A COMPUTATIONAL STUDY ON THE EFFECTS OF FEATURE ENHANCEMENT IN TOPIC MODELLING

MR. SIRIWAT LIMWATTANA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF ENGINEERING
(COMPUTER ENGINEERING)
FACULTY OF ENGINEERING
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY THONBURI
2021

A Computational Study on The Effects of Feature Enhancement in Topic Modelling

Mr. Siriwat Limwattana B.Eng. (Computer Engineering)

A Thesis Submitted in Partial Fulfillment
of the Requirements for
the Degree of Master of Engineering (Computer Engineering)
Faculty of Engineering King Mongkut's University of Technology Thonburi
2021

Thesis Committee

……..... Chairman of Thesis Committee
(Asst. Prof. Varin Chouvatut, Ph.D.)

……..... Member and Thesis Advisor
(Asst. Prof. Santitham Prom-on, Ph.D.)

……..... Member
(Asst. Prof. Khajonpong Akkarajitsakul, Ph.D.)

……..... Member
(Asst. Prof. Kejkaew Thanasuan, Ph.D.)

……..... Member
(Lect. Jaturon Harnsomburana, Ph.D.)

Thesis Title	A Computational Study on The Effects of Feature Enhancement in Topic Modelling
Thesis Credits	12
Candidate	Mr. Siriwat Limwattana
Thesis Advisors	Asst. Prof. Dr. Santitham Prom-on
Program	Master of Engineering
Field of Study	Computer Engineering
Department	Computer Engineering
Faculty	Engineering
Academic Year	2021

Abstract

In recent studies, many NLP tasks could gain better performance by applying the word embedding as the representation of words. In this research, we propose Deep Word-Topic Latent Dirichlet Allocation (DWT-LDA), a new process for training LDA with word embedding. DWT-LDA augments the sampling process of an original LDA by incorporating word embedding technique to allow the model to capture topics based embedding. A neural network is applied to the Collapsed Gibbs Sampling process as another choice for word topic assignment. A dataset crawled from Pantip.com and Amazon Customer Review. To quantitatively evaluate our model, the topic coherence framework, topic diversity, and topic quality were used to compare between our approach and LDA. The experimental result on both Thai and English dataset indicate that DWT-LDA performs better than LDA on both datasets.

Keywords: Feature Extraction/ Natural Language Processing/ Topic Modeling

หัวข้อวิทยานิพนธ์	การศึกษาเชิงปริมาณในการพัฒนาฟีเจอร์สำหรับการสร้างแบบจำลองหัวข้อ
หน่วยกิต	12
ผู้เขียน	นายสิริวัทก์ ลิ้มวัฒนา
อาจารย์ที่ปรึกษา	ผศ. ดร.สันติธรรม พรหมอ่อน
หลักสูตร	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
คณะ	วิศวกรรมศาสตร์
ปีการศึกษา	2564

บทคัดย่อ

จากงานวิจัยเมื่อไม่นานมานี้ งานวิจัยด้านกระประมวลผลภาษาธรรมชาติถูกพัฒนาประสิทธิภาพขึ้นมาจากอดีตผ่านการนำเทคนิคการแทนค่าของคำด้วยเวกเตอร์ขนาดจำกัด แต่ถึงกระนั้น การวิจัยและพัฒนาด้านแบบจำลองหัวข้อในภาษาไทยยังมีไม่มาก โดยงานวิจัยนี้มุ่งศึกษา และพัฒนาแบบจำลองหัวข้อแบบใหม่โดยประยุกต์ใช้การแทนค่าของคำด้วยเวกเตอร์ขนาดจำกัดเพื่อพัฒนาขีดความสามารถของแบบจำลองหัวข้อ โดยตั้งชื่อเทคนิคที่คิดค้นขึ้นมาใหม่ว่า Deep Word-Topic Latent Dirichlet Allocation(DWT-LDA) ซึ่ง DWT-LDA ถูกพัฒนาต่อยอดมาจาก LDA โดยการเปลี่ยนแปลงกระบวนการสุ่มหัวข้อ โดยเพิ่มการสุ่มหัวข้อจากการแทนค่าของคำด้วยเวกเตอร์ขนาดจำกัดผ่านการเรียนรู้ของเครื่องด้วยโครงข่ายประสาทเทียมเพื่อทำการกำหนดการกระจายตัวของหัวข้อภายใต้คำนั้นๆ ซึ่งเทคนิคดังกล่าวถูกใช้เป็นตัวเลือกที่สองสำหรับการสุ่มหัวข้อในกระบวนการ Collapsed Gibbs Sampling สำหรับการเรียนรู้แบบจำลองหัวข้อ ในส่วนของข้อมูลที่ใช้ในการทดลอง เราได้ทำการเก็บข้อมูลจากกระทู้ต่างๆบนเว็บไซต์ Pantip.com และ ได้ใช้ชุดข้อมูลการวิพากษ์สินค้าจากเว็บไซต์ Amazon.com เพื่อทดสอบประสิทธิภาพการเรียนรู้ของแบบจำลองในทั้งภาษาไทย และภาษาอังกฤษ โดยวัดผลจาก 1) การคำนวณความสอดคล้องของหัวข้อ 2) ความกระจายของหัวข้อ 3) คุณภาพของหัวข้อ ซึ่งผลการวิจัยพบว่าวิธีการที่ถูกพัฒนาขึ้นในงานวิจัยนี้สามารถเพิ่มประสิทธิภาพของแบบจำลอง LDA ได้ทั้งภาษาไทย และภาษาอังกฤษ

คำสำคัญ : การประมวลผลภาษาธรรมชาติ / การสกัดคุณลักษณะ / การสร้างแบบจำลองหัวข้อ

CONTENTS

PAGE

CONTENTS (Cont'd)

LIST OF TABLES

TABLE **PAGE**

LIST OF FIGURES

LIST OF TECHNICAL VOCABULARY AND ABBREVIATIONS

BoW	=	Bag-of-Word
DWT-LDA	=	Deep Word-Topic Latent Dirichlet Allocation
ETM	=	Embedded Topic Model
FSL	=	FastText-based Sentence - Latent Dirichlet Allocation
HTML	=	Hypertext Markup Language
LDA	=	Latent Dirictlet Allocation
LF-LDA	=	Latent Feature - Latent Dirichlet Allocation
LSI	=	Latent Semantic Indexing
NLP	=	Natural Language Processing
NLTK	=	Natural Language Toolkit
NMF	=	Non-negative Matrix Factorization
PLSA	=	Probabilistic Latent Semantic Analysis
PyThaiNLP	=	Thai Natural Language Processing in Python
ReLU	=	Rectified Linear Unit
SVD	=	Singular Value Decomposition
TNG	=	Topic N-grams
WE-LDA	=	Word Embedding - Latent Dirichlet Allocation

CHAPTER 1 INTRODUCTION

1.1 Statement of Problem

With the massive amount of data existing nowadays, relying on human to read and understand all documents in order to comprehend all the minutes details in everyday life has become difficult, if not impossible. Finding a way to systematically extract information from the textual data to help identify important information thus become crucial.

Natural Language Processing (NLP) is a research area which focus on analyzing human languages to understand the underlying information. With the growth of web technology, it makes social network, blog and discussing forum become a casual in our daily life. Most online popular platforms give no boundary to people to express their thought, ideas, and knowledge that understanding these data is a valuable source of information. As of 2018, there was a report stated that 90 percent of data in the world was generated within the last 2 year [1]. To imagine the fast pace of data, the report also showed that 456,000 tweets and 510,000 Facebook comments were posted every minutes. As data grow faster, more data are generated that NLP will become important in data analytics.

Topic modeling is one of the challenging tasks in NLP that focuses on processing unstructured text data to automatically cluster similar document together and identify important word clusters known as topics. One of the most successful topic modeling algorithms is Latent Dirichlet Allocation (LDA) [2]. LDA is a generative model that learns the representation of documents from frequencies of each word used in that document using the Bag-of-Word representation. It treats one word apart from each other, although some of them may closely related or be a synonym. So, the shortcoming of LDA is that it discards the semantic information of words. Therefore, finding a technique which utilizes word-level information to apply with LDA may improve its performance.

After an efficient word embedding technique was proposed by Mokolov et al. in 2013 [3], Skip-gram with negative sampling becomes a powerful word representation that uses less training time and able to embed the semantic and syntactic information of words into embedding space [4]. This technique then has been successfully applied to various NLP

applications[5-7]. Thus, it is logical to test whether the incorporation of word embedding could improve the effectiveness of topic modeling or not.

This study aims to combine the original topic modeling with the word embedding technique to improve the topic modeling algorithm. A hybrid approach for training a Collapsed Gibbs Sampling based LDA with a neural network was tested. To compare the performance of our method and the original LDA, we conducted an experiment on two datasets, a text corpus from Pantip.com as a Thai dataset, and Amazon Customer Reviews as an English dataset [8]. To qualitatively compare the result, topic coherence, topic diversity and topic quality were used for the evaluation.

This thesis is organized as follow: Section 1 introduces the thesis objectives and problem statement. Section 2 presents the related works along with the discussion and literature review on the algorithms involved in this thesis, Section 3 describes our datasets, data preparation process, our modified algorithms including the experimental design, and the metrics used for evaluation, Section 4 presents the experimental results, and Section 5 is the discussion of this research and conclude the experiments of this research.

1.2 Objectives

1. To study on the feasibility of improving Latent Dirichlet Allocation algorithm with word embedding.
2. To study on the feasibility of applying Topic Modeling on Thai language.

1.3 Scopes

1. The experiment will be conducted based on 2 languages which are Thai and English dataset. 1) The Thai dataset will be collected from www.pantip.com which is a Thai-language based online discussion forum. We crawled the first thread of posts across all forums in total of 58,304 posts. 2) The English dataset is sampled from Amazon Customer Review dataset [8] which consist of customer reviews on their purchased products which categories are beauty, clothing shoes and jewelry, grocery and gourmet food, sports and outdoor, and video games in total of 77,749 reviews.
2. Experiment on improving the topic model by adding new features into the modified algorithm.

