evaluation between extracted clusters and topics at once. The advantage and effectiveness of this system presented with short text documents. In addition this model is capable to obtain the user interest dynamically by combining time distributions and also set of contents can cluster to get more better accuracy by applying side information's e.g., time information.

Keywords: *Short text document, Topic Modeling, Latent Dirichlet Allocation, Gibbs Sampling, Text Segmentation, Word2vec*

1. Introduction

Corpus is known as a collection of text. A corpus is a combination of multiple chunks of text which collectively called documents. The document can be an article, post on social media, library book or any spooky storybook. Each document contains a separate term that is common words. A word or words are treated as a single entity. Text corpus appears in text analysis, often referred to the bag-of-words (BOW) model, which involves the arrangement of words in a contents represented in terms of frequencies. The number of words in the corpus describes the dimensionality of the corpus. The main reason behind topic modeling is chunks of text shows different meaning topics and themes with several semantic to the viewer or readers. Topic modeling reduces the repetitive contents of text corpus to meaningful topics. However, document clustering is also used as a wide scope for unsupervised analysis of document whereas topic modeling aim is to find structure in the text through describing a text corpus by a lower-dimensional set of topics. Topic modeling is used to find the underlying themes, the meaning of a corpus to classify, annotate the corpus documents based on the topics which ultimately help in summarize, organize and search the records. Topic segmentation is another method which is directly related to topic modeling which aims to differentiate segments of topically coherent text within a single document.

The segmentation is refereed as the splitting a document into a segments [1]. The segment is also called a passage or segment boundary [1, 2]. In-text segmentation, splitting document can be helpful in many ways for text analysis. The main reason for text segmentation is to ease in handling the segments as they are more coherent and smaller than whole document [2]. Similarly, these segments can be used as a component of analysis and access. Text segmentation is also used to process text in information retrieval [3], sentiment mining [4, 5], opinion mining [6, 7], language detection [8], and emotion extraction [9]. Text segmentation is a significant task in natural language processing (NLP), e.g. knowledge discovery, information retrieval (IR) and text summarization [10] which is responsible for improving the readability of lengthy corpora of documents.

Moreover, a well written and structured tests which are logically organized in terms of topics and paragraphs are also getting benefit from text segmentation in at least two ways; (i) By extracting relevant chunks of the document to provide ease in readability and (ii) Provide the automated text segmentation system that insight to the author on the latent structure of content. Over the last several years, many researchers have been proposed in the field of text segmentation and topic modeling. Most of the methods are unsupervised which exploit information related to lexical chain, based on the fact that words which are related or similar tend to be frequent in topically coherent segments and segment boundaries often correspond to a change in the vocabulary [11, 12, 13]. These methods do not need training of data and can be applied directly to any text from any domain, where word boundaries can be identified. The main deficiency in these systems is that when the segment boundaries are correctly estimated, the segments are not labeled with any topic information.

Information retrieval is the method of finding the information from the documents based on queries, e.g., list of keywords etc. [14]. The primary and most commonly used model for automatically indexing documents and information retrieval is the classical vector space model (CVSM) which depend on term frequency-inverse document frequency (TF-IDF) structure of corpus [15]. Every term in the document has assigned a value based on the term frequency in the document which is multiple of inverse of the document frequency. Similarly comparing the query and TF-IDF indexes of a document using cosine similarity and evaluate the output of how far the document is in term of query efficiency [16, 17, 18]. However, there are many drawbacks of CVSM, e.g., synonymy and polysemy. Synonymy is the concept in which two different words have the same meaning in term of semantics, e.g., horror and scary. Similarly, polysemy is a single term having different semantic in term of context, e.g., lie to someone or lie on the ground. Therefore in automatic information retrieval, a system that does not follow polysemy and synonymy failed to respond user query, e.g., user query horror, and there is word fear instead of horror, so in this case, the system was unable to acknowledge the query [19, 20].

Latent Dirichlet allocation (LDA) is a generative probabilistic model used for corpus. LDA is an extended of probabilistic latent semantic indexing (PLSI) model used for the documents developed by Blei et al. [21]. In topic modeling, LDA is used as an industrial standard and commonly used approach. LDA maintain the structure of PLSI where every document is a combination of topics, and a combination of terms signifies the topics. The LDA development enables a developer to develop a system for complex topic models that contributes to the same term topic combination and document topic combination structure but protracted the abilities of the model [22, 23, 24].

PLSI topic distribution is replaced by LDA using a generative process, where θ is represented as a topic combination of a document drawn from a Dirichlet distribution parametrized by α. PLSI and LDA structure are similar in three ways; (i) Illustrate the N, which is denoted as terms in a document from a Poisson distribution. (ii) For every N term w_n, list the topic z_n of the term from the multinomial distribution of topics θ. Sample the term w_n with probability $p(w_n|z_n, \beta)$, which is based on topic z_n using multinomial distribution.(iii) list the multinomial topic distribution, which is denoted as θ from a Dirichlet distribution symbolized as α in a document.

In Figure 1, the M represents the documents and θ is the topic distribution of a document which is listed by α denotes Dirichlet parameter. There are N terms in every document which created by listing a topic z from θ and then listing the terms from the topics multinomial term distribution symbolized as β based on Z [21].

One of the essential parts in text summarization is results which show the process of summarization if it covers all critical themes which a document contains. This mentions the document as a progression of subtopics based on segments (e.g. words). Another important topic used in the text segmentation is IR that contains some substantial requirements, e.g., structure stories from broadcast news transcription, which can be beneficial for retrieval. Nowadays without applying text segmentation process, e.g., in news broadcast, we need to search the whole broadcast to get access to the specific story, but in another way, if the news is segmented and labeled, we can access to that particular story directly. Text segmentation is also used to segment topics and sub-topics through the document and display only the proper and related parts of the text through the search process.

Motivation

In this paper, we used Latent Dirichlet Allocation to generalize topics in short text documents. Developed system automatically evaluate the number of clusters for various users which post or comment in social media websites, and can deduct the presented model by using collapsed Gibbs sampling to capture the evaluation of clusters and topics in same time. The remaining of this paper is divided into five sections which are as following: Section 2 explains the related work on sentence topic segmentation based on individual sentences and unsupervised topic modeling related to the majority of state-of-the-art topic models. Section 3 gives the detailed system methodology, design architecture and overall transaction process of the proposed system, Section 4 elaborates on the implementations and result of the proposed method, and section 5 concludes the paper.

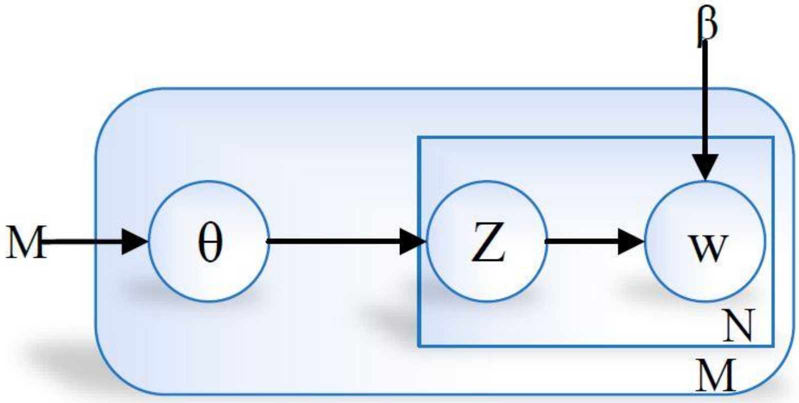


Figure 1. General diagram for creating process for LDA

2. Literature Review

Topic analysis divided into two basic tasks: topic identification and text segmentation. This process is beneficial in most text processing algorithms, likewise text automatic indexing, which is used in an information retrieval system. The goal of topic identification and text segmentation is to find differences between topics and sub-topics in documents.

