An Information-Theoretic Approach to Data Selection for Generative Topic Modeling

Michael Evan Santoso, Bhone Tay Zar Kyaw, Valentinus Roby Hananto, and Victor Kryssanov
Graduate School of Information Science and Engineering, Ritsumeikan University, Osaka 5678570, Japan
E-mail: {is0534is, is0645xv}@ed.ritsumei.ac.jp, hananto@fc.ritsumei.ac.jp, kvvictor@is.ritsumei.ac.jp

***Abstract*—Topic modeling is widely used to extract meaningful themes from large text corpora. Typically, it is assumed that all documents contribute equally to topic extraction, ignoring the fact that corpus composition directly determines the topic model performance. The entire document dataset is then processed without considering variations in the document quality that can affect the generated topic model performance. This paper introduces the concept of a topic model well-formed in terms of information theory, and proposes a data selection algorithm for topic modeling. Experimental results on the Reuters and Wikipedia corpora demonstrated that Latent Dirichlet Allocation (LDA) models trained on data selected using the well-formed topic model outperform baseline models. These results confirmed that removing noisy or unrepresentative documents from the generative process results in improving the topic quality. The paper briefly discusses the main findings and highlights directions for future study.**

I. Introduction

The exponential growth of text collections has necessitated the development of automated methods for extracting meaningful information from the texts to understand their underlying semantic structures and themes. To address this challenge, numerous NLP studies explored topic modeling techniques that help one obtain latent themes characterizing extensive textual data [1]–[3]. Topic modeling generally refers to the process of identifying human-interpretable and concise descriptions of document collections. The authors of [4] provided a representative survey of topic modeling applications, and emphasized the importance of generated topic quality, positioning it as a key element of topic modeling research.

Aiming to improve topic quality, several topic modeling methods have been proposed. Latent Dirichlet Allocation (LDA) [5] established a foundational framework for many probabilistic topic models, making document-topic and topic-word assignments with the Dirichlet distribution. Later advancements of LDA include Hierarchical Dirichlet Process (HDP) [6] that eliminated the fixed number of topics (i.e., topic size) requirement of LDA, and Correlated Topic Models (CTM) [7] that use the logistic normal distribution for handling inter-topic correlations. It should also be noted that over the years, various algorithms, such as batch-based online LDA [8], variational Bayes approaches [9], and online Gibbs sampling variants [10], were developed to improve the computational efficiency of generative topic modeling. As an alternative to the probabilistic framework, neural approaches to topic modeling employ embedding models to parameterize topic models (e.g., see [11] and [12]). Model-agnostic topic modeling methods then attempt to improve topic quality through data optimization [13]. All these methods, however, do not guarantee obtaining well-formed topics, and results of their application is often hard to unambiguously interpret or even to comprehend.

Another line of research focused on data selection methods, aiming to improve topic modeling outcomes through selection of "representative" or "important" texts for the model training corpus. This approach challenges the common assumption that all documents in a corpus would contribute equally to the finding of meaningful topics. It appears only natural to assume that the corpus composition directly impacts the topic model quality and performance, since noisy, uninformative, and irrelevant documents may degrade the coherence and interpretability of the generated topic model. Consequently, data selection methods become crucial, enabling selection of corpora that would enhance, rather than hinder, the topic model performance.

To the authors' best knowledge, however, research on data selection methods, remains scarce. One notable attempt in this direction by Zeng et al. [14] was to deploy the Zipf distribution as a mathematical basis for establishing a well-formed topic model for data selection to improve the modeling performance. This study, however, lacked a theoretical foundation as it did not justify the use of the Zipf distribution for modeling well-formed topics. Inspired by ideas formulated in [14], the presented work proposes a data selection algorithm, based on a well-formed topic model, using the apparatus of the information theory.

The projected contributions of this study are as follows:

1) A formal definition of a well-formed topic, based on information theory.
2) A data selection algorithm, based on the proposed well-formed topic model.

The remainder of the paper is structured as follows. Section II provides the definition of a well-formed topic. Section III describes the data selection algorithm, based on the proposed topic model. Section IV outlines the experimental setup used to validate the approach. Results obtained in the experiments are discussed in Section V, where limitations of this study and directions for future research are also formulated. Finally, Section VI summarizes the key findings.

