Automatic Topic Clustering Using Latent Dirichlet Allocation with Skip-gram Model on Final Project Abstracts

Hendra Bunyamin
Informatics Engineering
Maranatha Christian University
Bandung, Indonesia
hendra.bunyamin@it.maranatha.edu

Lisan Sulistiani
Informatics Engineering
Maranatha Christian University
Bandung, Indonesia
lisans1601@gmail.com

***Abstract*—Topic model has been an elegant method to discover hidden structures in knowledge collections, such as news archives, blogs, web pages, scientific articles, books, images, voices, videos, and social media. The basic model of topic model is Latent Dirichlet Allocation (LDA) and this paper utilizes LDA to automatically cluster topics from final project abstract collection. We compare two methods, that are LDA as a unigram model and LDA with Skip-gram model. Our results are evaluated by an expert on readily available categories. Overall, words from each topic are indeed keywords describing each topic; moreover, the combination of LDA and skip-gram model are capable to capture key phrases from each topic.**

***Index Terms*—topic model, latent dirichlet allocation, skip-gram model, final project abstracts**

I. Introduction

An increasing number of final project reports in Maranatha Christian University (MCU) library calls for an organization; specifically, how final project reports are organized will certainly help students find topics for their final projects. The ways project reports are categorized can be based on their topic structures. Blei [1] suggests searching topics through a collection of reports should start from finding topics instead of inputing mundane keywords. After choosing the topic, user may proceed to examine reports with the same topic.

Hidden topic structures from a collection of reports can be discovered by topic models. Conceptually, topic models are probabilistic models that learn semantic structures from a collection of documents based on hierarchical Bayesian analysis [2]–[7]. Topic models have been largely applied into various types of documents, e.g. emails [8], scientific abstracts [4], [6], and news archives [9].

This paper aims to investigate how topic models can be utilized to discover hidden topic structures in final project abstracts, specifically at Maranatha Christian University. As depicted in Fig. 1, average number of visitors from Faculty of Psychology is increasing monotonically. Therefore, we opt to utilize abstracts from Faculty of Psychology as our dataset; additionally, the choice of dataset complements our long-term goal that is assisting pyschology students in finding topics for their final projects.

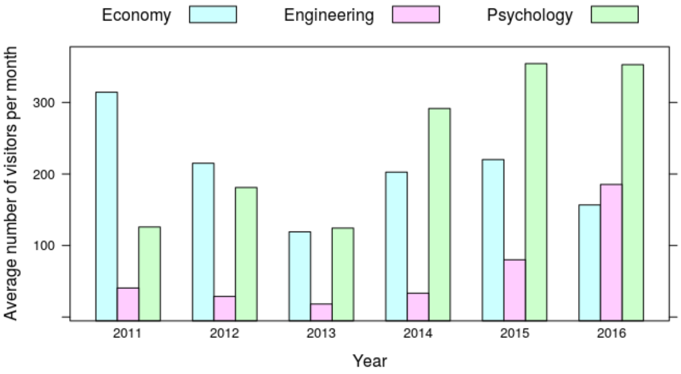

Fig. 1. Average number of visitors per month per year.

II. Related Work

This section explores fundamental concepts of automatic clustering, specifically Latent Dirichlet Allocation and Skip-gram model.

A. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an expansion model from probabilistic latent semantic analysis (PLSA) [3]. Particularly, LDA enhances PLSA model by defining a complete generative process [4]. LDA is a mixture model that employs convex combinations from distributed component sets to model observations. A convex combination is a linear combination of components where all coefficients are non-negative and sum to 1. In LDA, one word w is generated by a convex combination of topics z. Moreover, mixture models specify that the probability of one word instantiating a term t is

$$P(w=t)=\sum_k P(w=t|z=k)P(z=k), \qquad (1)$$

978–1–5386–0787–9/17/$31.00©2017 IEEE

with $\sum_k P(z=k)=1$ and each mixture component $P(w=t|z=k)$ is a multinomial distribution with each term corresponding to a latent topic $z=k$ from a text corpus. Mixture proportions consist of probability of topics $P(z=k)$. Focusing on Equation 1, objectives of LDA inference are

1) finding term distribution $P(t|z=k)=\vec{\varphi}_k$ for each topic k and
2) finding topic distribution $P(z|d=m)=\vec{\vartheta}_m$ for each document m.

Sets of parameters $\underline{\Phi}=\{\vec{\varphi}_k\}_{k=1}^{K}$ and $\underline{\Theta}=\{\vec{\vartheta}_m\}_{m=1}^{M}$ are latent-semantic representation of words and documents. Bayesian network view of LDA is shown in Fig. 2. Table I elaborates all variables in LDA model.

