Journal of Computer Science

Original Research Paper

Poisson-Gamma Latent Dirichlet Allocation Model for Topics with Word Dependencies

[1,2]Ibrahim Bakari Bala and [1]Mohd Zainuri Saringat

[1]*Faculty of Computer Science and Information Technology,*
Universiti Tun Hussein Onn Malaysia, Parit Raja 86400, Johor, Malaysia
[2]*Universal Basic Education Commission, Wuse Zone 4, Abuja Nigeria*

Article history
Received: 06-08-2020
Revised: 02-12-2020
Accepted: 03-12-2020

Corresponding Author:
Mohd Zainuri Saringat
Faculty of Computer Science
and Information Technology,
Universiti Tun Hussein Onn
Malaysia, Parit Raja 86400,
Johor, Malaysia
Email: zainuri@uthm.edu.my

Abstract: This paper introduces the Poisson-Gamma Latent Dirichlet Allocation (PGLDA) model for modeling word dependencies in topics modeling. The Poisson document length distribution has been used extensively in the past for modeling topics with the expectation that its effect will fizzle out at the end of the model definition. This procedure often leads to downplaying the effect of word correlation with topics and thus reducing the precision or accuracy of retrieved documents in such a situation. Therefore, we propose a new class of model that relaxes the words independence assumption in the existing Latent Dirichlet Allocation (LDA) model by introducing the Gamma distribution that can capture the correlation between adjacent words in a document. The Poisson document length distribution and Gamma correlation distribution are then convoluted to form a new mixture distribution for modeling word dependencies. Model parameter estimation was achieved via Laplacian approximation of the log-likelihood. The new model was then evaluated using the 20 Newsgroups and AG's News datasets. The applicability of the model was assessed using the F_1 score. The results of the evaluation showed appreciable supremacy of PGLDA over LDA.

Keywords: Poisson Distribution, Gamma Distribution, Topic Model, Latent Dirichlet Allocation

Introduction

A topic is defined as a random variable with a unique probability distribution over a fixed vocabulary (Jiang *et al.*, 2015; Wang and Zhang, 2016; Chen, 2017). A topic is made up of different words in a vocabulary. In the same vein, a document is also made up of several topics. The most important thing about topic modeling is determining the distribution of topics over the document and consequently determining the distribution of words over each topic. Mathematically, topic modeling involves working with the *N X K matrix of document and topics and subsequently K X V matrix of topics and words, where N*, *K*, *V* are the number of documents, topics and words, respectively (Liu *et al.*, 2016; Zhao *et al.*, 2019).

The first step in topic modeling is to define a generative process for simulating documents. Each document is simulated as follows: For each word in a document; choose a topic assignment and subsequently select a word from the topic. LDA and PLSA are the foundation models in topic modeling, but more valid and relevant models have been developed in recent times (Liu *et al.*, 2016). Thus, to develop an extended topic model, it is crucial to understand LDA.

In PLSA, *d* represents the document identity, while the topic is defined as *z* the word is represented by *w* word and N_d is the size of a *d* which can be colloquially referred to as the number of words in a specified document. The conditional distribution $P(z|d)$ is defined for topic *z* in the document *d* while $P(w|z)$ is defined over words *a* in topic *z*. Thus, the PLSA algorithm below can be used to model words in documents.

Algorithm 1: PLSA Algorithm

```
1) For each document $d \in \{1,2,3,\ldots,N\}$:
2) For each word $w \in d \in \{1,2,3,\ldots,N\}$:
3) Simulate $z \sim P(z|d)$
4) Simulate $w \sim P(w|z)$
```

In the contrast, for LDA, the two conditional probability distributions, $P(z|d)$ and $P(w|z)$ are presumed to follow multinomial distributions such that the topics in the entire documents have common Dirichlet prior

Science Publications

distribution $P(\alpha)$ and the word conditional distributions on topics have common Dirichlet prior $P(\beta)$ (Xue, 2019). The step proceeds by selecting appropriate prior parameters α and β for a document *d*, which will in-turn formed a conditional distribution of *K* topics with parameters θ.

The distribution for conditional distribution follows a multinomially distributed random vector that originated from a Dirichlet distribution $Dir(\theta|\alpha)$. Similarly, for *k* topic, *V* words conditional distribution are formed and it also follows the multinomial distribution $Mult(w|z,\beta)$. Parameter estimation and hypothesis testing for the LDA model can be readily achieved due to the conjugacy property existing between Dirichlet distribution and the multinomial likelihood. The graphical visualization for the PLSA model is shown in Fig. 1 while that of the LDA model is shown in Fig. 2.

Poisson document length distribution has been used extensively in the past for modeling topics with the expectation that its effect will fizzle out at the end of the model definition (Wang and Zhang, 2016; Wang *et al.*, 2016). Also, the Poisson assumption implies that the words in a document are independent or unrelated to another. This assumption is unrealistic in nowadays topic modeling. Group membership often occurs in modeling words in texts, thus violating the Poisson assumption (Inouye *et al*., 2014b). Inouye *et al.* (2014a) showed that some words that serve as hub word exist, which in turn determines the kinds of words that will follow in a document.