Topic modeling presents a statistical way to recognize hidden parts of the document. In better explanation, it gathers words into trace clusters which are called topic [25]. One of the challenges in a massive collection of documents which is available as a large corpus of data and digital technology is, unique analysis amount of document, and by using the

cutting-edge technique from LDA, topic modeling and design document rely on latent topics, can measure as a pattern of the word [26].
One of the various study area related to topical passages is topic-based text segmentation or boundary detection. The purpose of this work is to compare the before and after of each potential boundary through text segments. Originally it divides the text into a range of three to five sentences, and the output of the process are vectors. The similarity between vectors is calculated by cosine similarity to recognize potential topic boundaries [27].

For performing automatic text, summarization Salton and Singhal [28] applied text structure knowledge by passage extraction. Within this method, they generate intra-document links by using the technique of inter-document link generation. On the other hand, text segmentation applied in spoken dialogs.

Authors in [29], proposed lexical chaining based on coarse-grained segmentation of CNN news transcripts. Automatic speech recognition system (ASR) introduced by Christensen and Kolluru [30] which is based on the Maximum Entropy model for lecture and topic segments. Furthermore, is an investigation of automatically predicting segment boundaries problem proposed by Huseh and Moore [31] in spoken multi-part dialogs by utilization of lexical cohesion-based model. Topic segmentation research area relevant to dialog domain has almost the same technique to text segmentation, but they are reconciled to spoken language features.

Sentence distance matrix presented by Ji and Zha in [32] which represents the cohesion in the information of sentences, where each entry corresponds to the similarity between a sentence pair and also adopt a programming technique to find the optimal topical boundaries. Their process continued by transforming matrix against a gray scale image and for increasing semantic cohesion sharpen topic boundaries they exploit a method of anisotropic diffusion.

Text regions identification average step is, passage retrieval, which is available in the QA system, and it includes the answer of user questions [33] that is also the goal of improving QA. Produce passage for QA is one of two fixed length or variable length. In another way, modified sentences passages may participate or fragment [34] and also produce by semantic clues such as sentence similarities, user questions relevant and conjunctions [35].

There are a lot of algorithms used in the text segmentation process, but some of them are based on, and identified topic that works identically to the word-based segmentation system. However, to calculate topic distribution through latent Dirichlet allocation (LDA) needs latent semantic analysis to map from high dimensional material to low dimensional semantic space and also text segmentation method should discover the relationship between sub-topical and, in next step make mapping between topics and segments [36, 37, 38]. Many existing works use LDA to find the document collection, but generally, the LDA model does not contain the document structure information and cannot reliable for extracting more information from text [39]. However, to achieve a topic modeling perspective in text segmentation task, there are some contributions as below:

- The topic modeling approach has more acceptable and major efficiency than standard baseline method, which shows if the training data is accessible or not.
- On-line document segmentation based on dynamic programming (DP) that reduce the computational cost related to the topic model.
- This process tested in 3 types of the dataset which is standard dataset and later more realistic and with more amount dataset to find out the issues in text segmentation function, at last, part of TRECVid 2003 evaluations dataset for ending this way [40].

Using topic ID in LDA and topic modeling parameters presents the repetitive words by applying Bayesian methodology and also using text segmentation for comparing the word-based process to display on three types of algorithms, i.e., TextTiling [41], C99 [42], and TopicTiling [43].

TopicTiling algorithm contrast the text blocks with bag-of-words vectors using topic ID and applies the LDA derivation method. Moreover, by using the hypothesis of relaxed bag-of-words, some topic models illustrate the higher qualitative and quantitative performance, by holding the words during the process, it brings forward overhead computational [43].

Table 2. Notation used in this paper

Authors	Objective	Limitation
Hyo Jung Oh, Sung Hyon Myaeng, Myung Gil Jang (2007) [44]	Describe the relationship of semantic documents as sentences and specify Topics out of certain sentences.	Difficult to recognize the topics by classification method.
Hemant Misra Francois Yvon Olivier Cappe Joemon Jose (2010) [45]	Survey of unsupervised topic modeling using latent Dirichlet allocation (LDA) and multinomial mixture (MM), to break the document into relevant parts.	High computation cost and not matching the train and test domain.
Martin Riedl Chris Biemann (2012) [46]	To do text segmentation using latent Dirichlet allocation and information retrieval system.	Automatically specify the number of segments and subject to adjust author Priority.
Martin Riedl Chris Biemann (2012) [47]	Propose specific topic ID using LDA in hidden documents and modify the efficiency of topics.	Finding an optimal setting for the Window parameter automatically, especially when the target segment is unknown.
Shoaib Jameel Wai Lam (2013) [48]	Model detection using unsupervised learning to obtain topics with various parts like segment topics or word topics.	Applying an optional type of topics to the system process.
Xuwqi Cheng Xiaohui Yan Ji-afeng Gue (2014) [49]	Short type of documents related to biterm topic model (BTM) which is immediately obtained words to patterns.	Scalability of the inference algorithms in internet webpages and expand the huge type of dataset.
Qi Gue Mark J.Gierl (2019) [50]	Proposing a systematic method based on students misapprehension to feasible choices and quotation from registered replies.	Difficult to attend different parts action with creating distractors in the test set.

2.1 Text Segmentation Domain-Dependent Process

2.1.1 Word-based Text Segmentation

Word-based segmentation is one of the issues in the text segmentation process. The main problem is dividing the Document content into words. There are a few ways for extracting word occurrence [51] to distribute paragraphs to multi-paragraph subtopics. Moreover, text segmentation process contains three steps named "paragraph tokenization," allocate similarity score," and "subtopic boundary detection [52]." Although word-based segmentation can eliminate synonyms between terms. In this paper word embedding used to verify the relationship between terms using cosine similarity.

2.1.2 Topic-based segmentation

The main goal of topic segmentation is to detect boundaries of topic blocks in the document. It is also a significant task in text semantic analysis. Topic segmentation technique has worked well in long document contents, but in short, text documents it's quite difficult to use this process because there is not enough information for extraction out of short document contents [53].

2.1.3 Neural network-based segmentation

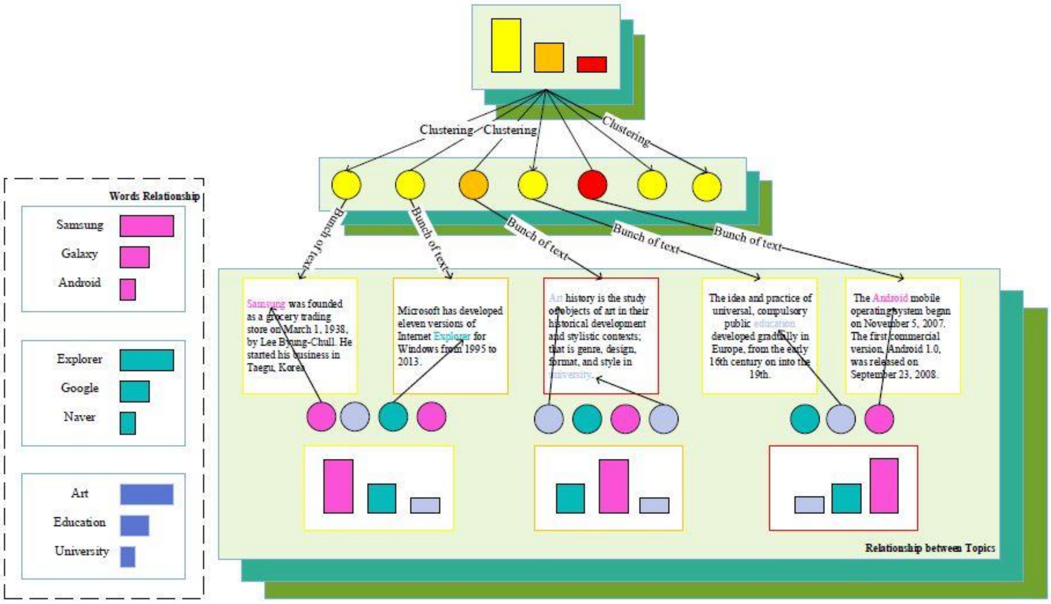
Recently, neural network system becomes a good idea to overcome the problem of sparsity and word representation. Neural network-based segmentation is named as word-embedding, and it is a real-value vector. Word embedding is enabled to find the relationship between words by calculating the distance between word vectors [54].