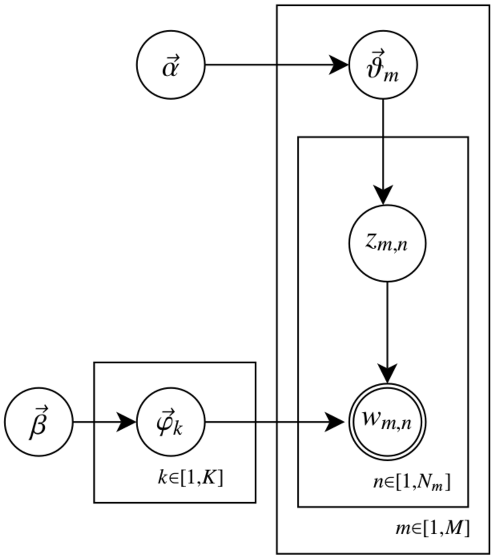

Fig. 2. Bayesian Network from Latent Dirichlet Allocation [10].

TABLE I
VARIABLES IN LDA MODEL [10].

M	number of documents to be generated (scalar).
K	number of topic or mixture components (scalar).
V	number of terms t in vocabulary (scalar).
$\vec{\alpha}$	hyperparameter on mixing proportions (K-vector or scalar if symmetric).
$\vec{\beta}$	hyperparameter on mixture components (V-vector or scalar if symmetric).
$\vec{\vartheta}_m$	parameter notation for $P(z|d=m)$, topic mixture proportion for document m. One proportion for each document, $\underline{\theta}=\{\vec{\vartheta}_m\}_{m=1}^{M}$ ($M\times K$ matrix).
$\vec{\varphi}_k$	notasi parameter untuk $P(t|z=k)$, mixture component untuk topik k. One component for each topic, $\underline{\Phi}=\{\vec{\varphi}_k\}_{k=1}^{K}$ ($K\times V$ matrix).
N_m	Length of document modelled by Poisson distribution [4] with constant parameter ξ.
$z_{m,n}$	mixture indicator that chooses a topic for n-th word from m-th document.
$w_{m,n}$	term indicator for n-th word from m-th document.

B. Skip-gram Model

Techniques to learn high-quality word vectors from huge data sets with billion of words, and with millions of words in the vocabulary has been initiated by measuring the quality of the resulting vector representations; however, those similar words can have multiple degrees of similarity [11]–[13].

Moreover, Mikolov et al. [13] uses a word offset technique where simple algebraic operations are performed on the word vectors; the resulting word vector is literally a result of those algebraic operations. For example, *vector("King") - vector("Man") + vector("Woman")* results in a vector that has a representation of the word *Queen*.

Skip-gram model is a variant of neural network language model (NNLM) and trained in two steps: firstly, continuous word vectors are learned using a simple model, and then the N-gram NNLM is trained on top of these distributed representations of words [11], [12]. To train skip-gram model on huge data sets, a large-scale distributed framework called DistBelief can be utilized [14].

III. RESEARCH METHODOLOGY

Principally, our approach consists of two steps; firstly, we make the data set ready to be input of the algorithms and secondly, we apply the algorithms to cluster topics automatically.

A. Preparing the Data Set

Firstly, final project abstracts from year 2000 to year 2013 in pdf format are converted into txt format by employing Apache Tika [15]. Number of successful converted and readable abstracts is 1,930. Typically, a final project has several information as follows: name, student id, title, supervisor 1, supervisor 2, and a topic.

The preprocessing that is applied into abstracts consists of removing page numbers, removing empty lines, and removing sentences containing several words ("Universitas Kristen Maranatha", "ABSTRACT" or "ABSTRAK", "DAFTAR ISI", "DAFTAR BAGAN DAN SKEMA", "DAFTAR TABEL", "DAFTAR LAMPIRAN", non-ASCII characters).

In order to evaluate our approach, our data set has already had labels assigned by psychology lecturers; in total, the number of topics is 85. Nevertheless, many topic names have the same concepts; for example, "pio", "psikologi industri organisasi", and "psikologi organisasi dan industri" are the same topic—i.e., *industrial organizational psychology*. We merge many names that have the same content with a help of an expert and come up with 6 general topics as explained in Table II.

B. Applying Skip-gram and LDA models into the Data Set

We run two algorithm settings—i.e., applying LDA algorithm without skip-gram and with skip-gram into final project abstracts. Specifically, we apply online learning to clustering final project topics [16] in our first setting. As for the second, we discover bigram language model from abstracts by utilizing skip-gram model algorithm, and then apply online learning to discover hidden topic structures [13].