979-8-3315-7206-8/25/$31.00

II. Well-formed topic model

A well-formed topic can be seen as a structured representation that encapsulates a coherent, meaningful, and interpretable theme. To elaborate further, a well-formed topic must possess semantic integrity and consistency, the attributes that mirror the cognitive capacities inherent in human comprehension. Semantic integrity here refers to the internal coherence of meaning within a topic, ensuring that all elements maintain meaningful connections. On the other hand, semantic consistency emphasizes clear and distinguishable boundaries between different topics. For the purposes of this study, we will assume that semantics (i.e., linguistic meaning) are conceptualized and expressed, using the notion of "concept", as it is defined in the semantic theory [15], [16]. Quite naturally, a concept can be viewed as an instance of encoded information [17]. A well-formed topic can then be defined as a set of concepts that contain information encoded efficiently.

To formalize the above definition, we will make the following assumptions:

1) Concepts are represented by words, the smallest units of information comprising texts.
2) Different concepts can share the same representation, and a single concept can have multiple (different) representations.
3) A topic is a mixture of concepts, with each concept contributing to the topic theme to a certain degree.

Let z be a discrete random variable denoting the frequency of words used to represent a single (fixed) concept. To characterize the behavior of z, we will seek to estimate $f_c(z)$, $z \in \mathbb{Z}^{0+}$, its probability mass function (PMF) that is the probability that a word appears with frequency z, representing concept c. The information theory then dictates that the least biased estimate possible on the given information is, in many cases, the distribution that maximizes the Shannon entropy $\mathcal{H}(z) = -\sum_{z=1}^{\infty} f_c(z) \ln(f_c(z))$ [18]. For a "closed" corpus (i.e. no new concepts to emerge from the text over time), the mean frequency $\bar{z} \equiv const.$ Also, $f_c(z)$, where $z > 0$ (otherwise, one would need to include into the model infinitely many words with zero frequency), must satisfy the normalizing condition $0 \leq f_c(z) \leq 1$. Under these constraints, one can derive the model $f_c(z)$ of a single concept that maximizes the amount of encoded information (see, for instance, [19] for derivation details):

$$f_c(z) = (e^{\lambda} - 1)e^{-\lambda z}. \tag{1}$$

Given that a topic is typically comprised of multiple concepts rather than a single one, a well-formed topic model can generally be written as follows:

$$f_{\text{wft}}(z) = \sum_{j=1}^{M} \phi_j \left(e^{\lambda_j} - 1\right) e^{-\lambda_j z}, \tag{2}$$

where M is the number of concepts in the mixture, and ϕ_j defines the contribution of j-th concept to the topic.

III. The data selection algorithm

The algorithm developed in this study is motivated by earlier efforts in Zipf distribution-based topic modeling [14], and it relies on the theoretical foundation of the maximum entropy principle [20]. The algorithm thus aims to select data while maximizing the amount of information (encoded with words) within each (latent) topic.

A. Overview

The data selection process consists of three steps: constructing word frequency vectors, fitting these vectors to the well-formed topic model, and evaluating the model's goodness-of-fit (see Fig. 1). First, frequencies of words in every document of the corpus are computed. Parameters of the well-formed topic model are then estimated, and the Akaike Information Criterion (AIC) [21] is used to select the best model. Finally, a topic quality metric is applied to assess how closely the word distribution of each document aligns with the expected distribution under the well-formed topic model assumption. The selection process is detailed in Algorithm 1, as it identifies documents that better align with the well-formed topic model (2). The algorithm thus defines a document filtering mechanism to improve the quality of the corpus used for generating a topic model.

B. Parameter estimation

A two-step Expectation-Maximization (EM) procedure [22], [23] is employed to obtain estimates of the parameters of the discrete hyperexponential distribution $f_{\text{wft}}(z)$. In the Expectation step, the expectation of the fitted distribution is calculated. $\xi_{i,k}$, the expectation of the well-formed topic model (2) is defined as follows:

$$\xi_{i,k} = \frac{\phi_k \left(e^{\lambda_k} - 1\right) e^{-\lambda_k z_i}}{\sum_{j=1}^{M} \phi_j \left(e^{\lambda_j} - 1\right) e^{-\lambda_j z_i}}, \tag{3}$$

where i is the data index for k-th component (i.e. concept) of the topic, and ϕ_k is the weight of the k-th component in the topic.