2.2 Text Segmentation Domain-Independent Process

Domain-dependent text segmentation has high performance in open source data type, but there is some opportunity in case of un-sufficient data and problems in evaluating document semantic similarities. To overcome these problems [55], word2vec model is demonstrated words semantic relationships based on vector space. Afterwards, dynamic programming extracts the segmentation boundaries and shows the similarity measurement. This process requires training process minimum amount and comparing with domain-dependent, the performance of domain-independent is insignificant.

3. Methodology

In this study, a semi-automatic process was proposed to analyze the social media contents of on-line sources. Collected contents are used to apply in topic modeling system architecture to find the similarity between topics. Moreover, topic modeling is to categorize contents into various parts using document stemming and semantic web standard extensions. The collected contents are related to comments or posts which one user uploaded. Each documents contains different type of words. Figure 2 show that each cluster connected to one document. Each cluster has a specific relationship with topics. The clusters that contribute with one document is just related to one user and it’s not assign with other user’s information. Here topic modeling is based on LDA system which already evaluate the number of clusters in each document. The chosen topics for each category is based on the relationship between words.



Samsung was founded as a grocery trading store on March 1, 1938, by Lee Byung-Chull. He started his business in Taegu, Korea.

Microsoft has developed eleven versions of Internet Explorer for Windows from 1995 to 2013.

Art history is the study of objects of art in their historical development and stylistic contexts; that is genre, design, format, and style in university.

The idea and practice of universal, compulsory public education developed gradually in Europe, from the early 16th century on into the 19th.

The Android mobile operating system began on November 5, 2007. The first commercial version, Android 1.0, was released on September 23, 2008.

Figure 2. System Architecture of Proposed Topic Modeling System

3.1 Notation

In Table 2 frequently used notations are summarized.

Table 2. Notation used in this paper

Name	Description
E_i	Word Embedding
M	Document Length
D_i	Vector Representation
T_i	LDA Topic Modeling
C_n	Candidate Blocks
P_d	True Segmentation
D_m	Distance Probability

3.2 Pre-processing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. In the proposed system we create the bag-of-words of each document depending on the text format and extract the repetitive parts from the text, and through that, remove all punctuations and stop words which in other word is cleaning a dataset. For the next step, data integration which is to combine the data to the meaning full parts used, and data transformation to convert all dataset in the same format and finally, data discretization to get the minimal loss of dataset that the process is shown in Figure 3.

ISSN: 2005-4238 IJAST

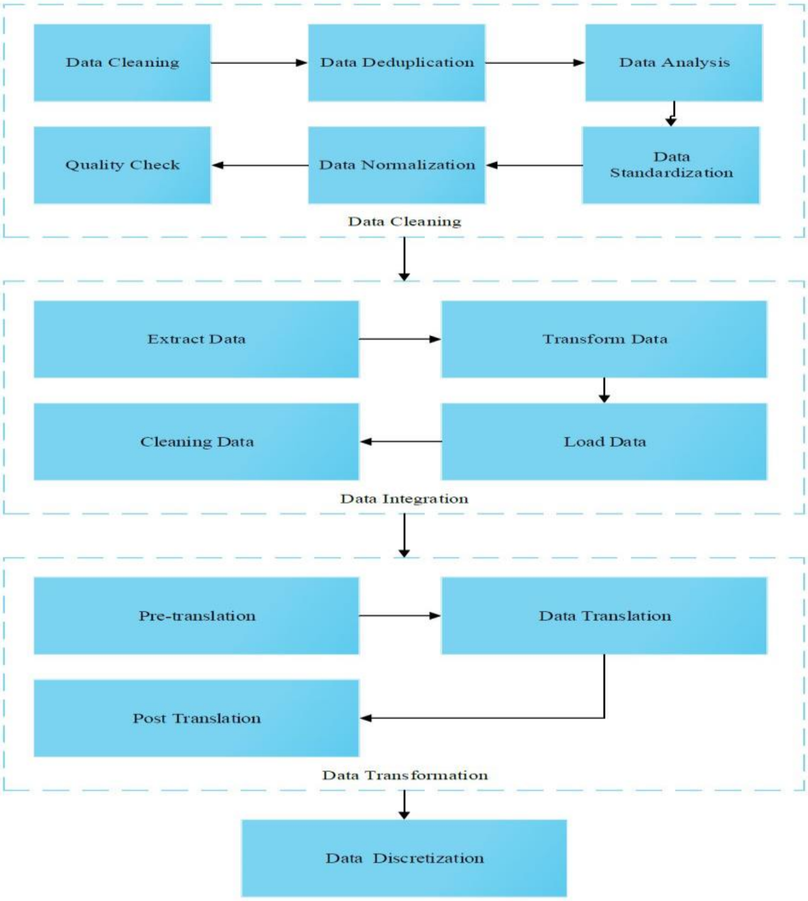


Figure 3. Data Pre-processing Structure

3.3 Input Social Media Contents

For this study, online text contents, e.g., face book, News and Tweeter, selected as a data source. Based on the user information data's divided into separate categories. The proficiency area related to these contents includes a wide range of detail and description. The filters and queries were operated to obtain information that only covered software and application development for big data. Table 3 describes the data analysis process in three different categories.

Table 3. Overview of Data Analysis Process

Data Collection
Specify search-query measures for online social media dataset Download relevant details of data Build a data category
Data Pre-processing
Transform the uppercase words to lowercase words Remove stop words, numbers, and punctuations Make a matrix of document terms Load the matrix of document terms
Data Analysis
Implement the LDA topic modeling system Define the stability of extracted topics Topic categorization

3.4 Text Segmentation Proposed System

The proposed study comprised of four modules, i.e., word input, features, content representation, pre-processing, and similarity measurement. In word input, the document dataset is processed using CoreNLP's MaxEnt sentence tokenizer. The CoreNLP's MaxEnt sentence tokenizer is responsible for extracting sentences from the documents. Afterwards, the output of the tokenizer is further processed for extracting words from the sentences using word extraction. Finally, the words are passed to the next module as an input. The feature module takes words as input and finds the similarities between words in term of meaning, and synonyms. The feature module further comprised of two sub-modules, i.e., word embedding, and LDA topic modelling. These modules are used to find the similarity of words in the document. The output of the feature module is a vector of words represented as $v_{1 \ldots n}$. The next module is the content representation, which is a combination of LDA and word embedding. The input of this module is word vectors, and output is processed words in term of similarity. The last module comprised of pre-processing, K-Mean and logistic regression calculation, cosine similarity, local minima identifier, candidate block, and unsupervised clustering. The pre-processing includes stemming, stop words, and tokenization, etc. After preprocessing the output is passed to the machine learning algorithm, i.e., K-Mean and logistic regression. The logistic regression is a statistical and logistic model is responsible for labelling data and describe the relationship among dependent binary variables and nominal variables, ratio level independent, ordinal, and interval variables.

Similarly, K-mean aims are to group the data using the vector quantization. K-mean used for unsupervised learning for unlabeled data. The labelled data is then passed to cosine similarity to calculate the similarity between words vector and the cosine of the angle between them. The computed percentage of similarity between words using cosine similarity is pass to local minima as an input to identify the sub-topical boundaries. The output of local minima identifier pass to candidate blocks which use the similarity score as a group of documents and merge to unsupervised clusters which contain same meaning segments and finally by applying text segmentation it automatically categorize text into coherent topics. The detailed process is shown in Figure 4.