Finally, the results of those algorithm settings are evaluated by an expert specializing in psychology final project topics.

TABLE II
EIGHTY-FIVE TOPICS MERGE INTO SIX GENERAL TOPICS.

1. Psikologi Pendidikan: *(Educational Psychology)*	**4. Psikologi Klinis:** *(Clinical Psychology)*
Psikologi Pendidikan-Sosial	Psikologi Klinis-Sosial
Psikologi Pendidikan-Perkembangan	Psikologi Klinis
Psikologi Pendidikan	Psikologi Klinis-Perkembangan
Psikologi Pendidikan-Industri	Psikologi Klinis-Positive
2. Psikologi Industri Organisasi: *(Industrial Organizational Psychology)*	Psikologi Klinis-Psikologi Industri dan Organisasi
Psikologi Industri dan Organisasi	**5. Psikologi Perkembangan:** *(Developmental Psychology)*
Psikologi Industri-Klinis	Psikologi Perkembangan
Psikologi Sosial-Industri	Psikologi Perkembangan-Sosial
Psikologi Industri-Perkembangan	**6. Lainnya:** (*Others*)
3. Psikologi Sosial: *(Social Psychology)*	Psikologi Eksperimen
Psikologi Sosial	Psikologi Kepribadian
Psikologi Lintas Budaya	Psikologi Positif
Psikologi Sosial-Lintas Budaya	Spiritual
Psikologi Sosial-Budaya	Positif-Integratif
Psikologi Sosial-Klinis	Psikologi Kesehatan
Psikologi Sosial-Perkembangan	Manajemen Industri

The expert compares the results with the six general psychology topics shown in Table II.

IV. RESULTS

In the first algorithm setting, we employ two experiments. The first experiment is executed by firstly removing words whose frequencies are one and more than 2,000. As shown in Table III, there are some similar words in several topics, for example, the word ”reliabilitas” which refers to reliability is shared in all topics. Moreover, words in each topic yet cannot distinguish themselves as keywords of the topic.

TABLE III
TEN WORDS WITH THE HIGHEST PROBABILITY FOR EACH TOPIC IN THE FIRST SETTING

Topic 1	**Topic 2**	**Topic 3**
aspek	universitas	derajat
teori	fakultas	kuesioner
validitas	kerja	validitas
hubungan	saran	rendah
teknik	derajat	efficacy
kota	psikologi	teknik
saran	teknik	work
responden	kuesioner	reliabilitas
deskriptif	reliabilitas	saran
reliabilitas	hubungan	responden
Topic 4	**Topic 5**	**Topic 6**
derajat	karyawan	kerja
efficacy	derajat	validitas
saran	kerja	aspek
kuesioner	kota	reliabilitas
validitas	deskriptif	saran
reliabilitas	dimensi	teori
dimensi	teori	uji
responden	kuesioner	kuesioner
rendah	hubungan	hubungan
teori	reliabilitas	responden

In the second experiment, we firstly remove all words in the intersection between every two topics. As shown in Table IV, LDA algorithm has discovered several words that are indeed keywords of each topic. Words in topic 1 such as kelas (*class*), studi (*study*), and universitas (*university*) are keywords of ”Educational Psychology” topic. Words in topic 2 such as *stress*, emosional (*emotional*), and *efficacy* are keywords of ”Clinical Psychology” topic. Words in topic 3 such as remaja (*teenager*), rumah (*home*), and motivasi (*motivation*) are keywords of ”Developmental Psychology” topic. Words in topic 4 such as sosial (*social*), korelasi (*correlation*), and perilaku (*behavior*) are keywords of ”Social Psychology” topic. Words in topic 6 such as pt (*company*), dukungan (*support*), and *sampling* are keywords of ”Industrial Organizational Psychology” topic. Finally, words in topic 5 is a mixed of the six general topics; therefore, topic 5 is assigned ”Others” topic.

TABLE IV
TEN WORDS WITH THE HIGHEST PROBABILITY FOR EACH TOPIC IN THE SECOND SETTING

Topic 1	**Topic 2**	**Topic 3**
kelas	efficacy	value
belajar	work	remaja
sma	perawat	motivasi
universitas	kompetensi	universitas
program	sumber	rumah
perusahaan	sampling	sampling
pengolahan	universitas	stres
motivasi	stress	studi
studi	pengolahan	emosional
rancangan	emosional	pengolahan
Topic 4	**Topic 5**	**Topic 6**
korelasi	anak	universitas
pengolahan	bidang	remaja
engagement	sma	korelasi
perilaku	remaja	anak
rancangan	style	pt
anak	pengolahan	rancangan
sampling	kelas	pengolahan
berkisar	sampling	sampling
sosial	universitas	berkisar
of	of	dukungan