3. Qualitatively compare the result on three metrics which are topic coherence, topic diversity, and topic quality.
4. Use Latent Dirichlet Allocation as a baseline model for evaluation.

CHAPTER 2 LITERATURE REVIEW AND THEORY

2.1 Related Work

A brief history of topic modeling technique [9, 10] started from Latent Semantic Indexing (LSI) [11], it was proposed by Deerwester et al. in 1990. With LSI, a document assumed to be a mixture of latent topics, which was derived from the frequency of words used in the document by Singular Value Decomposition (SVD). SVD decomposes a document-term matrix into term-topic, document-topic, and singular value matrix. The drawback of LSI is that it lacks statistical interpretation of the value from the generated matrix.

A later study in 1999 by Thomas Hofmann, Probabilistic Latent Semantic Analysis (PLSA)[12] was introduced to mitigate LSI statistical vulnerability by using a statistical technique to learn the lower-dimensional representation of documents. It is a generative model in which a topic is sampled from the topic distribution over its document and word is then sampled from the topics. The graphical model of PLSA is illustrated in Figure 2.1. As the probability of a document is a factor, PLSA cannot handle an unseen document.

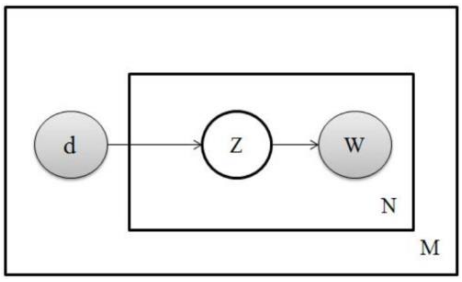

Figure 2.1 Graphical model representation of PLSA model [12]

To mitigate this problem, Latent Dirichlet Allocation (LDA) [2] is proposed as a generalization model of PLSA. Instead of considering a document as a factor, it uses two Dirichlet priors to model the document-topic and word-topic distribution. This approach enables LDA to handle the unseen data which led it to become a standard approach for nowadays topic modeling. Along with the generations of topic modeling, all mentioned technique only uses Bag-of-Word as the input of the model.

Later in 2013, word embedding was proposed by Mikolov et al. [3, 4]. Instead of representing a word as a one-hot encoding vector, it embedded a word into a lower dimension vector. Moreover, it also showed its success in capturing the relationship

between words. Therefore, word embedding starts a new era in Natural Language Processing including topic modeling.

In 2015, Nguyen et al. proposed a latent feature topic model (LF-LDA) [5] which applies word embedding to the topic-word inference process. The process is improved by creating latent features for each topic to be used with word embedding using a Bernoulli distribution that controls the sampling process. The result showed that their proposed algorithm could achieve a higher NMPI score on all datasets.

In 2017, Yao et al. proposed Word Embedding LDA (WE-LDA) [7]. According to the algorithm, LDA is run as an initializer for discovering topics. Then, the topical words are selected to generate a must-link knowledge base on their embeddings. Finally, they include the must-link into Gibbs Sampling, to generate more coherent topics. They run their algorithm to compare with multiple algorithms including LDA and LF-LDA [5]. The result showed that their approach could generate more coherent topic, compared to the other seven models.

In 2019, Dieng et al. proposed a novel Embedded Topic Model (ETM) [5] that constructs the topic embedding from the word embedding space, and generates the document topic mixture using the variational autoencoder model. This model allows the word embedding to be learned during the topic modeling training process, and also allows using a pre-trained embedding. The experiment showed that this model was more robust to stop words when using a pre-trained word embedding compared to training the word embedding during the topic modeling inference process.

In 2020, Zhang et al. proposed the FastText-based Sentence-LDA (FSL) [13] model, which does an augmentation during the sampling process. They pre-compute the distance between words with cosine similarity to find the closest word of each word in the dictionary. The closest word is used as an alternative word for drawing words from topics. The switching between the original and the augmented word is controlled by a Bernoulli distribution. According to the literature, their technique could generate higher topic coherence among the other two baseline models. This approach inspires us the feasibility of not considering a certain word, so we decide to do an experiment with raw word embedding.

According to our survey, there is a few topic modeling in Thai language. Asawaroengchai et al. [14] experiment on comparing the performance between LDA and Topic N-grams

model (TNG) to investigate the feasibility of handling composite word in Thai. The analysis on perplexity showed that TNG performs better than LDA. Another recent research in Thai has been done by Pitichotchokphokhin et al. [15], they compared the performance of Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation on Thairath dataset with the number of topics from 5 to 35 topics. The result showed that NMF could achieve better performance over LDA. In the aspects of Thai topic modeling, we found that there is a research gap on experimenting with topic modeling which is incorporated with word embeddings. Therefore, we conduct this research to examine the feasibility of applying word embedding in Thai topic modeling.

2.2 Theory

2.2.1 Bag-of-Word

Bag-of-Word (BoW) represents documents using a word-document matrix. It requires a finite set of known words called dictionary or vocabulary set. Each position of the matrix stores the number of times that the word is seen in the document, so the sequence of words are discarded. As shown in Table 2.1, the word "Program" appears on Doc 1 and Doc 2 for 5 and 2 times respectively while information of preceding and proceeding word are unknown.

Table 2.1 Document representation using Bag-of-Word [10]

Term	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
Computer	0	3	2	0	1
Printer	1	2	8	0	1
Program	5	3	0	1	0
Unix	0	3	0	0	1
Microprocessor	2	0	1	0	3

2.2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative model proposed by Blei et al. [2]. It is designed on an assumption that each document contains a mixture of topic, and each topic contains a mixture of words which defines the possibility of words to be in a topic. Each document M contains a distribution over k topics as θ with a Dirichlet prior α. Each word w in a document is randomly drawn from each topic $\phi_{1..k}$ with a Dirichlet prior β. For

the graphical model of LDA, see Figure 2.2. The generative process for a document is as follows:

For each word w_{ij} in document D_j

1. Draw topic $z_{ij} \sim Multinomial(\theta_j)$
2. Draw word $w_{ij} \sim Multinomial\left(\phi_{z_{ij}}\right)$

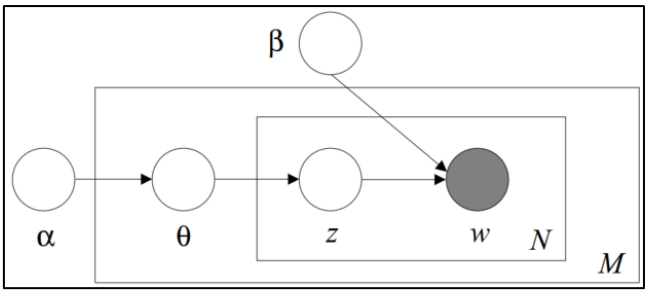

Figure 2.2 Graphical model representation of LDA model [2]

One of the widely used approximation techniques is Collapsed Gibbs Sampling [16]. Instead of inferencing all posterior distributions, it directly samples the topic z to the observed word w to estimate θ and ϕ as shown in Algorithm 2.1. The approximation of ϕ of word w and topic k is shown in Equation 2.1, θ of topic k of j^{th} document is shown in Equation 2.2, and the probability of topic to be assigned is shown in Equation 2.3.

$$\phi_{wk} = \frac{N_{wk} + \beta}{N_k + W\beta} \tag{2.1}$$

$$\theta_{kj} = \frac{N_{kj} + \alpha}{N_j + K\alpha} \tag{2.2}$$

$$P(z = k|w, \alpha, \beta) = \frac{\phi_{wk} * \theta_{kj}}{\sum_{k=1}^{K} \phi_{wk} * \theta_{kj}} \tag{2.3}$$

Algorithm 2.1 Collapsed Gibbs Sampling for Latent Dirichlet Allocation

Require: a corpus of document D
Require: i^{th} word in the document W_{di}
Require: number of word in each document N_d
Require: frequency of each word in each document N_{di}
Require: frequency of topic assigned to each document N_{kj}
Require: frequency of word assigned to each topic N_{wk}
Require: topic assignment of each word z_{dij}

for d in 1... D **do**
  for i in 1... N_d **do**
    $w = W_{di}$
    for j in 1... N_{di} **do**
      $\hat{k} = z_{dij}$
      $N_{w\hat{k}} = N_{w\hat{k}} - 1$
      $N_{d\hat{k}} = N_{d\hat{k}} - 1$
      $\hat{k} \sim$ Multinomial(P($z|z^{-ij}, w, \alpha, \beta$))
      $z_{dij} = \hat{k}$
      $N_{w\hat{k}} = N_{w\hat{k}} + 1$
      $N_{d\hat{k}} = N_{d\hat{k}} + 1$
    end for
  end for
end for

2.2.3 Neural Network

Neural network is inspired by the mammalian brain which can perform multiple tasks. The cells work in a direction that receives inputs from dendrites, processes the inputs in the cell body, and sends the output via synapses as shown in Figure 2.3. In supervised machine learning, a neuron [17] is defined by the connection strength as weights to the inputs, then aggregate the inputs into one value and ship out to another neuron as illustrated in Figure 2.4

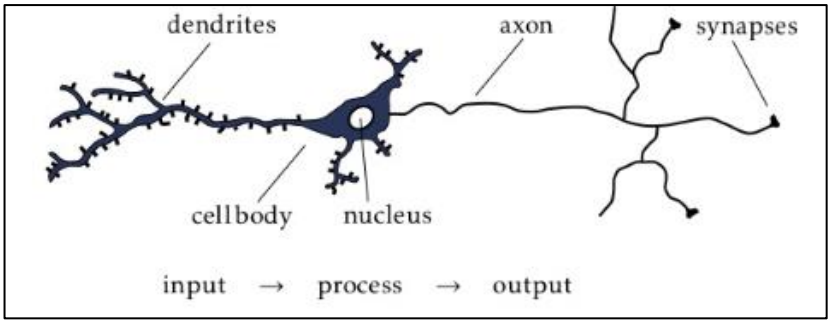

Figure 2.3 The schematic image of a neuron. [18]