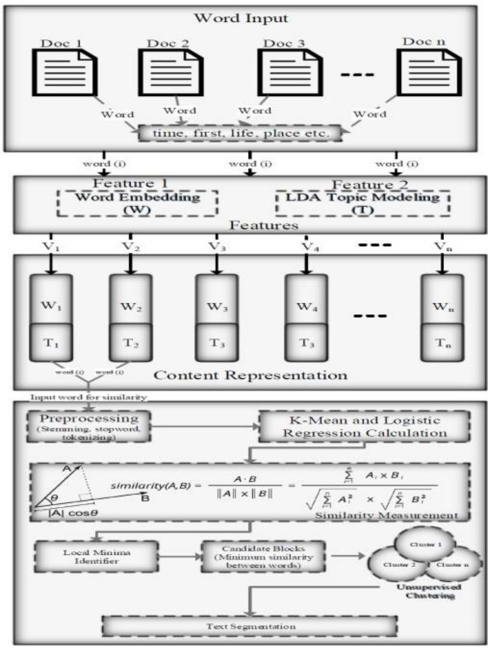


Figure 4. Visualization of Proposed Text Segmentation Structure

3.5 Content Representation

In this section, Feature engineering detail process needed to apply in data content Doc_n. Specially, vector representation is required for each tokenized part of text to quantify the data material. Proposed Glove embedding which used for word representation and topic models which manufacture vector representation e_i and t_i for each Doc_n. Subsequently, output of both process are combined as one vector $v_i = [g_i^{\wedge T} , t_i^{\wedge T}]^{\wedge T}$. Complete process presented in Figure 4.

3.5.1 Features

In topic generalization, two different features techniques which are very famous in similarity domain from which we choose word embedding and latent Dirichlet allocation. The main reason for selecting these features is to find the same meaning, repetitive, and hidden part of words which have similarity to others. These features are described in detail in below subsection.

1. Word embedding e_i: Word embedding e_i is one of the features used in the process of finding similarity between words because it contains a big dictionary of words which makes it easier to find the same meaning words through the dataset. Different type of methods is used to calculate the word embedding, e.g., TF-IDF, glove embedding, N-Gram, and prediction-based embedding. However, in the proposed system, we have used glove embedding, which is used to calculate the summation of each document d_1 to m and transform the document into one-dimensional segments. The reason for using Glove embedding is that it can provide the particular meaning of words in the proposed model without centralizing training process.

$$h_i = \frac{1}{M}\sum_{m=1}^{M}(E_{im}) \tag{1}$$

2. LDA topic modeling t_i: Latent Dirichlet allocation (LDA) is one of the famous generative statistical models uses in topic generalization permit to set of perceptions to illustrate by unobserved or hidden parts [56]. LDA is used to find the similarity between the subtopics. The output of word embedding is passed to LDA as an input for further processing to generalize the subtopics to one main topic. Figure 5 represents the detailed working of LDA.

   In the proposed methodology, LDA is used based on a case by case process, which means document separations and the number of generated topics with LDA process has the best performance and right segmentation in a large corpus of data. Similarly, short type of text data can contain just one single text or limited length of the document which make the segmentation process being unknown and LDA topic modeling is to optimize the value and also give optimal options for parameters.
3. Gibbs sampling: Is a Markov chain Monte Carlo algorithm (MCMC) that is to capture monitoring sequences that almost are from a particular probability distribution. Based on this process the topics which sampled through logical coalition out of topics and words relationship are defined as Equation 2.

   $$A(x, y, \theta, \text{Ø}|\alpha,\beta) = A(x|y, \text{Ø})* A(\text{Ø}|\beta)* A(y|\theta)* A(\theta|\alpha) \tag{2}$$

   A is defined as topic probability, y is presented as topics and words are presented by Ø. Equation 2 is representing the mixture of LDA and Gibbs sampling where topic y is sampled through logical coalition out of topic probability.

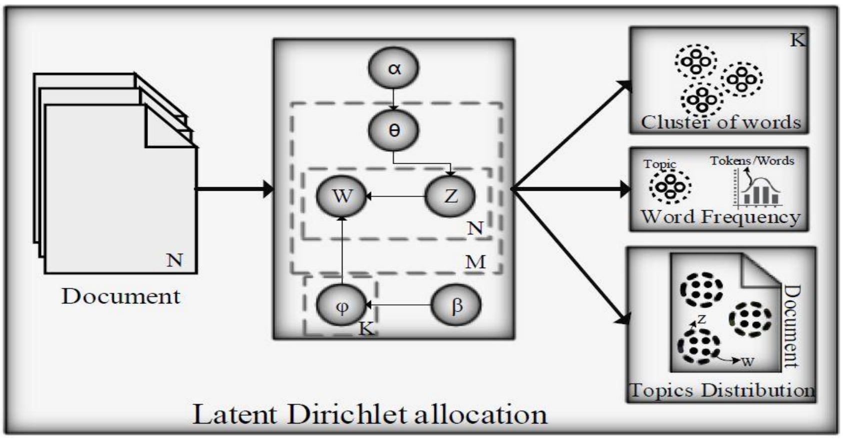


Figure 5. LDA Topic Modeling

3.5.2 Feature Combination

Output of natural language processing (NLP) e_i and t_i are combined and become as $v_i = [g_i^{\wedge T}, t_i^{\wedge T}]^{\wedge T}$. Figure 6 plots the combination of word2vec and LDA features in proposed system.

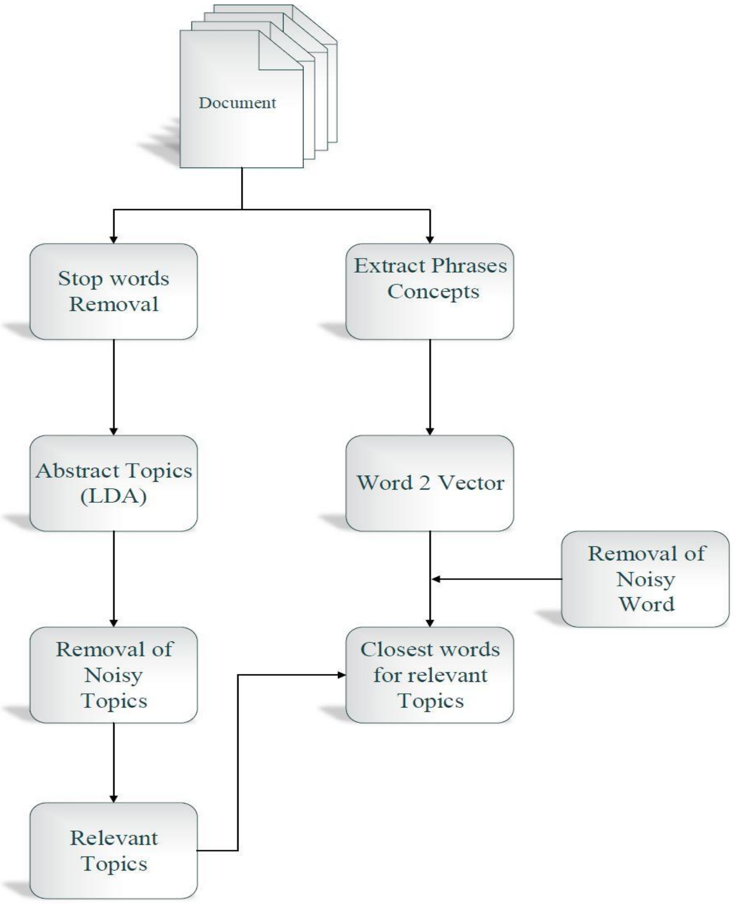


Figure 6. Combination of LDA and Word2vec Approach

3.6 Terms Aggregation

Further processing is to find the sub-tropical boundaries among text contents. To do this, First similarity between the outputs of v_i is evaluated to recognize differences between text distributions. Second is to recognize candidate blocks applying local minima identifier through text distributions. Noticed candidate blocks are eventually applied in unsupervised clustering. Finally, clustering between candidate blocks specifies document text segmentation.

3.6.1 Segment similarity ($C_{Similarity}$)

In sequence to locate potential break point between segments cosine similarity calculated for each contiguous segments based on vectors denotes as v_i. Domain-independent text segmentation algorithm also used this process to realize the approximate performance. Cosine similarity between two segments evaluated as:

$$Similarity = \cos(\theta)\frac{X.Y}{\| X \|\| Y \|} = \frac{\sum_n^{i=1} X_iY_i}{\sqrt{\sum_n^{i=1} X_i^2}\sqrt{\sum_n^{i=1} Y_i^2}} \quad (3)$$

$$C_{Distance} = \frac{\cos^{-1}(cosinesimilarity)}{\pi} \quad (4)$$

$$C_{Similarity}= \{1- C_{Distance}\} \quad (5)$$

It demonstrates the similarity between the immense value of two adjacent segments and local minima identifier. The considerable decrease of the similarity score and recognizes the sub-topical boundaries of two contiguous segments which are notably different in the textual relationship.