In the Maximization step, MLE parameter estimates of ϕ_k and λ_k are sought, using the following estimators:

$$\widehat{\phi}_k^{(o+1)} = \frac{1}{n} \sum_{i=1}^{n} \xi_{i,k}^{(o)}, \quad \widehat{\lambda}_k^{(o+1)} = \frac{\sum_{i=1}^{n} \xi_{i,k}^{(o)}}{\sum_{i=1}^{n} z_i \xi_{i,k}^{(o)}}, \tag{4}$$

where n denotes the number of data points; the superscript (o) indicates the current state, and $(o+1)$ - the updated state of the parameter estimate. The estimation is done iteratively

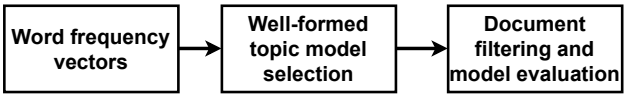

Fig. 1. Overview of the data selection process.

until the distribution parameters converge. To select the optimal model, the Akaike criterion $AIC\left(M,\phi,\lambda\right) = -2Q\left(\phi,\lambda\right) + 2M$, where $Q(\phi,\lambda)$ is the log-likelihood function, and M is the number of exponential mixtures in (2), is used.

C. Document filtering

The chi-square statistic χ is used to evaluate how well the model obtained through EM approximates the actual data distribution:

$$\chi\left(f_{EM}, f_{true}\right) = \sqrt{\frac{\sum_{z=1}^{n}\left[f_{EM}(z) - f_{true}(z)\right]^2}{n}}, \quad (5)$$

where z denotes the word frequency, f_{EM} is the model obtained, and f_{true} is the empirical distribution. A smaller χ value stands for a better agreement between the model and actual data, suggesting that the given document aligns better with the (well-formed) topic structure. The metric is further used as a selection criterion in the document selection algorithm (Algorithm 1).

IV. Experiments

A. Data

To evaluate the proposed algorithm, experiments were conducted, using two popular datasets: Reuters [24] and Wikipedia [1]. The Reuters corpus contains newswire documents from 1987, and it is commonly used for benchmarking in text classification and topic modeling. The set comprises 21,578 documents. After removing duplicates, 19,043 documents remained, featuring 120,673 unique tokens and an average document length of 425 words. The Wikipedia corpus consists of 1205 articles with 67,127 unique tokens and an average of 933 words per document. Both datasets were subjected to preprocessing that included text cleaning, tokenization, stop word filtering, part-of-speech filtering, lemmatization, and n-gram construction.

B. Evaluation metrics

The performance of topic models trained on the selected corpora was evaluated, using the perplexity and topic coherence (C_v) scores. Perplexity is a likelihood-based metric that allows for evaluating how well a topic model generalizes to unseen documents [5], [25]. It calculates the exponentiated average log-likelihood of held-out test data. Lower perplexity indicates better predictive performance and a closer fit to the underlying data distribution. The coherence score C_v [26] allows for assessing semantic similarity between words within topics, and between topics themselves. The C_v metric demonstrated the highest correlation with human judgment, compared to other topic coherence measures, and it presently serves as a standard evaluation metric to assess the topic model quality. Higher values of C_v indicate better topic model performance.

[1]https://github.com/blacfli/data-selection-algorithm/blob/main/dataset/wiki.csv

Algorithm 1 Data selection procedure

procedure WELL-FORMED TOPIC FIT(data $X : \{x_1, ..., x_i\}$)
  Initialize: weight parameter: $\{\phi_1, ..., \phi_{j-1}\}$, rate parameter: $\{\lambda_1, ...\lambda_j\}$, and set aic to be an empty list
  for all $j : \{1, \cdots, M\}$ **do**
    while $\phi^{(o)}, \lambda^{(o)}$ is not converged **do**
      $\phi^{(o+1)}, \lambda^{(o+1)} \longleftarrow$ calculate using (4)
    end while
    $aic \longleftarrow$ Compute $AIC\left(M,\phi,\lambda\right)$
  end for
  Best model $\widehat{M}, \widehat{\phi}, \widehat{\lambda} \longleftarrow \underset{aic}{\arg\min}\ (M,\phi,\lambda)$
  return $\widehat{M}, \widehat{\phi}, \widehat{\lambda}$
end procedure
procedure DOCUMENT SELECTOR(documents $docs$, threshold ρ)
  Initialize: set model-selected corpus (I) s to be an empty list
  for $i : \{0, \cdots, \text{len}(doc) - 1\}$ and each $doc \in docs$ **do**
    Sort the frequency of word in the document *doc*, and arrange them into vector $\boldsymbol{z}$ in descending order
    Build the word frequency distribution $f_{\text{true}}(z)$ by using vector $\boldsymbol{z}$
    Estimate the well-formed topic model $\widehat{f}_{\text{wft}}(z)$ parameters $(\lambda_1, ..., \lambda_M)$, and $(\phi_1, ..., \phi_{M-1})$ using **procedure WELL-FORMED TOPIC FIT** with the input vector $\boldsymbol{z}$
    if $\chi\left(\widehat{f}_{\text{wft}}(z), f_{\text{true}}(z)\right) \leq \rho$ **then**
      $s \leftarrow doc\,[i]$
    end if
  end for
  return s
end procedure