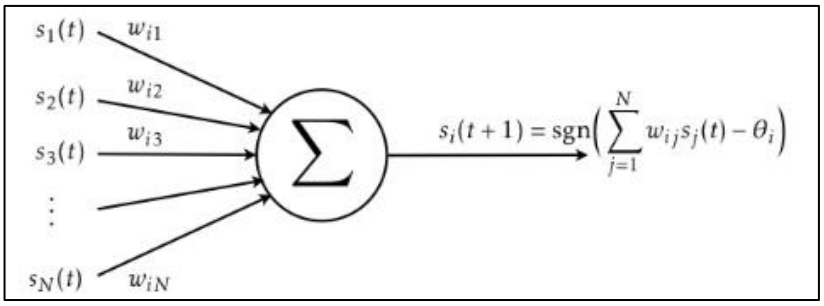

Figure 2.4 Schematic diagram of a McCulloch-Pitts neuron. [18]

Rosenblatt [19] suggests that the neuron should be connected into layers as a feed-forward neural network. It only processes data from the preceding layer and sends the output to the proceeding layer without connection within the same layer. As a single layer, the matrix operation of a single layer is defined on Equation 2.4 where ϕ is the activation function, x is the inputs from the preceding layer, w is the weights to the specific preceding node, and b is the bias. To solve complex problems, the feed-forward network is stacked into multiple layers which are referred as Multi-Layers Perceptron as illustrated in Figure 2.5. To handle the nonlinear problem, an activation function can be put at the output of a neuron which acts as a post-process of data within a neuron. For example, a Rectified Linear Unit or ReLU can be used as an on and off switch for the output that negative value will be set to 0 while there is no effect on the positive value. To perform a classification task, a Softmax activation is used to map the n dimensional output vector into probability vector [20]. Each position of the vector represents the probability of that class.

$$y = \phi(z) = \phi(w^T x + b) \tag{2.4}$$

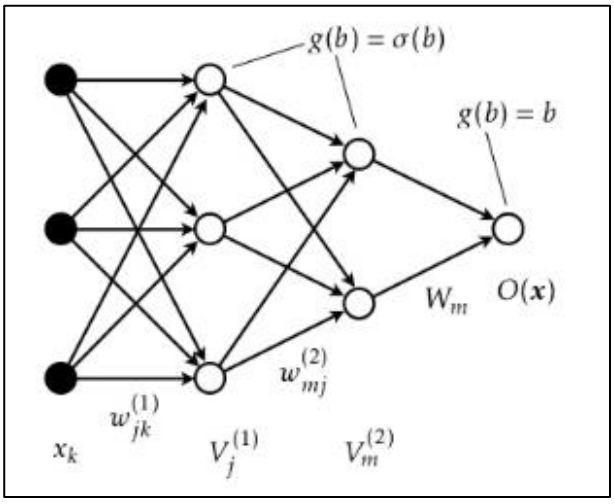

Figure 2.5 The multi-layer perceptron with 2 layers. [18]

The key success of a neural network is the ability to automatically learn the input which adapts with errors and adjust the weight into a strong machine learning model. The backward propagation process is used to adjust the weight of the inputs. Once the forwarding path is finished, the actual label of the data is compared with the output and compute the loss function and perform the backward propagation to reduce the loss. Stochastic gradient descent is one of the basic techniques do the task. It performs making subsets of the training data into batches for training.

2.2.4 Word Embedding

The finding of word embedding by Mikolov et al. in 2003 [3, 4] is a big step in Natural Language processing. The embedding is learned by giving nearby words to predict the central word using feed-forward Neural Network Language Model architecture as shown in Figure 2.6. He also proposes a Skip-gram model which is trained by a shallow network that takes a word as input to predict the context word. Both approaches succeed in representing words in the vector space which captures the syntactic and semantic word relationship from a very large unlabeled text dataset.

By adopting the negative sampling method on Skip-gram, the training process of Skip-gram with negative sampling is to make a shallow network to distinguish between the contexts word and random sampling. It shows an improvement in the speed of model training that performs a logistic regression instead of a Softmax layer. It also improves the quality of the embedding vectors, especially for rare words. Therefore, it has become a standard practice for word representation in current research in NLP.

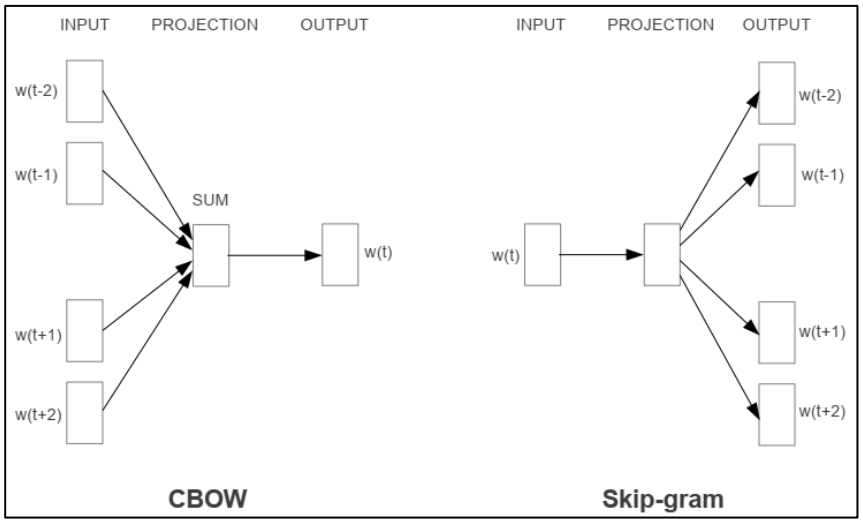

Figure 2.6 The model architecture of word embedding [4]

CHAPTER 3 MATERIAL AND METHODOLOGIES

3.1 Dataset

3.1.1 Pantip dataset

Pantip.com is one of the most popular online discussion boards in Thailand. In the site, people are discussing or sharing a wide range of topics on Pantip (e.g., science, finance, traveling, loves, religions, cooking, sports, and dramas). Text data of the posts were download from the first post written by the owner of each thread to ask or to share something across all forums. A total of 58,304 posts were collected, and the data preparation process is as follows:

1. Eliminate HTML tags.

   Raw data of each document collected from a website is a HTML page. To get a plain text, HTML tags need to be handled to reconstruct the content.

2. Mask the numbers.

   The numbers were masked with a special tag <number> to reduce the bias to a specific number.

3. Normalize the character.

   In Thai language, people frequently use the sub-character to easily type complex characters that look identical and readable by humans. Although those words closely look the same, the sequence of characters is different. Detecting these patterns and handling them with the correct character can reduce the dictionary size and the word is more consistent. For example, replacing double เ (เ) with แ, and -ี and า with -ำ. The normalization tool used in this research is provided by PyThaiNLP [21].

4. Token segmentation.

   Text was segmented into a sequence of tokens. Each token was treated as a word for the analysis. In this research, a pre-trained Attacut [22] is used to tokenize texts into tokens.

5. Remove stop words.

   Stop word is an extremely common word which less contribute to the meaning of a sentence. To maintain a small set of meaningful vocabulary, stop words

was removed from the vocabulary. In this research, stop word removal is based on the list provided in PyThaiNLP [21].

6. Remove common words.

   Words, presented on almost all documents, are expected to be common word. To maintain a small set of meaningful vocabulary, filtering out words which exist in 90 percent of documents could remove common words.

7. Remove rare words.

   Words which hardly exist in document are considered as rare words. To maintain a small set of meaningful vocabulary, filtering out words which exist in fewer 10 documents could remove rare words.

3.1.2 Amazon Customer Review dataset

Amazon.com is a global e-commerce platform in which customers may give a positive or negative review of the product they bought. The dataset was published in [8]. We randomly sample 20 percent of the dataset which a document length is greater than 100 words from various categories including beauty, clothing shoes and jewelry, grocery and gourmet food, sports and outdoor, and video games. The data preparation process is as follows:

1. Token segmentation using

   Text was segmented into a sequence of tokens. Each token was treated as a word for the analysis. In this research, a tool provided in NLTK [23] was used to tokenize texts into tokens.

2. Lowering the character

   In text processing, vocabulary set is a case-sensitive set. To remain a case-insensitive vocabulary, all characters was converted into lower-case.

3. Remove stop words

   Stop word is an extremely common word which less contribute to the meaning of a sentence. To maintain a small set of meaningful vocabulary, stop words was removed from the vocabulary. In this research, stop word removal is based on the list provided in NLTK [23] and a public list on GitHub [24].

4. Remove common words.

Words, presented on almost all documents, are expected to be common word. To maintain a small set of meaningful vocabulary, filtering out words which exist in more than 90 percent of documents could remove common words.

5. Remove rare words.

   Words which hardly exist in document are considered as rare words. To maintain a small set of meaningful vocabulary, filtering out words which exist in fewer than 10 documents could remove rare words.

6. Remove unknown words.

   In this research, a pre-trained embedding was used for English words. Therefore, words which have not been existed in the pre-trained word embeddings was removed from the vocabulary set.

3.2 Exploratory Data Analysis

3.2.1. Pantip Dataset

1. Raw topic distribution

   According to raw data, the topics are not equally distributed. It is decreasing gradually among the ranked topics as illustrated in Figure 3.3. This confirms that our dataset is diverse among topics.

2. Raw document length and vocabulary size

   Before the data was pre-processed, the dataset had the vocabulary size of 402,568 words. An average document length was 392 and a median document length was 213 words. Figure 3.2 shows the distribution of raw document length.

3. Final document length and vocabulary size

   After the data was pre-processed the final vocabulary size of 43,244 words. An average length of the documents was 156 words and a median of 83 words. Figure 3.2 shows the distribution of final document length.