3.6.2 Candidate blocks

Candidate blocks are collection and organization of multiple similarity scores results out of the document, which is shown as c_1, c_2, c_n that each block leastwise one segment. The transactions are the output of document structure and show the correspondence that contains particular blocks. This process needs a breakpoint to exist that defines as local minima which explained in Figure 4. Local minima identifier is used as possible segmentation boundaries because it can recognize the maximum semantic shifts. This process is occurring in all local minima until all segments of Doc_n group to the candidate blocks.

3.6.3 Unsupervised Clustering

The last section of the proposed system is block clustering which in this system unsupervised clustering used through doing machine learning algorithms. Block clustering defined as $Cluster_n$. Extracted segments with the same meaning group into clusters. Unsupervised clustering is applied to purify text segmentation which ultimate boundaries output of individual candidate blocks. Clusters are contained a group of segments with the same meaning.

Specifically, the proposed system is famous to affinity propagation that doesn't need to specify the number of clusters. Affinity propagation captures the similarity between match data points. Real-valued messages are exchanged between data points until a final set of exemplars and corresponding clusters emerges [57].

4. Implementation and Experimental Results

In this section, we evaluate our text segmentation methodology, determine dataset and experimental settings.

4.1. Development Environment

The development environment of the proposed system is summarized in Table 4. All experiments and results of the system are carried out using Intel(R) Core(TM) i7-8700 CPU @3.20GHz 3.19 GHz processor with 32 GB memory. For sentence extraction, CoreNLP's MaxEnt sentence tokenizer is used. Combination of word embedding and latent Dirichlet allocation (LDA) features are used to find the relevant words and categorize them into related topics. Similarly, cosine similarity is used to calculate the similarity between sub-topics. The library and framework used in the proposed system is Jupiter notebook. The programming language used in the designing of this system is WinPython--3.6.2.

Table 4. Development Environment of Proposed Context-Independent Text Segmentation

Component	Description
Programming language	WinPython{3.6.2
Operating system	Windows 10 64bit
Browser	Google Chrome, opera
Library and framework	Jupyter notebook
CPU	Intel(R) Core(TM) i7-8700 CPU @3.20GHz 3.19 GHz
Memory	32GB
Machine learning algorithm	K-mean and Logistic Regression
Similarity Algorithm	Cosine Similarity
Distribution Modeling Algorithm	CoreNLP's MaxEnt
Similarity Features	Word embedding and latent Dirichlet allocation (LDA)

4.2. Dataset

In the proposed study, we used the social media contents dataset. Dataset contains texts and comments from "Tweeter", "Face book" and "News". Each user has a separate category based on the comments and posts uploaded in social media. CoreNLP MaxEnt sentences tokenizer used to create sentences out of chunking larger documents. Dataset grouping is shown in Table 5 with the average segment length and document size, which is effective to estimate the performance of system.

Table 5. Dataset Statistics

Data Type	Doc	Avg. Segment Length	Avg. Min Segment Length	Avg. Max Segment Length
Twitter	7900			
Face Book	6044	26.73	5	861
News	5635			

4.3. Evaluation Metrics

Presented dataset has pre-specified sentence segmentation. To evaluate the true segmentation (P_d) and WindowDiff (WD) metrics applied to measure out the true segmentation between proposed algorithm and applied segmentation process. True segmentation determine as hypothesized segmentation Hyp and evaluated segmentation determine as reference segmentation (Ref). P_d evaluated based on word movement size. d is defined as size of word to half of the true segmentation average size and collect the penalties among words in case of end process if it doesn't end with same segments. Equation 6 specify P_d as:

$$P_d\,(\mathrm{Hype},\mathrm{Ref}) = \Sigma_{x<=y<=n}\; D_m(x,y)\,(S_{Ref}(x,y) \oplus S_{Hyp}(x,y)) \qquad (6)$$

Where d defined as an average of the true segmentation, both S_{Ref} and S_{Hyp} are indicator functions in case of x and y sentences are part of true segmentation with the same segment. Distance probability defines as $D_m(x,y)$ to show every possible distance through randomly chosen documents.

WindowDiff (WD) represents the words fixed-size through the document and penalizes number of boundaries based on Hyp segmentation when the words are not match in true segmentation Ref. Equation 7 specify WD as:

$$WindowDiff_d(Hype, Ref) = \frac{\sum_{x=1}^{M-d} |r(x,d) - h(x,d)|}{M-d} \qquad (7)$$

where, r(x,d) is present the Ref segmentation boundary number which contain x and x+d sentences, while h(x,d) present the Hyp segmentation boundary number which contains the x and x+d sentences (M shows the number of sentences and d presents number of word size.)

4.4. Parameter Alteration and Baselines

4.4.1 Word embedding dimension

The word embedding dimension concerning to quantify how word embedding w_i can be instructive for concluding sentences, recognize segment boundaries, and semantic relationships. The detailed word embedding process is represented in Figure 7.

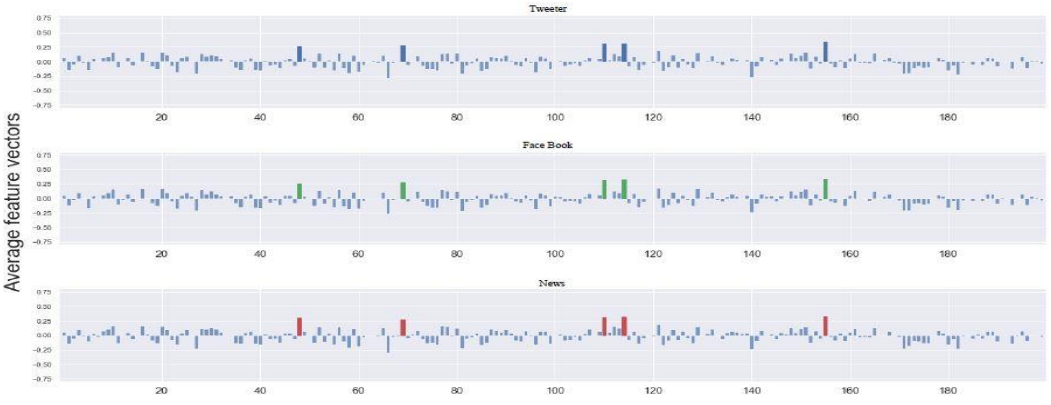


Figure 7. Word Embedding Dimension

In order to measure the algorithm efficiency of given modified dimension of word embedding, and to estimate the trade-off between longer training time and higher dimension embedding, we used Glove embedding pre-trained 50d, 100d and 300d on Wikipedia with 400k vocabularies as well as Principal Component Analysis (PCA) and new metric of 50d embedding. Word2vec feature distribution, performance, and also running time particularity is summarized in Table 6.

Table 6. Dimension lead to performance with a small run time

Word Embedding	w2w-test-score	w2w-train-sxore	Rank-test-score	Param-penalty	R.t (s)
PCA (0)	-0.6795	-0.6712	1	I1	40.3
Glove Embdding	-0.69	-0.68	2	I2	

PCA represents as 0 and Glove Embedding as 1. by comparing the process of training and testing the word2vec feature, results shows that Glove embedding has 1 percent higher result than PCA in training set and 2 percent higher in the test set. The output of the process indicates that Glove Embedding performance is more effective in this analysis because PCA has limitation in data distribution and also it requires a large number of processed units based on the size of document.