C. Experimental setup

Experiments were conducted to evaluate the effect of the proposed data selection on the generated topic quality. Three versions of the model training corpus were compiled: I) the proposed model-selected documents, II) the Zipf-based model-selected documents (replicating [14]), and III) a reference corpus containing all documents without selection. Corpus (I) was constructed as described in Algorithm 1.

Each of the three corpora was used to generate LDA models, using the Collapsed Gibbs Sampling (CGS) method [27]. The obtained models were evaluated against the reference corpus (which, thus, served as both, the training and testing data). The number of topics was varied from 1 to 200 with Dirichlet priors set to $\alpha = 0.1$ and $\beta = 0.01$, and an initial seed value of 100. Fig. 2 provides an overview of the experimental procedure.

V. Results and Discussion

First, three different selection threshold values of ρ have been considered (see Fig. 3): the first elbow ($\rho \approx 0.016$), the

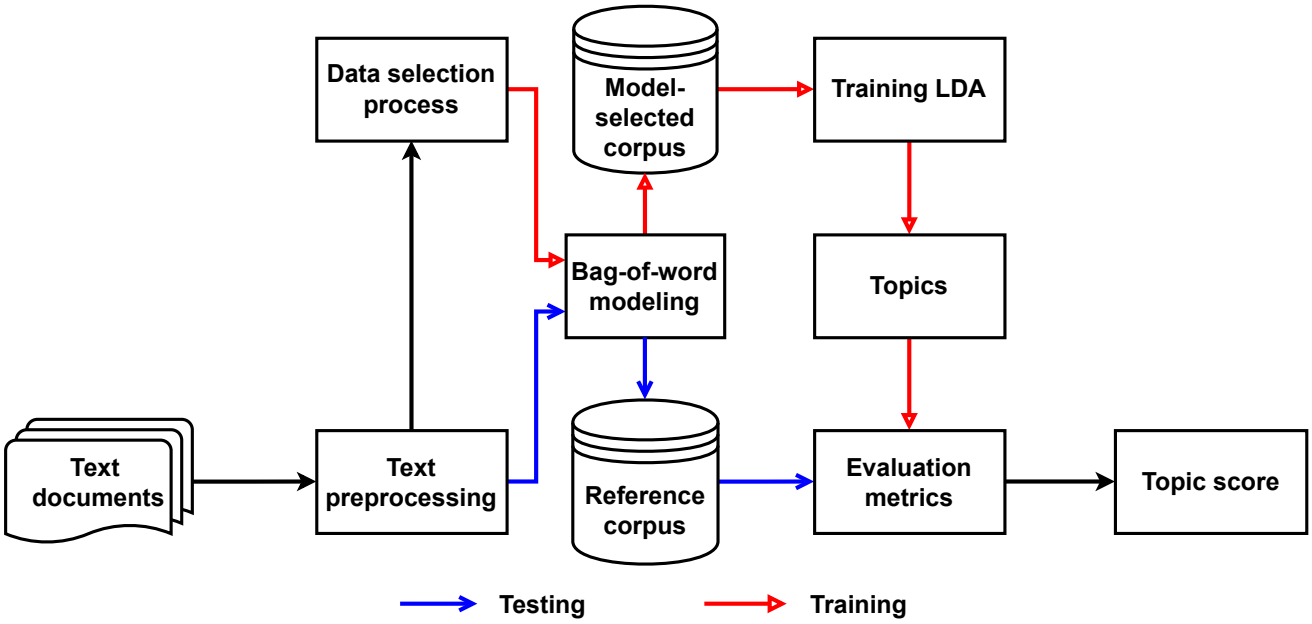

Fig. 2. Experimental framework for testing the proposed data selection algorithm.