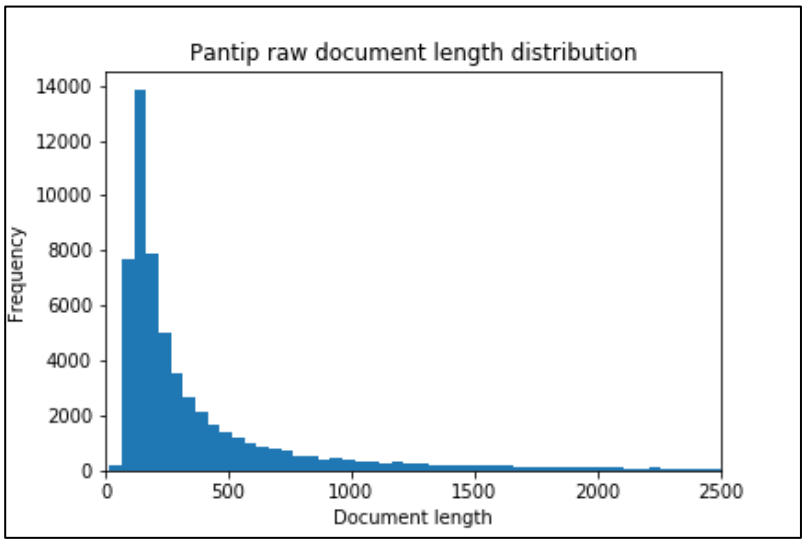

Figure 3.1 The distribution of raw document length on Pantip dataset

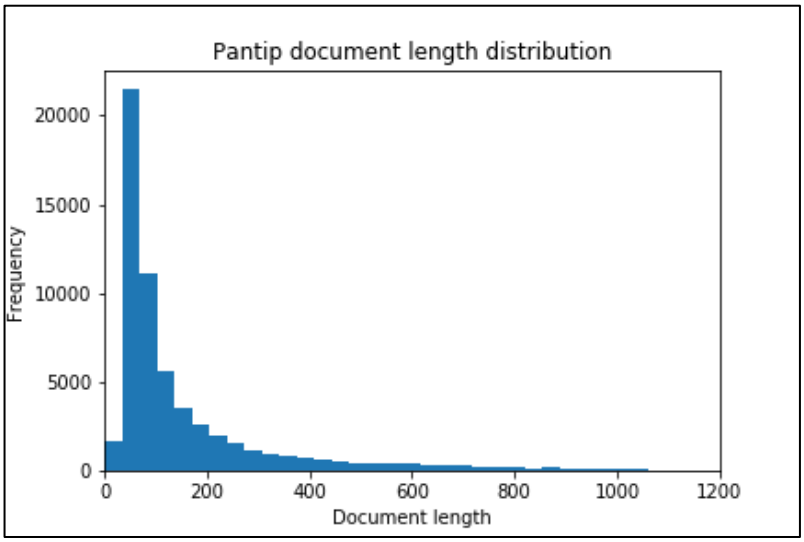

Figure 3.2 The distribution of final document length on Pantip dataset

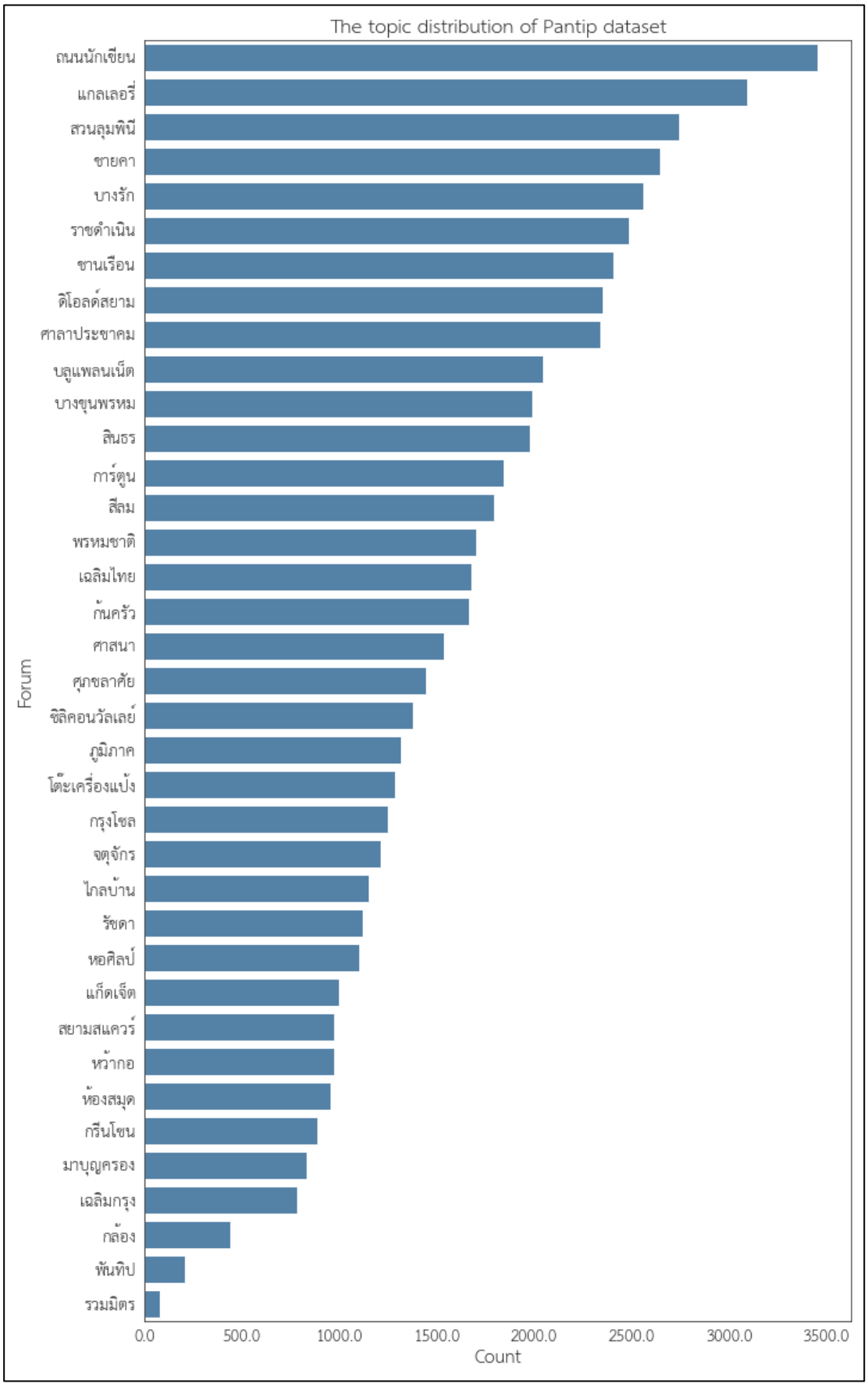

Figure 3.3 The distribution of physical topics on Pantip dataset.

3.2.2. Amazon Customer Review Dataset

1. Raw topic distribution

   Amazon Customer Review is a huge dataset that contains many categories. The reviews are not equally distributed among categories which are gradually decreasing among the ranked categories as illustrated in Figure 3.4. According to the size of the dataset, we only sampled 20 percent of each category from those reviews with a minimum of 100 words to be used as the dataset for this research. Therefore, the final size of our dataset is 77,749 reviews.

2. Raw document length and vocabulary size

   Before the data was pre-processed, the dataset had the vocabulary size of 286,442 words. An average document length was 261 and a median document length was 178 words. Figure 3.5 shows the distribution of raw document length.

3. Final document length and vocabulary size

   After the data was pre-processed the final vocabulary size of 21,377 words. An average length of the documents was 76 words and a median of 52 words. Figure 3.6 shows the distribution of final document length.

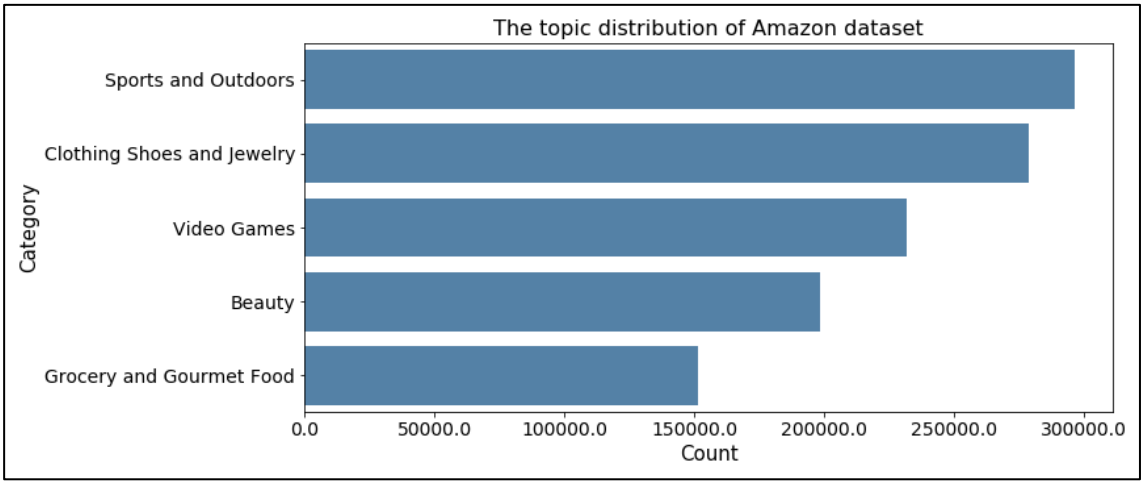

Figure 3.4 The distribution of physical topics on Amazon Customer Review dataset.