Table 7. Word Embedding Measurement on Book Dataset

Word Embedding Model	Performance		
	WD(%)Lowest	Pd(%)Lowest	Runtime(m)
glove-400K-50d	8.5	25.8	3.38
glove-400K-100d	8.9	25.2	3.41
glove-400K-300d	9.2	23.1	3.68
glove-400K-300d - 50d	7.3	21.9	3.40
glove-2.2M-300d	6.1	6.2	3.68
glove-2.2M-300d - 50d	8.1	22.9	3.40

Table 7 presents the effect of word embedding on book dataset performance. Inputs demonstrate word embedding model, and it is per minute in the training set. High dimension embedding is effective to increase the prediction accuracy.

4.4.2 Topic Numbers

To detect abstract or topics through a document, topic modeling is required, which is one of the statistic models in natural language processing. LDA used to categorize document contents into specific topics. In data pre-processing steps we followed tokenization, remove stop words, apply word lemmatized which is to change the third-person value to first-person and change all past scenes to present, stemming and finally make a dictionary of words base on bag-of-words which shows how many time a word occurs in training set. One of the disadvantages of LDA topic modeling is figuring out topic numbers and should set it manually. To do this and to consider the validity, we set latent dimension over 0, 5, 10, 15, 20 and 200 to calculate text segmentation proficiency on text dataset whereas it has an accurate segmentation and shown in Table 8.

Table 8. Sentence topics for all domains

Dataset	Topic Number	Top Words
Twitter	18	long, draw, Gilman, skull, one
	19	Foot, inch, three, six, could
	39	Fact, get, upon, gold, pocket
	63	Quit, become, circus, consider, purpose
Face Book	66	Knew, second, could, latter, however
	0	Far, long, make, one, could
	1	Fear, though, old, would, upon
	2	One, could, Night, see, host
	32	Strange, self, bury, gaze, relief
	104	Gone, dream, complete, son, guess
	29	Chamber, Raymond, sister, protector, worm
New	30	Despair, warm, would, grief, relieve
	61	Father, affect, heart, love, strength
	108	Never, love, together, miser, would
	119	Suggest, power, case, devote, resolve
Total 3 domains	150	Number of unique topic names: 75

To simulate the LDA model in the proposed system, we apply it in 3 different categories which some parts of that contain short text document that will discuss it more in the experiment section. Figure 8 shows the probability output of 150 topics.

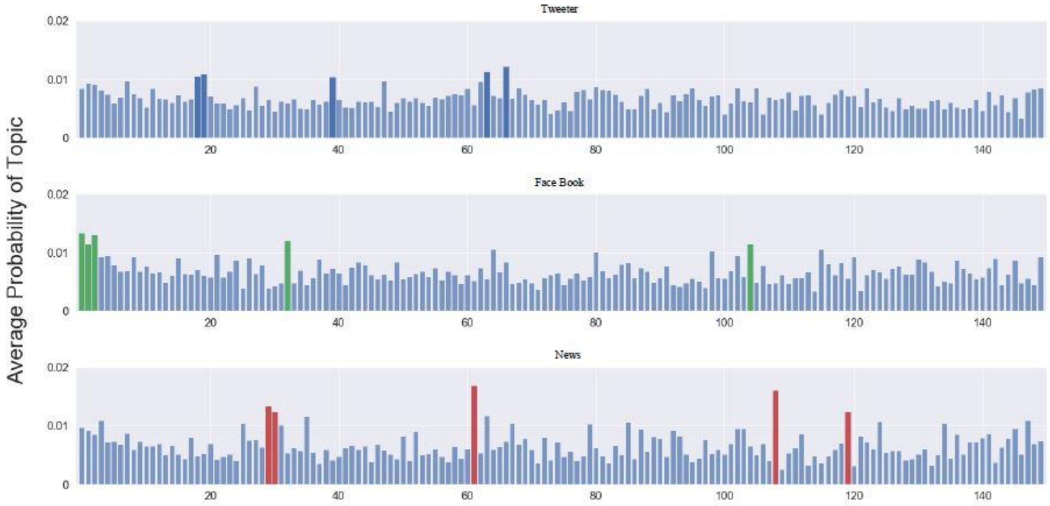


Figure 8. Visualization of value correlation across a number of topics used LDA model

As shown in Figure 8, each row represents the output of one category which is marked "Tweeter", "Face Book", "News". Each category contains 150 topics that in higher probability, five topics have more similar content than others.

To fit topics in text data we applied pyLDAvis which shown in Figure 9. pyLDAvis is to show the probability of words occur in topic. It extracts information from a fitted LDA topic model to inform an interactive web-based visualization.

Figure 9 shows the measurement of terms and detail information about topic terms and also presents the importance of topics based on the data type. To do this in the first step, we use tokenized words which collected as a dictionary and convert to the bag-of-words corpus. This dictionary reports each word appearing time (The probability of how many times a single word appears). This word gives information about the topic content. The circled area shows the topic importance in the whole corpus; the distance between circles presents the topic similarity. Separately in every topic, right side histogram shows the 30 top related terms. The blue color represented as the probability of relevant words in each topic and red color present as estimated terms which appeared in that topic.

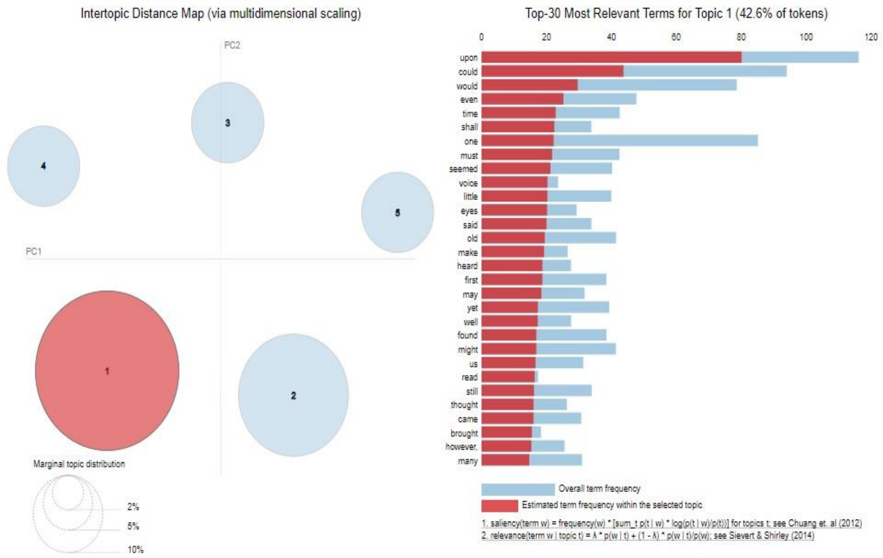


Figure 9. Relevant Terms for Topic 1

Table 9 shows the number of topics which generated out of LDA process and output of topic details based on the probability of words which occur in contents. By comparing the result of Table 9, topic one has the most significant amount of relevant words in document content which means that the highest information or most crucial part is aggregated in topic one.

Table 9. Probability of relevant words in different topics

Number of topics	Related words probability
Topic 1	42.6%
Topic 2	25%
Topic 3	10.8%
Topic 4	10.8%
Topic 5	10.8%

Table 10 presents the effectiveness of LDA model. Generally, processed topics have minimum accuracy on the proposed system. Training time on large documents takes around 3 to 5 minute. Therefore, text segmentation has lower training time if the number of topics in LDA is smaller.

Table 10. Modifying Number of Topics in LDA Model

Number of Topics	Performance		Run Time (m)
	WD (%) Lowest	Pd (%) Lowest	
T = 0	5.1	6.2	3.41
T = 5	6.9	11.1	5.49
T = 10	5.1	6.2	4.96
T = 15	6.9	22.9	5.12
T = 20	6.9	25.2	5.38
T = 50	7.2	22.9	5.94
T = 100	6.9	25.2	6.46
T = 150	7.1	22.9	6.48

4.4.3 Similarity measurement

In statistics, similarity function is processing to find the resemblance out of sentences, and usually, such measures are opposite of distance metrics. To continue the previous section outputs for finding the similarity between 5 topics in each document and to get more standard result we used one of the natural language processing similarity algorithm which is popular as cosine similarity to find the separation between words/sentences/segments and also combination of word2vec and LDA features used to find the effectiveness of this features in similarity progression, which is summarized in Table 11 in details.