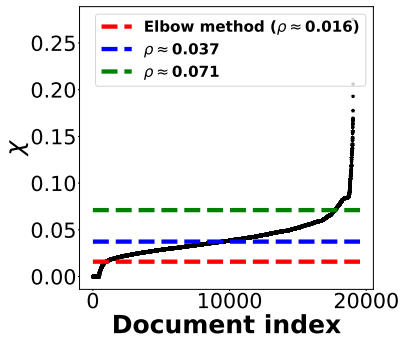

Fig. 3. Three different threshold values for data selection process.

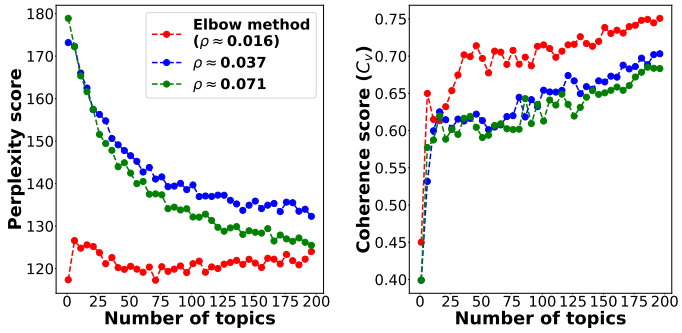

Fig. 4. Perplexity (left) and coherence C_v (right) scores for the three different threshold values (ρ) using the Reuters dataset.

midpoint between the first and second elbows ($\rho \approx 0.037$), and the second elbow ($\rho \approx 0.071$). As one can see from Fig. 4, the first elbow consistently yielded superior results on the Reuters dataset, achieving lower perplexity and higher coherence C_v scores across various topic counts. Therefore, $\rho \approx 0.016$ was used for further computing.

Results of the data selection are presented in Fig. 5, where the well-formed topic metric was used to decide on the inclusion of documents in corpus (I). The blue dots indicate the metric values of excluded documents, while the red dots—of those selected for the model generation. For the Reuters dataset, 952 documents were selected (out of 19,043), and for the Wikipedia dataset—61 articles (of 1,205). Table I provides a text example from the excluded and selected documents.

Figs. 6 and 7 compare the performance of three topic modeling approaches tested: LDA trained on the full dataset (original), LDA trained on the Zipf model-selected corpus, and LDA trained on the proposed model-selected corpus (i.e., hyperexponential pre-trained LDA). As one can see from the figures, the proposed approach allowed for outperforming the other models in almost all cases. For the Reuters dataset, hyperexponential pre-trained LDA consistently achieved lower perplexity and higher C_v scores across different counts of topics. For the Wikipedia dataset, it showed competitive results for the whole topic number range with superior coherence scores particularly high when the number of topics exceeded 50. Results obtained in the experiments thus allowed us to verify the assumption that not all documents in the corpus contribute equally to forming a topic. Manual inspection conducted randomly on documents excluded through the selection procedure confirmed that these documents typically have noisy or overly specific content that would, apparently, distort the topic formation (see Table I).

One limitation of the presented study is that the experimental

TABLE I
AN EXAMPLE OF EXCLUDED AND SELECTED DOCUMENTS FROM REUTERS DATASET

Removed document	Selected document
Computer Terminal Systems Inc said it has completed the sale of 200,000 shares of its common stock, and warrants to acquire an additional one mln shares, to <Sedio N.V.>of Lugano, Switzerland for 50,000 dlrs. The company said the warrants are exercisable for five years at a purchase price of .125 dlrs per share. Computer Terminal said Sedio also has the right to buy additional shares and increase its total holdings up to 40 pct of the Computer Terminal's outstanding common stock under certain circumstances involving change of control at the company. The company said if the conditions occur the warrants would be exercisable at a price equal to 75 pct of its common stock's market price at the time, not to exceed 1.50 dlrs per share. Computer Terminal also said it sold the technolgy rights to its Dot Matrix impact technology, including any future improvements, to <Woodco Inc>of Houston, Tex. for 200,000 dlrs. But, it said it would continue to be the exclusive worldwide licensee of the technology for Woodco. The company said the moves were part of its reorganization plan and would help pay current operation costs and ensure product delivery. Computer Terminal makes computer generated labels, forms, tags and ticket printers and terminals.	Pacific Southwest Airlines said its average load factor during February was 54.9 pct, down from 56.1 pct a year earlier. In the first two months of the year the load factor totaled 51.5 pct, down from 54.0 pct a year ago. Revenue passenger miles in February totaled 327.6 mln, compared to 295.5 mln. So far this year, revenue passenger miles totaled 640.2 mln, compared to 600.5 mln. Available seat miles in February totaled 596.6 mln, up from 526.8 mln a year ago. Year to date available seat miles totaled 1.24 billion, compared to 1.11 billion a year ago.