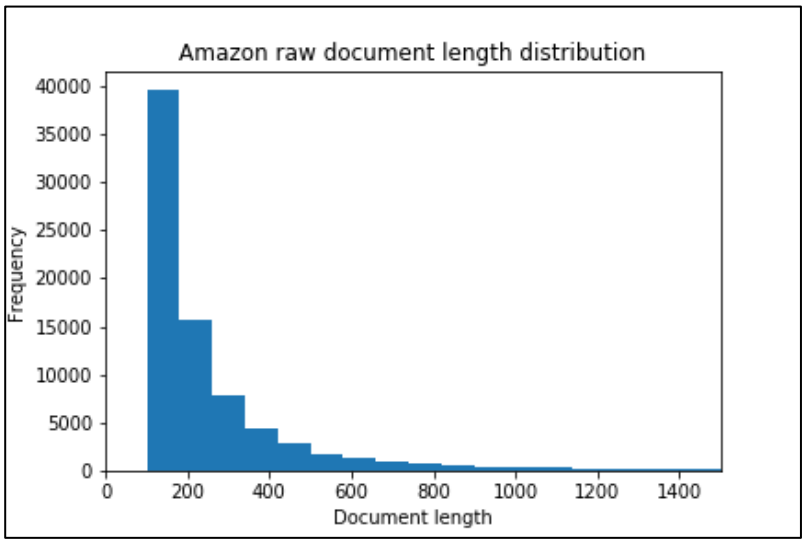

Figure 3.5 The distribution of raw document length on Amazon dataset

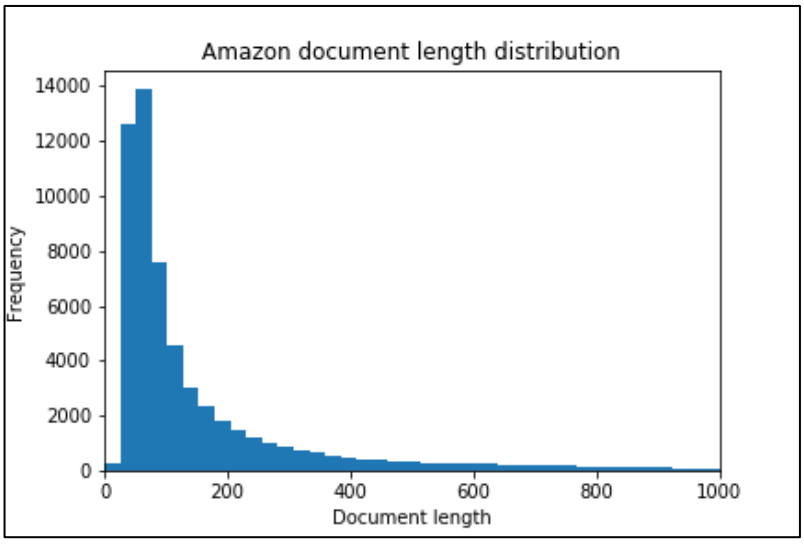

Figure 3.6 The distribution of final document length on Amazon dataset

3.3 Proposed Method

This research proposes a new topic modeling technique as Deep Word-Topic Latent Dirichlet Allocation (DWT-LDA) which has an augmentation on the sampling process using a neural network. Our model is designed to allow the LDA to gain more knowledge on words by applying information from word embedding that the sampling process of topics is involved a neural network which takes word embeddings as the input. The model is divided into 2 steps as follows:

1. Initialize

The first step is a standard LDA using Collapsed Gibbs Sampling method that aims to initialize the topics as shown in Algorithm 2.1. For the initialization, this step needs to be run until it is converged.

2. Topic Enhancement

The second step is a modified version of LDA which is designed to enhance the learned topics by applying the knowledge of words from a pre-trained word embedding. A neural network as shown in Figure 3.7 was trained on the topic assignment of each word which is placed on the sampling process of topic $P(z|w)$. It is designed to learn the association between the topics from LDA and the contextual information which the embedding layer allows the model to learn the topic assignment based on the embedding vector of the word. Since similar words are likely to be close to each other on the embedding space, it increases the chance of being predicted in the same topic that the network input will be closely similar to each other. To maintain a Dirichlet prior, β is added to the predicted probability of the network. A Bernoulli distribution λ is added as switching between topic assignment $P(z|w)$ from using φ and neural network to avoid repeatedly using all of its own predictions as the training data for a later iteration. Therefore, the word-topic agreement φ_{wk} on Equation 2.1 of the original LDA is changed to $\hat{\varphi}_{wk}$ on Equation 3.2. Hence, the conditional probability on Equation 2.3 is changed to Equation 3.3 and the Collapsed Gibbs Sampling for LDA is changed to Algorithm 3.1.

```
Model: "topic_assignment"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
word (Embedding)             (None, None, 300)         7376100
_________________________________________________________________
dense1 (Dense)               (None, None, 1000)        301000
_________________________________________________________________
dropout1 (Dropout)           (None, None, 1000)        0
_________________________________________________________________
dense2 (Dense)               (None, None, 1000)        1001000
_________________________________________________________________
dropout2 (Dropout)           (None, None, 1000)        0
_________________________________________________________________
topic (Dense)                (None, None, 30)          30030
=================================================================
Total params: 8,708,130
Trainable params: 1,332,030
Non-trainable params: 7,376,100
_________________________________________________________________
```

Figure 3.7 The architecture of neural network-based word-topic assignment

$$\lambda_{ij} \sim Bernoulli(\lambda) \tag{3.1}$$

$$\hat{\phi}_{wk} = (1 - \lambda_{ij}) * \frac{N_{wk} + \beta}{N_k + W\beta} + \lambda_{ij} * \frac{NN(w)_k + \beta}{K\beta + \sum NN(w)} \tag{3.2}$$

$$P(z = k | w, \alpha, \beta, \lambda_{ij}) = \frac{\hat{\Phi}_{wk} * \theta_{kj}}{\sum_{k=1}^{K} \hat{\phi}_{wk} * \theta_{kj}} \tag{3.3}$$

Algorithm 3.1 The modified Collapsed Gibbs Sampling for topic enhancement process

Require: a corpus of document D
Require: i^{th} word in the document W_{di}
Require: number of word in each document N_d
Require: frequency of each word in each document N_{di}
Require: frequency of topic assigned to each document N_{kj}
Require: frequency of word assigned to each topic N_{wk}
Require: topic assignment of each word z_{dij}

for d in 1... D **do**
  for i in 1... N_d **do**
    $w = W_{di}$
    for j in 1... N_{di} **do**
      $\hat{k} = z_{dij}$
      $N_{w\hat{k}} = N_{w\hat{k}} - 1$
      $N_{d\hat{k}} = N_{d\hat{k}} - 1$
      $\lambda_{ij} \sim$ Bernouli(λ)
      $\hat{k} \sim$ Multinomial(P($z|z^{-ij}, w, \alpha, \beta, \lambda_{ij}$))
      $z_{dij} = \hat{k}$
      $N_{w\hat{k}} = N_{w\hat{k}} + 1$
      $N_{d\hat{k}} = N_{d\hat{k}} + 1$
    end for
  end for
end for

To run the topic enhancement process, the neural network needs to be initialized with the learned word-topic assignment from the first step for some iterations. In this research, it is empirically set to 20 iterations. After the network is initialized, each iteration of the training procedure is done by running an iteration of modified Collapsed Gibbs Sampling as Equation 3.3. After each iteration, the neural network is trained for several epochs with the topic assignment tracking from the modified algorithm. The whole training procedure is illustrated in Figure 3.8.

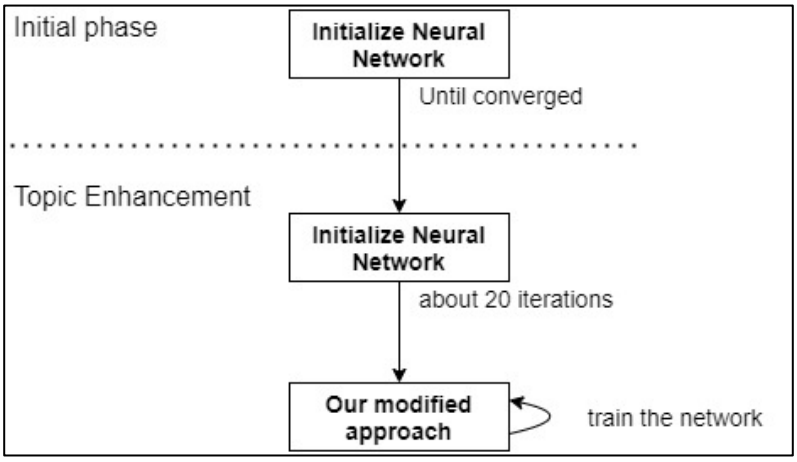

Figure 3.8 The training procedure of our modified algorithm

3.4 Experimental Design

The experiment of our model is designed according to our training phase. We choose the LDA as a baseline model to compare with DWT-LDA. Since our initial phase is an original LDA, so we make a copy of learned model from the initial phase to continue training LDA to compare with the topic enhancement process as illustrated in Figure 3.9. These copies enable us to compare the training process with and without our approach on the same topic initialization.

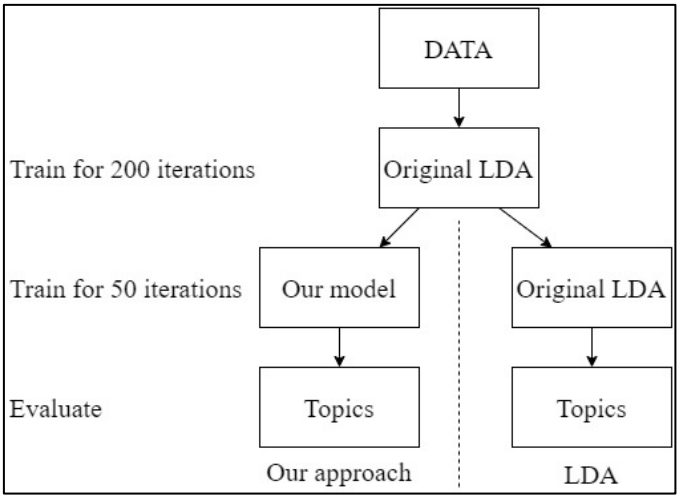

Figure 3.9 The evaluation process of on our approach and the baseline algorithm

To conduct the experiment on both datasets, the training parameter is empirically set to k from the range from k = 5 to k = 180 where β, α and λ were set to $\frac{1}{k}, \frac{1}{k}$ and 0.6 respectively. The boundary of word in Thai language is unclear since we put words close

to each other, the segmentation of word is a challenging task. So, we decide to train the Thai embedding ourselves because we can control the tokenization algorithm to use the same method throughout this research. We train the embedding with the skip-gram model on a very large corpus across multiple domains (e.g., Wikipedia, News, constitutional and etc.). For the English dataset, we use a pre-trained word embedding on Google News Corpus published by Google Open Source.