Table 11. Verifying Similarity Measurement in Book Dataset

Train-id	Text	Doc Type	Tokenized Sent	LDA-F	w2v F	cos similarity
0	This process, however, afforded me no means of	Twitter	[this, process, however, afforded me no means o]	0.0014	0.099	0.043
1	In his hand was a gold snuff-box, from wh	Facebook	[in his left hand was a gold snuff-box, from w]	0.0016	0.047	0.58
2	It never once occurred to me that the fumbling	News	[it never once occurred to me that the fumblin]	0.0044	0.14	0.52

Train_id represents the number of inputs as 0,1, and 2. Input texts are based on raw data which is divided into categorized sentences. LDA and w2v features show the highest probability of similarity in contents that first input with 0.0044 in LDA feature have the highest likelihood of similarity and same in w2v with 0.14. Finally, cosine similarity represents maximum similarity which Face Book (input number one) has 0.58 percent similar contents comparing with the other two type.

4.4.4 Clustering Methods

Final segments are merging based on candidate blocks. Common way to evaluate the depth score which measures the deepness of a minimum by looking at the highest coherence scores on the left and on the right, and then search for maxims. Segments x and x+1 depth score evaluated as:

$$d_{x,x+1}= 1/2\ (hl_{x,x+1}+(-c_{x,x+1})+ hr_{x,x+1}+(-c_{x,x+1})) \tag{8}$$

similarity score is defined as $c_{x,x+1}$. $hl_{x,x+1}$ and $hr_{x,x+1}$ estimating the maximum similarity. The highest depth score of true segmentation uses as segment boundaries.

4.5 Text Segmentation Evaluation Baselines

To represent the proposed method performance, we used the comparison of multiple text segmentation baselines. Domain-Independent text segmentation baselines are "TextTilling", "C99", "LCseg" and "U00" and domain-dependent baselines are "F04", "M09" and "TopicTilling".

4.5.1 Comparison with Baselines

Table 12 presents the comparison of the proposed topic modeling method with other state-of-art methods on social media contents using P_d metric. The proposed method in domain-independent which is shown as Group A has no training set however, it needs a specification of some hyper-parameters. Domain-dependent which is shown as Group B unlike Group A requires a training set.

Table 12. Comparison of Various Algorithms in Literature based on Social Media Contents

Group	Baselines	P_d (%)
Domain-Independent (A)	TextTiling C99 LCseg U00 Proposed Method	56.36 21.61 9.71 8.86 6.19
Domain-Dependent (B)	F04 M09 TopicTiling	5.31 3.83 1.99
Machine Learning Algorithms (C)	K-Mean Logistic Regression	95% 97%

Generally, proposed system output performance on domain-independent text segmentation can combine with other domain-dependent algorithms by combining features, e.g. trained word embedding from a neural network system. Figure 10 shows the comparison of K-Mean and Logistic Regression machine learning algorithms.

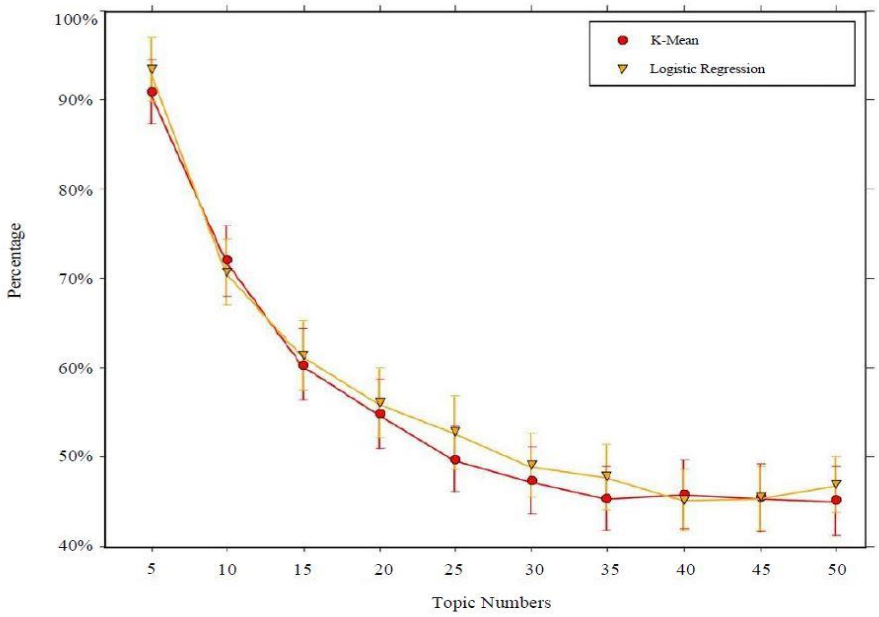


Figure 10. Comparison of K-Mean and Logistic Regression Machine Learning Algorithms

5. Conclusion

In this study, proposed system is a new topic modeling process to analyze short text documents in social media websites based on user information that is called topic models mixture. the proposed system can use a huge data type to get more document relationship by combining texts and also detect more information from same data type and same user. Experiments shows that proposed system can capture topic evaluations and clusters and accomplish confusing baselines. Moreover, this model display that it can discover more relevant topics from contents.

References

[1] Martin Scaiano, Diana Inkpen, Robert Laganiere, and Adele Reinhartz. Automatic text segmentation for movie subtitles. In Canadian Conference on Artificial Intelligence, pages 295–298. Springer, 2010.

[2] Hyo-Jung Oh, Sung Hyon Myaeng, and Myung-Gil Jang. Semantic passage segmentation based on sentence topics for question answering. Information Sciences, 177(18):3696–3717, 2007.

[3] Xiangji Huang, Fuchun Peng, Dale Schuurmans, Nick Cercone, and Stephen E Robertson. Applying machine learning to text segmentation for information retrieval. Information Retrieval, 6(3-4):333–362, 2003.

[4] Irina Pak and Phoey Lee Teh. Text segmentation techniques: a critical review. In Innovative Computing, Optimization and Its Applications, pages 167–181. Springer, 2018. [5] Huosong Xia, Min Tao, and Yi Wang. Sentiment text classification of customers reviews on the web based on svm. In 2010 Sixth International Conference on Natural Computation, volume 7, pages 3633–3637. IEEE, 2010.

[6] Chuanhan Liu, Yongcheng Wang, and Fei Zheng. Automatic text summarization for dialogue style. In 2006 IEEE International Conference on Information Acquisition, pages 274–278. IEEE, 2006.

[7] Deanna J Osman and John L Yearwood. Opinion search in web logs. In Proceedings of the eighteenth conference on Australasian database-Volume 63, pages 133–139. Australian Computer Society, Inc., 2007.

[8] Moayad Yousif Potrus, Umi Kalthum Ngah, and Bestoun S Ahmed. An evolutionary harmony search algorithm with dominant point detection for recognition-based segmentation of online arabic text recognition. Ain Shams Engineering Journal, 5(4):1129– 1139, 2014.

[9] Yun Wu, Yan Zhang, Si-ming Luo, and Xiao-jie Wang. Comprehensive information based semantic orientation identification. In 2007 International Conference on Natural Language Processing and Knowledge Engineering, pages 274–279. IEEE, 2007.

[10] Muthusamy Arumugam. Processing the textual information using open natural language processing (nlp). Available at SSRN 3361108, 2019.

[11] Freddy YY Choi. Advances in domain independent linear text segmentation. arXiv preprint cs/0003083, 2000.

[12] Marti A Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. Computational linguistics, 23(1):33–64, 1997.

[13] Masao Utiyama and Hitoshi Isahara. A statistical model for domain-independent text segmentation. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pages 499–506, 2001.

[14] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. Journal of the American society for information science, 41(6):391–407, 1990.

[15] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):613–620, 1975.

[16] Ignacio Arroyo-Ferna´ndez, Carlos-Francisco M´endez-Cruz, Gerardo Sierra, Juan-Manuel Torres-Moreno, and Grigori Sidorov. Unsupervised sentence representations as word information series: Revisiting tf–idf. Computer Speech & Language, 56:107–129, 2019.