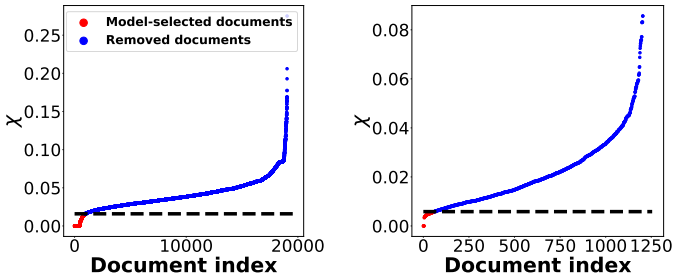

Fig. 5. Selection results score for the Reuters (left) and Wikipedia (right) datasets.

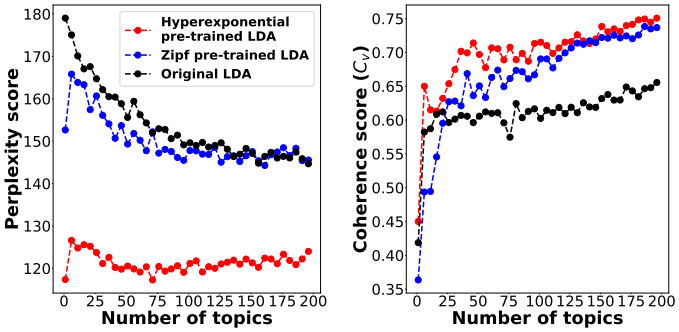

Fig. 6. Perplexity (left) and coherence C_v (right) scores of the tested topic models, using the Reuters dataset.

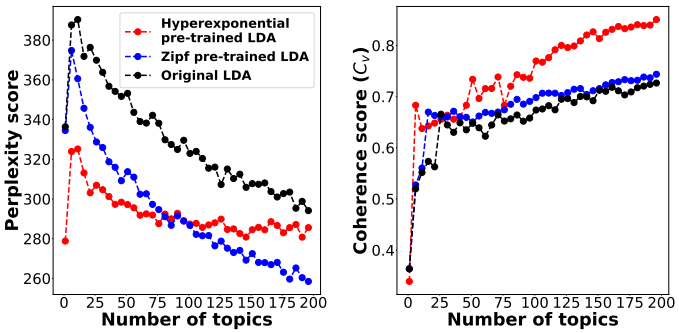

Fig. 7. Perplexity (left) and coherence C_v (right) scores of the tested topic models, using the Wikipedia articles.

evaluation was limited to LDA with a single set of hyperparameters. The generalizability of the proposed method to other topic modeling algorithms (e.g., CTM, HDP, and neural topic models) remains to be tested. Additionally, the data selection algorithm could not reliably handle short documents (with fewer than ten unique words), failing in parameter estimation. This could, however, be mitigated by removing short documents during preprocessing.

Finally, although the proposed method allowed for improving topic modeling through filtering the training data, the internal modeling process of LDA remained unchanged. As LDA models topics using a multinomial distribution over words, it often leads to overly general or less interpretable topics. Future work should investigate integrating the hyperexponential model directly into the topic-word modeling process, potentially leading to more semantically interpretable topic representations.

VI. CONCLUSIONS

The presented study proposed an approach to improving topic modeling through data selection, based on a well-formed topic model, which was also introduced in the study. To validate the approach, experiments have been conducted on the Reuters and Wikipedia datasets. It was shown that the deployment of the well-formed model for data selection allows for addressing topic composition issues that typically impede performance of probabilistic topic modeling methods, such as LDA.

This study, therefore, contributes to the field of computational text analysis by providing a formal definition of a well-formed topic. The proposed data selection algorithm utilizes this definition and allows for systematic corpus construction when building a topic model.