3.5 Evaluations

To compare the performance between our approach and the baseline model, two metrics were used to qualitatively compare the result of our model. One is the Cv score from the topic coherence framework which metric is widely used in evaluating the quality of topic model. Another is a metric called topic diversity which measures the uniqueness of top keywords among topics. As a combination of topic coherence and topic diversity, we use another metric named topic quality to measure the performance on both metrics.

3.5.1 Topic Coherence

Topic Coherence is a framework for evaluating topic model proposed by Röder et al. [25]. It defines the evaluation process in terms of segmentation, probability estimation, similarity estimation, and aggregation as illustrated in Figure 3.10. According to the literature, the experiment showed that Cv score is the most correlated to human rating. Therefore, we use Cv as a metric in this research.

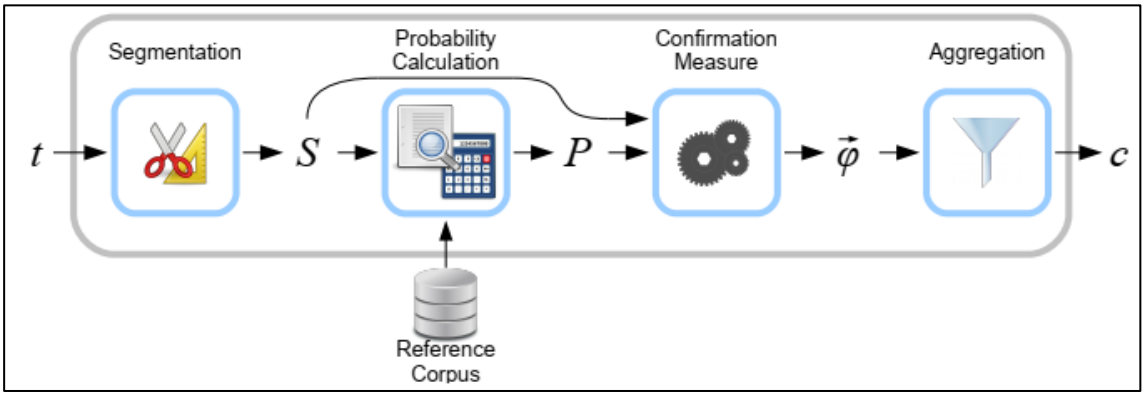

Figure 3.10 An overview of topic coherence framework [25]

Cv is calculated by segmenting the top word in each topic with one and the full set of its topic denoted as W' and W^*, then calculate the agreement of segmented set within the

sliding window of 110 words using Boolean sliding window which discard the distance of word. Next, the Normalized Pointwise Mutual Information (NPMI) and Cosine similarity are calculated as the confirmation measure against each set. Finally, all confirmation measure is aggregated with arithmetic mean as a single value for Cv score.

$$S_{set}^{one} = (W', W^*)|W' = w_i; w_i \in W; W^* = W \tag{3.4}$$

$$PMI = log\ \frac{P(W', W^*)\ +\ \epsilon}{P(W')\ *\ P(W^*)} \tag{3.5}$$

$$NPMI = \frac{PMI}{-log(P(W', W^*)\ +\ \epsilon)} \tag{3.6}$$

$$S_{cos}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\ \vec{u}\ \|_2 \cdot \|\ \vec{w}\ \|_2} \tag{3.7}$$

3.5.2 Topic Diversity

Topic diversity is a novel metric proposed by Dieng et al. [6]. It measures the uniqueness of top keywords among topics that highly overlapped keywords are considered as bad topics. The value close to 1 indicates that the same word did not appear across topics which reduces the ambiguity on interpreting the topic. It is computed by the ratio of unique keywords and the total top keywords in all topics. In this research, the topic diversity is calculated from the top 25 keywords of each topic.

3.5.2 Topic Quality

Topic quality is a measure that combines the topic coherence score and topic diversity by multiplying the topic coherence and topic diversity. It is proposed along with the topic diversity by Dieng et al. [6],

CHAPTER 4 EXPERIMANTAL RESULT

4.1. Qualitative evaluation

After the model was trained with the training process as defined in chapter 3, the models are evaluated with topic coherence, topic diversity, and topic quality. Table 4.1 and 4.2. shows the qualitative evaluation of our model compared to the LDA.

For the topic coherence framework, Figure 4.1 shows that DWT-LDA achieved higher Cv score on both Pantip and Amazon Customer Review datasets than LDA. The topic coherence scores were slightly decreased with a small variation upon the increasing number of topics. The average difference between our approach and LDA was about 0.15 on Pantip, and 0.08 on Amazon dataset. However, the difference between each approach was decreasing with the increasing number of topics. On average, our method was able to yield the topic coherence score of 0.62 on Pantip dataset compared to the original LDA of 0.47, and 0.64 on Amazon dataset compared to 0.55.

For the topic diversity, Figure 4.2 shows that DWT-LDA archived higher topic diversity score on both Pantip and Amazon Customer Review datasets than LDA. The topic diversity scores were gradually decreased with a small variation upon the increasing number of topics. According to the graph, the difference remained stable that the average difference between our approach and LDA was about 0.28 on Pantip, and 0.24 on Amazon dataset. The result showed that our approach clearly outperformed the existing LDA on this metric.

As a single metric, topic diversity was considered to combine topic coherence and topic diversity. The result showed that our approach yielded higher score in all number of topics and on both datasets as illustrated in Figure 4.3.

Table 4.1 The evaluation of DWT-LDA compared to LDA on Pantip dataset

k	Topic Coherence		Topic Diversity		Topic Quality	
	LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA
5	0.37	0.57	0.66	1.00	0.25	0.57
10	0.40	0.60	0.60	0.97	0.24	0.58
15	0.45	0.60	0.56	0.93	0.25	0.56
20	0.46	0.67	0.52	0.91	0.24	0.61
25	0.47	0.63	0.53	0.88	0.25	0.55
30	0.47	0.68	0.48	0.84	0.22	0.57
40	0.47	0.64	0.44	0.84	0.21	0.54
50	0.49	0.63	0.42	0.79	0.21	0.50
60	0.48	0.64	0.40	0.76	0.19	0.49
70	0.47	0.61	0.39	0.75	0.19	0.46
80	0.49	0.63	0.37	0.71	0.18	0.45
90	0.48	0.63	0.36	0.69	0.17	0.44
100	0.49	0.61	0.34	0.68	0.17	0.41
110	0.50	0.63	0.32	0.66	0.16	0.41
120	0.49	0.63	0.30	0.62	0.15	0.39
130	0.49	0.62	0.31	0.61	0.15	0.38
140	0.49	0.60	0.29	0.63	0.14	0.38
150	0.48	0.60	0.28	0.62	0.14	0.37
160	0.48	0.63	0.28	0.60	0.14	0.38
170	0.49	0.60	0.27	0.60	0.13	0.36
180	0.49	0.60	0.27	0.60	0.13	0.36

Table 4.2 The evaluation of DWT-LDA compared to LDA on Amazon Customer Review dataset

k	Topic Coherence		Topic Diversity			
	LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA
5	0.51	0.56	0.73	0.81	0.37	0.46
10	0.50	0.71	0.56	0.93	0.28	0.66
15	0.55	0.67	0.59	0.92	0.32	0.62
20	0.56	0.68	0.60	0.91	0.33	0.62
25	0.56	0.67	0.58	0.91	0.33	0.61
30	0.57	0.66	0.56	0.86	0.32	0.57
40	0.55	0.65	0.54	0.87	0.30	0.56
50	0.57	0.66	0.52	0.82	0.29	0.54
60	0.56	0.66	0.51	0.80	0.28	0.53

k	Topic Coherence		Topic Diversity		Topic Quality	
	LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA
70	0.57	0.64	0.47	0.78	0.27	0.50
80	0.56	0.65	0.47	0.78	0.26	0.51
90	0.57	0.63	0.45	0.75	0.26	0.47
100	0.56	0.62	0.43	0.72	0.24	0.44
110	0.55	0.61	0.42	0.74	0.23	0.46
120	0.55	0.62	0.41	0.72	0.23	0.44
130	0.56	0.62	0.40	0.69	0.22	0.43
140	0.54	0.61	0.38	0.71	0.21	0.44
150	0.53	0.60	0.38	0.73	0.20	0.44
160	0.54	0.61	0.37	0.70	0.20	0.43
170	0.54	0.61	0.36	0.70	0.20	0.43
180	0.54	0.60	0.35	0.68	0.19	0.41