[17] Ting Zhang and Shuzhi Sam Ge. An improved tf-idf algorithm based on class discriminative strength for text categorization on desensitized data. In Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence, pages 39–44. ACM, 2019.

[18] Jinghuan Guo, Yong Mu, Mudi Xiong, Yaqing Liu, and Jingxuan Gu. Activity feature solving based on tf-idf for activity recognition in smart homes. Complexity, 2019, 2019.

[19] Libby Barak, Sammy Floyd, and Adele Goldberg. Modeling the acquisition of words with multiple meanings. In Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 216–225, 2019.

[20] Roberta A Sinoara, Jose Camacho-Collados, Rafael G Rossi, Roberto Navigli, and Solange O Rezende. Knowledge-enhanced document embeddings for text classification. Knowledge-Based Systems, 163:955–971, 2019.

[21] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, March 2003.

[22] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In Proceedings of the third ACM conference on Recommender systems, pages 61–68. ACM, 2009.

[23] Marie Lienou, Henri Maitre, and Mihai Datcu. Semantic annotation of satellite images using latent dirichlet allocation. IEEE Geoscience and Remote Sensing Letters, 7(1):28– 32, 2009.

[24] Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. In Advances in Neural Information Processing Systems, pages 917–925, 2012.

[25] Muzafar Rasool Bhat, Majid A Kundroo, Tanveer A Tarray, and Basant Agarwal. Deep lda: A new way to topic model. Journal of Information and Optimization Sciences, pages 1–12, 2019.

[26] Tobias Hecking and Loet Leydesdorff. Topic modelling of empirical text corpora: Validity, reliability, and reproducibility in comparison to semantic maps. arXiv preprint arXiv:1806.01045, 2018.

[27] Jay M Ponte and W Bruce Croft. Text segmentation by topic. In International Conference on Theory and Practice of Digital Libraries, pages 113–125. Springer, 1997. [28] Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic text structuring and summarization. Information processing & management, 33(2):193–207, 1997.

[29] Nicola Stokes, Joe Carthy, and Alan F Smeaton. Select: a lexical cohesion based news story segmentation system. AI communications, 17(1):3–12, 2004.

[30] Heidi Christensen, BalaKrishna Kolluru, Yoshihiko Gotoh, and Steve Renals. Maximum entropy segmentation of broadcast news. In Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., volume 1, pages I–1029. IEEE, 2005.

[31] Pei-Yun Hsueh, Johanna D Moore, and Steve Renals. Automatic segmentation of multiparty dialogue. In 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006.

[32] Xiang Ji and Hongyuan Zha. Domain-independent text segmentation using anisotropic diffusion and dynamic programming. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 322–329. ACM, 2003.

[33] Sanda Harabagiu and Steven Maiorano. Finding answers in large collections of texts: Paragraph indexing+ abductive inference. In Proceedings of the AAAI Fall Symposium on Question Answering Systems, pages 63–71, 1999.

[34] Charles LA Clarke, Gordon V Cormack, Thomas R Lynam, CM Li, and Greg L McLearn. Web reinforced question answering (multitext experiments for trec 2001). 2001. [35] Hyeon-Jin Kim, Hyo-Jung Oh, Ji-Hyun Wang, Chung-Hee Lee, and Myung-Gil Jang. The 3-step answer processing method for encyclopedia question-answering system: Anyquestion1. 0. In Annual Conference on Human and Language Technology. Human and Language Technology, 2004.

[36] Yan Xiao, Jacky Keung, Kwabena E Bennin, and Qing Mi. Improving bug localization with word embedding and enhanced convolutional neural networks. Information and Software Technology, 105:17–29, 2019.

[37] Liqun Shao. Text Automation: Title Generation, Summarization and Classification. PhD thesis, University of Massachusetts Lowell, 2018.

[38] Stefan Bunk and Ralf Krestel. Welda: Enhancing topic models by incorporating local word context. In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pages 293–302. ACM, 2018.

[39] Te-Ming Chang and Wen-Feng Hsiao. Lda-based personalized document recommendation. In PACIS, page 13, 2013.

[40] Hemant Misra, Fran¸cois Yvon, Joemon M Jose, and Olivier Cappe. Text segmentation via topic modeling: an analytical study. In Proceedings of the 18th ACM conference on Information and knowledge management,

pages 1553–1556. ACM, 2009. [41] Marti A Hearst. Multi-paragraph segmentation of expository text. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 9–16. Association for Computational Linguistics, 1994.

[42] Freddy YY Choi. Advances in domain independent linear text segmentation. arXiv preprint cs/0003083, 2000.

[43] Martin Riedl and Chris Biemann. Topictiling: a text segmentation algorithm based on lda. In Proceedings of ACL 2012 Student Research Workshop, pages 37–42. Association for Computational Linguistics, 2012.

[44] Hyo-Jung Oh, Sung Hyon Myaeng, and Myung-Gil Jang. Semantic passage segmentation based on sentence topics for question answering. Information Sciences, 177(18):3696–3717, 2007.

[45] Hemant Misra, Franc¸ois Yvon, Olivier Capp´e, and Joemon Jose. Text segmentation: A topic modeling perspective. Information Processing & Management, 47(4):528–544, 2011.

[46] Martin Riedl and Chris Biemann. Text segmentation with topic models. Journal for Language Technology and Computational Linguistics, 27(1):47–69, 2012.

[47] Martin Riedl and Chris Biemann. Topictiling: a text segmentation algorithm based on lda. In Proceedings of ACL 2012 Student Research Workshop, pages 37–42. Association for Computational Linguistics, 2012.

[48] Shoaib Jameel and Wai Lam. An unsupervised topic segmentation model incorporating word order. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pages 203–212. ACM, 2013.

[49] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. Btm: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 26(12):2928–2941, 2014.

[50] Jinnie Shin, Qi Guo, and Mark J Gierl. Multiple-choice item distractor development using topic modelling approaches. Frontiers in psychology, 10:825, 2019.

[51] Okumura Manabu and Honda Takeo. Word sense disambiguation and text segmentation based on lexical cohesion. In Proceedings of the 15th conference on Computational linguistics-Volume 2, pages 755–761. Association for Computational Linguistics, 1994.

[52] Marti A Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. Computational linguistics, 23(1):33–64, 1997.

[53] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems, pages 288–296, 2009.

[54] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In Aistats, volume 5, pages 246–252. Citeseer, 2005.

[55] Makoto Sakahara, Shogo Okada, and Katsumi Nitta. Domain-independent unsupervised text segmentation for data management. In 2014 IEEE International Conference on Data Mining Workshop, pages 481–487. IEEE, 2014.

[56] Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. Representation learning for very short texts using weighted word embedding aggregation. Pattern Recognition Letters, 80:150–156, 2016.

[57] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. science, 315(5814):972–976, 2007.

Authors

Zeinab Shahbazi received her B.S. in software engineering from Pooyesh University, IRAN. In March 2017, she moved to Republic of Korea for M.S studies and started working in internet laboratory, Chonbuk National University (CBNU). After completing her master in 2018, she moved to Jeju-do in March 2019 and started working as a Ph.D. research fellow in Machine Learning Laboratory (MLL), Jeju National University. Research interests include artificial intelligence and machine learning, natural language processing, deep learning and data mining.

Yung Cheol Byun received his B.S. from Jeju National University, Korea in 1993, M.S and Ph.D degrees from Yonsei University in 1995 and 2001. He worked as a special lecturer in SAMSUNG Electronics in 2000 and 2001. From 2001 to 2003, he was a senior researcher of Electronics and Telecommunications Research Institute and he promoted to join Jeju National University as an assistant professor in 2003, where he is currently a professor of Department of Computer Engineering. From 2012 to 2014, he had research activities at University of Florida as a visiting professor. His research interests include the areas of pattern recognition & image processing, artificial intelligence & machine learning, security based on pattern recognition, home network and ubiquitous computing, u-Healthcare and RFID & IoT middleware system.