We utilize two algorithms for our third setting—i.e. skip-gram model [13] and online LDA [16]. With the help of skip-gram model, online LDA can discover phrases that are keywords of each topic. Again, words in topic 1 such as kerja (*work*), profil (*profile*), and *individuated* are keywords of ”Industrial Organizational Psychology” topic. Words in topic 2 such as kemandirian emosional (*emotional intelligence*), kuesioner (*questionnaire*), and *rank spearman* are keywords of ”Clinical Psychology” topic. Words in topic 3 such as *purposive sampling* and berusia tahun (*how old are the sample?*) are keywords of ”Developmental Psychology” topic. Words in topic 4 such as mahasiswa (*student*), universitas x (*university x*), and universitas kristen (*christian university*) are keywords of ”Educational Psychology” topic. Words in topic 6 such as saran (*opinion*), subyek (*subjek*), and orangtua (*parents*) are keywords of ”Social Psychology” topic. Finally, words in topic 5 is a mixed of the six general topics; therefore, topic 5 is assigned ”Others” topic.

TABLE V
TEN WORDS WITH THE HIGHEST PROBABILITY FOR EACH TOPIC IN THE THIRD SETTING

Topic 1	**Topic 2**	**Topic 3**
aspek	item	sampel
derajat	kuesioner	profil
rendah	data	sesuai
telepon_genggam	kemandirian_emosional	populasi
orang	berdasarkan_pengolahan	peneliti
individuated	teori	menggunakan_metode
profil	holland	kuesioner
faktor	rank_spearman	metode
kerja	validitas	purposive_sampling
karakteristik	koefisien_korelasi	berusia_tahun
Topic 4	**Topic 5**	**Topic 6**
dimensi	rancangan	nokia
mahasiswa	brand_image	orangtuanya
universitas_x	mahasiswa	orangtua
derajat	minat	mahasiswa
psikologi	keputusan_membeli	saran
rendah	tipe	untuk_mengetahui
untuk_mengetahui	psikologi	derajat
mahasiswa_fakultas	kesimpulan	telepon_genggam
maranatha_bandung	faktor	anak
universitas_kristen	aspek	subyek

V. CONCLUSION

This paper intends to explore LDA and skip-gram model in order to automatically cluster topics on a collection of final project abstract documents. To the best of our knowledge, this combination of algorithms has never been explored before on a collection of final project abstracts, specifically in Indonesian language.

Furthermore, several experiments exhibit that LDA and skip-gram model have discovered specific keywords and keyphrases for each final project topic. However, we still need to define how to evaluate our approach more quantitatively instead of relying on domain experts. This issue shall be addressed in our future work.

Another future research direction we consider are comparing our model with TDE-TC [17], and LTSG [18]. Both TDE-TC and LTSG are recent models that utilizes Skip-gram and LDA models.

REFERENCES

[1] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, p. 391, 1990.

[3] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[5] W. Buntine and A. Jakulin, "Applying discrete pca in data analysis," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004, pp. 59–66.

[6] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[7] D. M. Blei and J. D. Lafferty, "Topic models," *Text mining: classification, clustering, and applications*, vol. 10, no. 71, p. 34, 2009.

[8] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," *Computer Science Department Faculty Publication Series*, no. 3, 2005.

[9] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 178–185.

[10] G. Heinrich, "Parameter estimation for text analysis," *University of Leipzig, Tech. Rep*, 2008.

[11] T. Mikolov, "Language modeling for speech recognition in czech," Ph.D. dissertation, Masters thesis, Brno University of Technology, 2007.

[12] T. Mikolov, J. Kopecky, L. Burget, O. Glembek *et al.*, "Neural network based language models for highly inflective languages," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4725–4728.

[13] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations." in *HLT-NAACL*, vol. 13, 2013, pp. 746–751.

[14] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le *et al.*, "Large scale distributed deep networks," in *Advances in neural information processing systems*, 2012, pp. 1223–1231.

[15] J. P. PDFBox, "Processing library," *Link: http://www. pdfbox. org*, 2014.

[16] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 856–864. [Online]. Available: http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf

[17] S. Hui and Z. Dechao, "A weighted topical document embedding based clustering method for news text," in *Information Technology, Networking, Electronic and Automation Control Conference, IEEE*. IEEE, 2016, pp. 1060–1065.

[18] J. Law, H. H. Zhuo, J. He, and E. Rong, "LTSG: latent topical skip-gram for mutually learning topic model and vector representations," *CoRR*, vol. abs/1702.07117, 2017. [Online]. Available: http://arxiv.org/abs/1702.07117