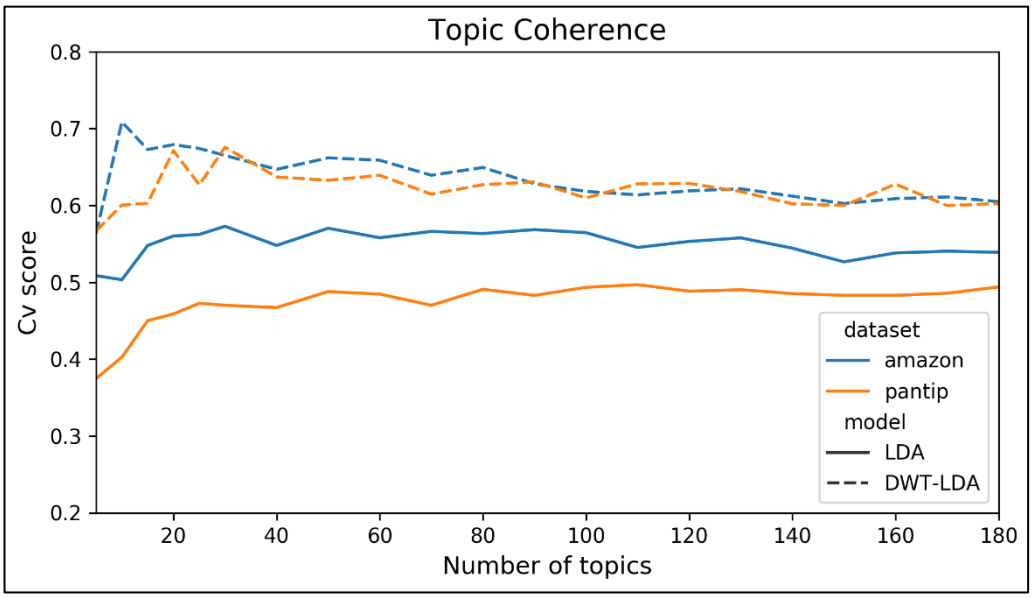

Figure 4.1 Topic coherence between DWT-LDA and LDA

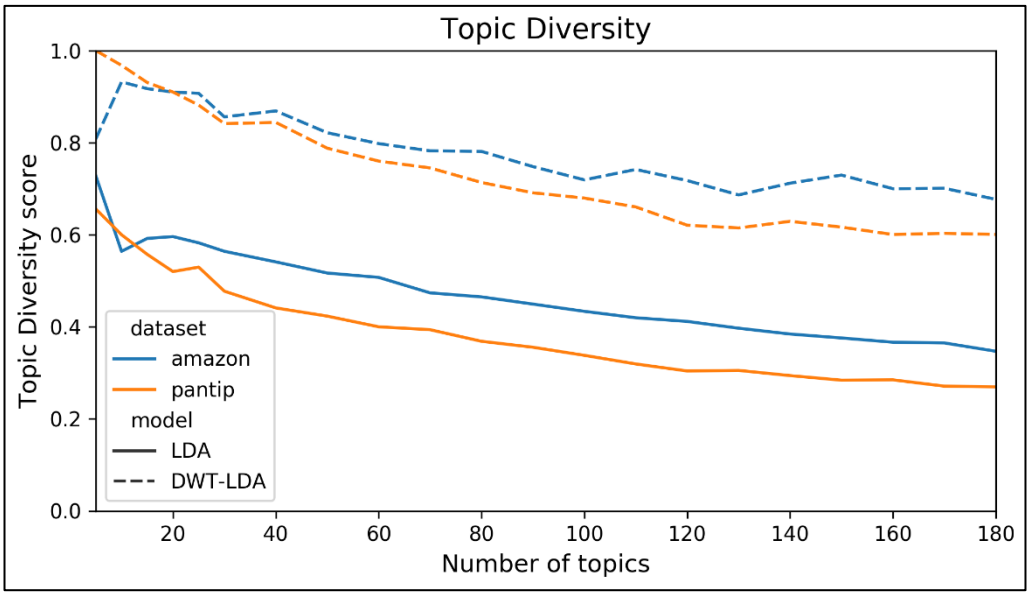

Figure 4.2 Topic diversity between DWT-LDA and LDA

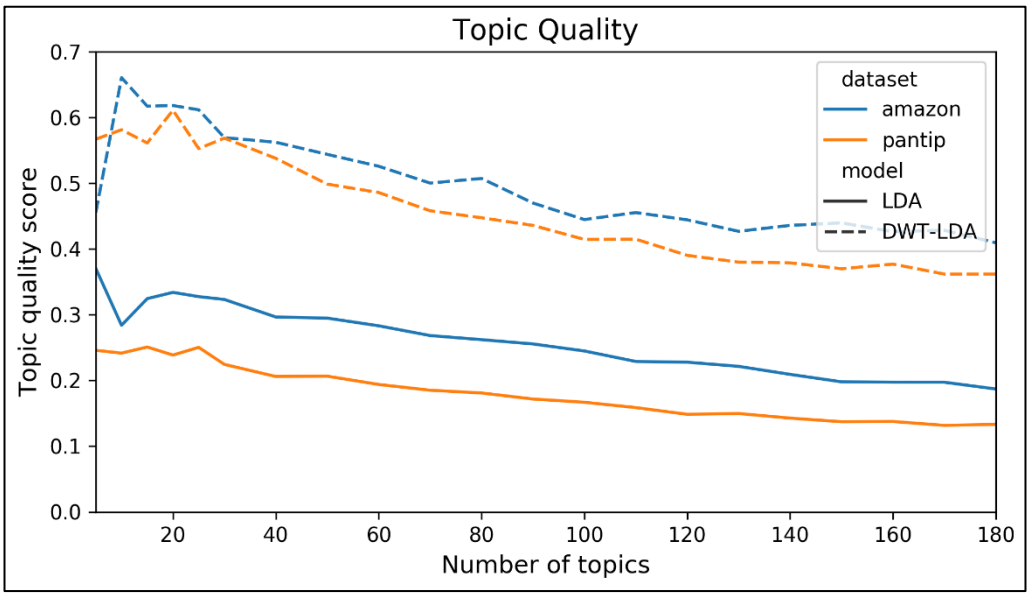

Figure 4.3 Topic quality between DWT-LDA and LDA

4.2. Discovered Topics

Along with the metrics, the learned topics was visually displayed to investigate the learned concept of topics. According to the topic diversity score, we expected our method to give more specific keywords to the topics. Some topics from LDA were more ambiguous while DWT-LDA gave more concrete topics where the keywords were less overlapped. The result from Pantip dataset showed that the word "ทำ" appeared on six topics while it was omitted on DWT-LDA as shown in Table 4.3. Moreover, the result from Amazon Customer Reviews as shown in Table 4.4 also confirmed the less redundant

keywords. For instance, the word “good” showed on almost all LDA topics but it did not show on the DWT-LDA. This behavior made the topics easier to interpret from the top keywords.

Table 4.3 The learned topic from Pantip dataset

Travel		**Politics**		**Skin Care**		**Restaurant**	
LDA	**DWT-LDA**	**LDA**	**DWT-LDA**	**LDA**	**DWT-LDA**	**LDA**	**DWT-LDA**
เดิน	เที่ยว	คน	พรรค	ผิว	ผิว	ร้าน	ร้าน
คน	เดินทาง	ทำ	สส	สี	หน้า	อาหาร	อร่อย
รถ	รถไฟ	พรรค	รัฐบาล	ตัว	สิว	กิน	ทาน
เที่ยว	สถานี	ไทย	เลือกตั้ง	หน้า	สี	น้ำ	ใส่
เดินทาง	เมือง	ปี	ประชาชน	ดี	ทา	ทำ	อาหาร
พัก	บิน	ประเทศ	นายก	ดู	ครีม	ทาน	เนื้อ
เวลา	ตั๋ว	เมือง	รัฐมนตรี	ทำ	เนื้อ	ใส่	รสชาติ
นั่ง	ทริป	เรื่อง	เมือง	น้ำ	สาร	อร่อย	หมู
ถ่าย	จอง	รัฐบาล	รัฐธรรมนูญ	คน	ใส	ดี	หวาน
รูป	รถ	ข่าว	คะแนน	ทา	กลิ่น	ข้าว	น้ำ
ดี	พัก	สส	กกต	สิว	ชุ่มชื้น	หวาน	ไข่
เมือง	สนามบิน	ตัว	ไทย	ลอง	บำรุง	เนื้อ	เมนู
ดู	แรม	ประชาชน	กฎหมาย	ครีม	แห้ง	สั่ง	รส
Buddhism		**Treatment**		**Finance**		**Relationship**	
LDA	**DWT-LDA**	**LDA**	**DWT-LDA**	**LDA**	**DWT-LDA**	**LDA**	**DWT-LDA**
ท่าน	จิต	หมอ	หมอ	เงิน	เงิน	ผม	ผม
ทำ	ธรรม	ยา	ยา	เดือน	บัตร	คน	แฟน
คน	ท่าน	อาการ	อาการ	ค่า	เดือน	ทำ	คุย
โลก	พระ	ทำ	โรค	ทำ	จ่าย	ตอน	คับ
รู้	ศาสนา	กิน	เลือด	บาท	บาท	รู้	คบ
ตัว	พระพุทธเจ้า	ตอน	ตรวจ	ซื้อ	แจ้ง	ตัว	เลิก
ชีวิต	กรรม	ตัว	รักษา	จ่าย	ธนาคาร	เรื่อง	ทัก
พระ	วัด	เดือน	ปวด	ขาย	โทร	ดี	ทะเลาะ
ใด	ทุกข์	โรค	พยาบาล	แจ้ง	หุ้น	ถาม	ก้

Buddhism		Treatment		Finance		Relationship	
LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA
สร้าง	บุญ	รักษา	ป่วย	ปี	ประกัน	เหมือน	จีบ
จิต	นิพพาน	ตรวจ	ผ่าตัด	โทร	กู้	เพื่อน	ผญ
ปี	สมาธิ	เลือด	ไข้	บัตร	บัญชี	ดู	เสียใจ
วัด	องค์	นอน	ฟัน	ตอน	ติดต่อ	เวลา	เทอ

Table 4.4 The learned topic from Amazon Customer Review dataset

Orders		Skin Care		Game Console	
LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA
amazon	amazon	skin	skin	games	controller
price	received	product	face	game	wii
buy	seller	face	product	controller	xbox
product	item	oil	cream	xbox	ps3
reviews	shipping	products	oil	play	games
time	service	dry	lotion	ps3	console
bought	customer	cream	products	wii	nintendo
money	ordered	feel	moisturizer	console	psp
good	arrived	smell	dry	great	sony
review	shipped	good	acne	version	gaming
quality	return	lotion	wash	playing	controllers
purchase	company	day	sensitive	screen	remote
shipping	replacement	scent	scent	psp	vita
buying	purchase	body	cleanser	nintendo	ps2
box	order	time	body	gaming	buttons
Bike		**Clothing**		**Makeup**	
LDA	DWT-LDA	LDA	DWT-LDA	LDA	DWT-LDA
bike	bike	size	shoes	color	brush
rack	rack	wear	wear	watch	color
easy	tire	fit	size	brush	polish
good	seat	shoes	fit	great	nail
work	road	comfortable	shoe	colors	nails
ride	bikes	pair	pair	love	mascara
seat	lock	great	comfortable	good	lashes