Future research will expand and extend the proposed approach to a broader range of topic modeling methods and datasets. Further investigation is warranted to replace the

multinomial distribution assumption in topic-word modeling with the proposed hyperexponential model to better capture semantic patterns. In addition, the current method limitation in handling short texts will require finding alternative solutions to incorporate their semantics in the topic model.

The source code and datasets used in this study are publicly available on https://github.com/blacfli/data-selection-algorithm

ACKNOWLEDGMENT

This work was supported, in part, by JST SPRING, Japan Grant Number JPMJSP2101.

REFERENCES

[1] R. Churchill and L. Singh, "The evolution of topic modeling," *ACM Computing Surveys (CSUR)*, 2021.

[2] U. Chauhan and A. Shah, "Topic modeling using latent dirichlet allocation: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1–35, 2021.

[3] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in sociology*, vol. 7, p. 886 498, 2022.

[4] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Information Systems*, vol. 112, p. 102 131, 2023.

[5] 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.

[6] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Sharing clusters among related groups: Hierarchical dirichlet processes," *Advances in neural information processing systems*, vol. 17, 2004.

[7] D. Blei and J. Lafferty, "Correlated topic models," *Advances in neural information processing systems*, vol. 18, p. 147, 2006.

[8] A. Banerjee and S. Basu, "Topic models over text streams: A study of batch and online unsupervised learning," in *Proceedings of the 2007 SIAM International Conference on Data Mining*, SIAM, 2007, pp. 431–436.

[9] M. Hoffman, F. Bach, and D. Blei, "Online learning for latent dirichlet allocation," *advances in neural information processing systems*, vol. 23, 2010.

[10] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 937–946.

[11] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[12] K. N. Keya, Y. Papanikolaou, and J. R. Foulds, "Neural embedding allocation: Distributed representations of topic models," *Computational Linguistics*, vol. 48, no. 4, pp. 1021–1052, 2022.

[13] A. Schofield, M. Magnusson, and D. Mimno, "Pulling out the stops: Rethinking stopword removal for topic models," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, short papers*, 2017, pp. 432–436.

[14] J. Zeng, J. Duan, W. Cao, and C. Wu, "Topics modeling based on selective zipf distribution," *Expert Systems with Applications*, vol. 39, no. 7, pp. 6541–6546, 2012.

[15] P. M. Pietroski, *Conjoining meanings: Semantics without truth values*. Oxford University Press, 2018.

[16] J. Szymanik, *Conjoining meanings: Semantics without truth values*, 2021.

[17] Y. Neuman, "A theory of meaning," *Information Sciences*, vol. 176, no. 10, pp. 1435–1449, 2006.

[18] E. T. Jaynes, "Information theory and statistical mechanics," *Physical review*, vol. 106, no. 4, p. 620, 1957.

[19] V. V. Kryssanov, F. J. Rinaldo, E. L. Kuleshov, and H. Ogawa, "Modeling the dynamics of social networks," in *E-Business and Telecommunication Networks: Third International Conference, ICETE 2006, Setúbal, Portugal, August 7-10, 2006, Selected Papers*, Springer Science & Business Media, vol. 9, 2008, pp. 40–51.

[20] D. K. Folye and E. Scharfenaker, "Bayesian inference and the principle of maximum entropy," *The American Statistician*, pp. 1–7, 2025.

[21] A. Chakrabarti and J. K. Ghosh, "Aic, bic and recent advances in model selection," *Philosophy of statistics*, pp. 583–605, 2011.

[22] M. Mialaret, P. Pereira, A. Sá Barreto, T. Pinheiro, and P. Maciel, "Automated phase-type distribution fitting via expectation maximization," *Journal of Reliable Intelligent Environments*, vol. 10, no. 4, pp. 339–355, 2024.

[23] Y. Teng and T. Zhang, "The em algorithm for generalized exponential mixture model," in *2010 International Conference on Computational Intelligence and Software Engineering*, IEEE, 2010, pp. 1–4.

[24] D. Lewis, "Reuters-21578 text categorization test collection," *Distribution 1.0, AT&T Labs-Research*, 1997.

[25] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. C. Lai, and R. L. Mercer, "An estimate of an upper bound for the entropy of english," *Computational Linguistics*, vol. 18, no. 1, pp. 31–40, 1992.

[26] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[27] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 569–577.