Bike		Clothing		Makeup	
LDA	**DWT-LDA**	**LDA**	**DWT-LDA**	**LDA**	**DWT-LDA**
wheel	ride	feet	feet	nail	colors
road	chain	shoe	socks	time	brushes
lock	pump	good	bra	polish	coat
great	riding	ordered	foot	black	makeup
pump	tires	bought	boots	nails	apply
riding	rear	love	wearing	product	eye
tire	pedals	wearing	ordered	nice	foundation
bit	miles	bit	toe	dark	lip

CHAPTER 5 CONCLUSION

5.1 Discussion

According to the evaluation, DWT-LDA could achieve higher score on all metrics of both datasets. As a result, Table 4.3 and 4.4 showed our model success in generating strong keyword to the topics that general keywords were not ranked on the top keywords. For example, the word "คน" and "ทำ" is less relevant to politics while they were listed on the first two strongest keywords of politics compared to "พรรค" and "สส" on DWT-LDA. The result also showed that general word on Amazon Customer Review dataset such as "good" and "great" were not listed on the top keywords.

Although our method was able to improve the original LDA, it highly depends on the LDA on the initial stage. The topic enhancement process learned the topic assignment on the first stage. Therefore, there is a possible limitation that this approach may fail to enhance the topics when the original LDA does not perform well. However, this possibility was not shown in our research.

5.2 Conclusion

In this research, we proposed a new algorithm that upgrade the original LDA by adding an augmentation to the topic assignment during the inference process. The word embeddings allowed our model to learn the word information from a larger corpus. Hence, the word-topic assignment is done by incorporating with word embedding to assign the topic. In this manner, the model learned the topics from the latent information of words instead of considering the close or related word apart as a completely different vector. An experiment was done to compare the performance of DWT-LDA and the original method on Thai and English datasets which were collected from Pantip.com and Amazon.com. The experimental result on topic coherence of Pantip dataset showed that our model could achieve approximately 0.2 on Cv score, higher than LDA. However, the difference between scores was reduced to 0.1 on higher number of topics on Pantip dataset while it slowly decreased from 0.12 to 0.07 on Amazon dataset. The topic diversity which considered the ratio of uniqueness among top keywords showed that DWT-LDA could generate approximately 0.35 and 0.31, higher than LDA on Pantip and Amazon dataset

respectively. Finally, the combined score showed that our approach could achieve higher topic quality than LDA on Pantip and Amazon datasets for 0.28 and 0.24 respectively. Moreover, the generated topics also showed that our model could generate more specific keywords to topics. So, we can conclude that word embedding not only improved the topic modeling in English language but was also able to improve the Thai topic modeling. For future work of both Thai and English topic modeling, some experiments on topic modeling with other embedding methods should be investigated to compare the performance of embedding on topic modeling task.

REFERENCES

1. Marr, B., 2018, **How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read**, [Online], Available: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read [2020, Septemter 12].

2. Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003, "Latent Dirichlet Allocation", **The Journal of Machine Learning Research**, Vol. 3, pp. 993–1022, doi: 10.1162/jmlr.2003.3.4-5.993.

3. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., 2013, "Distributed Representations of Words and Phrases and Their Compositionality", **Proceedings of the 26th International Conference on Neural Information Processing Systems**. Curran Associates Inc., 5-10 December 2013, Lake Tahoe, Nevada, USA, Vol. 2, pp. 3111-3119.

4. Mikolov, T., Chen, K., Corrado, G.S., and Dean, J.,2013. "Efficient Estimation of Word Representations in Vector Space". **The International Conference on Learning Representations (ICLR)**, 2-4 May 2013,Scottsdale, Arizona, USA.

5. Nguyen, D.Q., Billingsley, R., Du, L., and Johnson, M., 2015, "Improving Topic Models with Latent Feature Word Representations", **Transactions of the Association for Computational Linguistics**, Vol. 3, pp. 299-313, doi: 10.1162/tacl_a_00140.

6. Dieng, A.B., Ruiz, F.J.R., and Blei, D.M., 2020, "Topic Modeling in Embedding Spaces", **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 439-453, doi: 10.1162/tacl_a_00325.

7. Yao, L., Zhang, Y.S., Chen, Q., Qian, H., Wei, B., and Hu, Z., 2017, "Mining Coherent Topics in Documents Using Word Embeddings and Large-Scale Text Data", **Engineering Applications of Artificial Intelligence**, Vol. 64, pp. 432-439, doi: 10.1016/j.engappai.2017.06.024.

8. He, R. and McAuley, J., 2016, "Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering", **Proceedings of the 25th International Conference on World Wide Web**, 11-15 April 2016, Montréal, Québec, Canada, doi: 10.1145/2872427.2883037.

9. George, L.E. and Birla, L., 2018, "A Study of Topic Modeling Methods", **2018 Second International Conference on Intelligent Computing and Control**

Systems (ICICCS), 14-15 June 2018, Madurai, India, pp. 109-113, doi: . 10.1109/ICCONS.2018.8663152.

10. Barde, B.V. and Bainwad, A.M., 2017, "An Overview of Topic Modeling Methods and Tools", **2017 International Conference on Intelligent Computing and Control Systems (ICICCS)**, 15-16 June 2017, Madurai, India, pp. 745-750, doi: 10.1109/ICCONS.2017.8250563.

11. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., and Harshman, R.A., 1990, "Indexing by Latent Semantic Analysis", **Journal of the American Society for Information Science.**, Vol. 41, pp. 391-407.

12. Hofmann, T., 1999, "Probabilistic Latent Semantic Indexing" **Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. Association for Computing Machinery, 15-19 August 1999, Berkeley, California, USA, pp. 50–57.

13. Zhang, F., Gao, W., Fang, Y., and Zhang, B.,2020. "Enhancing Short Text Topic Modeling with Fasttext Embeddings". **2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)**, 12-14 June 2020, Fuzhou, China, pp. 255-259, doi: 10.1109/ICBAIE49996.2020.00060.

14. Asawaroengchai, C., Chaisangmongkon, W., and Laowattana, D.,2018. "Probabilistic Learning Models for Topic Extraction in Thai Language". **The 5th International Conference on Business and Industrial Research (ICBIR)**, 17-18 May 2018, Bangkok, Thailand, pp. 35-40, doi: 10.1109/ICBIR.2018.8391162.

15. Pitichotchokphokhin, P., Chuangkrud, P., Kalakan, K., Suntisrivaraporn, B., Leelanupab, T., and Kanungsukkasem, N.,2020. "Discover Underlying Topics in Thai News Articles: A Comparative Study of Probabilistic and Matrix Factorization Approaches". **2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)**, 24-27 June 2020, Phuket, Thailand, pp. 759-762, doi: 10.1109/ECTI-CON49241.2020.9158065.

16. Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M., 2008, "Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation", **Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining**, Association for Computing Machinery, 24-27 August 2008, Las Vegas, Nevada, USA, pp. 569–577, doi: 10.1145/1401890.1401960.

17. McCulloch, W.S. and Pitts, W., 1943, "A Logical Calculus of the Ideas Immanent in Nervous Activity", **The Bulletin of Mathematical Biophysics**, Vol. 5, No. 4, pp. 115-133, doi: 10.1007/bf02478259.

18. Mehlig, B., 2021, **Machine Learning with Neural Networks**, Cambridge University Press, pp. 6-12, 71-107, doi: 10.1017/9781108860604.

19. Rosenblatt, F., 1958, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain", **Psychological Review**, Vol. 65(6), pp. 386-408, doi: 10.1037/h0042519.

20. Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., and Dehmer, M., 2020, "An Introductory Review of Deep Learning for Prediction Models with Big Data", **Frontiers in Artificial Intelligence**, Vol. 3, No. 4, doi: 10.3389/frai.2020.00004.

21. Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., and Chormai, P., 2019, PyThaiNLP: Thai Natural Language Processing in Python, Zenodo, doi: 10.5281/zenodo.3519354.

22. Chormai, P., Prasertsom, P., and Rutherford, A.T., 2019, "Attacut: A Fast and Accurate Neural Thai Word Segmenter", **ArXiv**, Vol. abs/1911.07056

23. Bird, S., Klein, E., and Loper, E., 2009, **Natural Language Processing with Python : Analyzing Text with the Natural Language Toolkit**, 1. ed., O'Reilly, Beijing, China, Pages.

24. jimmyjames177414, Toro, E., and Sean, B., 2020, **Nltk's List of English Stopwords**, [Online], Available: https://gist.github.com/sebleier/554280

25. Röder, M., Both, A., and Hinneburg, A., 2015, "Exploring the Space of Topic Coherence Measures", **Proceedings of the Eighth ACM International Conference on Web Search and Data Mining**, Association for Computing Machinery, 2-6 February 2015, Shanghai, China, pp. 399-408, doi: 10.1145/2684822.2685324.

CURRICULUM VITAE

NAME	Mr. Siriwat Limwattana
DATE OF BIRTH	2 May 1996

EDUCATONAL RECORD

HIGH SCOOL	High School Graduation Mukdahan School, 2013
BACHELOR'S DEGREE	Bachelor of Engineering (Computer Engineering) King Mongkut's University of Technology Thonburi, 2017
MASTER'S DEGREE	Master of Engineering (Computer Engineering) King Mongkut's University of Technology Thonburi, 2021

SCHOLARSHIP/ RESEARCH GRANT	Research Grant for Graduate Student Big Data Experience Center Scholarship, 2018