Inouye *et al*. (2014a) proposed the Poisson Markov Random Field (PMRF) model to model dependencies of words. The approach defines the conditional distribution of current words using previous words. It also assumed that the parameter of word occurrence is a multivariate distribution that can be modeled using the Generalized Linear Model (GLM). The drawback of the approach is the complexity of the method of estimation arising from the use of multivariate distribution for the several rate parameters of words in the model. Therefore, there is a need to model the dependencies of words in topics arising from several documents with a simpler model that can be easily estimated.

Algorithm 2: LDA Algorithm

```
1) Simulate $N$ documents from Poisson; $Pois(N = n|\xi)$
2) For each topic $k \in \{1,2,3,\ldots,K\}$:
3) For each document $d \in \{1,2,3,\ldots,N\}$:
4) Simulate $\theta_d \sim Dir(\theta_d|\alpha)$
5) For each word $w \in d \in \{1,2,3,\ldots,K\}$:
6) Simulate $z_{dn} \sim Mult(z_{dn}|\theta_d)$
7) Simulate $w_{dn} \sim Mult(w_{dn}|z_{dn}\ \beta)$
```

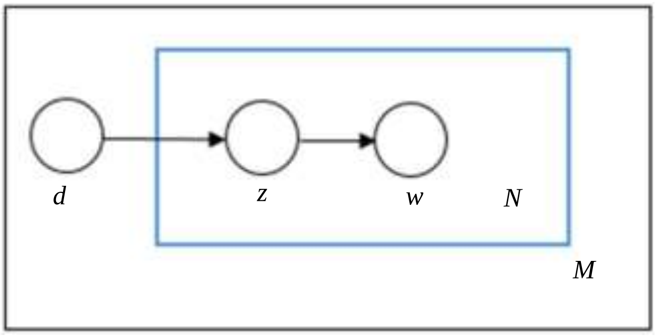


Fig. 1: Graphical model representation of PLSA. The boxes are "plates" representing replicates. The first layer plate (*blackpa lines*) represents documents, the second layer plate (*blue lines*) represents the repeated choice of topics and words within a document

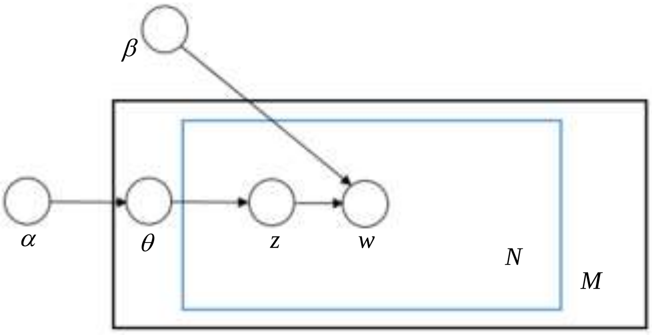


Fig. 2: Graphical model representation of LDA. The boxes are "plates" representing replicates. The first layer plate (*black lines*) represents documents, the second layer plate (*blue lines*) represents the repeated choice of topics and words within a document and the third layer plate (*brown lines*) represents the latent layer of document length that describes hidden relation of document length *N* and *z* or *w*

This paper aims to extend the original LDA to capture the word dependencies feature by convoluting the Poisson document length distribution and Gamma correlation distribution. The mixed distribution with LDA gave birth to the new document model, which we termed as PGLDA. PGLDA performance was tested using the popular 20 newsgroups and AG's News datasets (Wang and Zhang, 2016; Jiang *et al*., 2018; Albishre *et al*., 2015; Del Corso *et al*, 2005). The datasets have been used to classify the newsgroups using LDA.

Materials and Methods

Wallach (2006) presented one of the notable earliest modifications of the LDA model in terms of modification of the exchangeability assumption used by Bag-of-Word (BoW) models. Wallach (2006) developed a model that assumes there exists a correlation between adjacent topics. The approach involves the use of a hierarchical procedure of combining the latent topic models and *n*-grams statistical procedure. The author specifically extended the unigram topic modeling procedure to Dirichlet hierarchical bigrams model. Wallach (2006) reported that the combination of the unigram and bigram Dirichlet models is better than either of the two. The author's conclusion was inferred from the analysis of datasets consisting of 150 documents each. Gruber *et al.* (2007) corroborated (Wallach, 2006) in their paper by concluding that the exchangeability assumption is not practical and rare in real-life document modeling, especially when dealing with the contextual meaning of words. Hu *et al.* (2014) provided an alternative class of model that negates the belief of either using exchangeability assumptions or relaxing them. The authors concluded that the class of models only makes *apriori* fixes and not interactive, which makes them not applicable to most real-life document modeling.

Reisinger and Mooney (2010) focused on the modification of LDA to accommodate modeling of word absences. The LDA model was modified by updating the likelihood function to ensure the capturing of rare words. The multinomial likelihood was replaced with the Von-Mises Fishers distribution for the sampling of topics.

Furthermore, (Inouye *et al*., 2014a) differentiate between inter-topic correlation and intra-topic correlation. The authors reported that most of the existing models focused on inter-topic correlation rather than intra-topic correlation which is often inherent in long text documents. They define this class of correlation as word dependencies in the presence of a "hub" word. For example, in a document; *"The temperature of Johor today is high"*. The topic of the document is *"temperature"* and as well a word within the document. The hub word here is *"temperature"* as it is serving the dual purpose of being the word as well as the topic. This is the intra-topic correlation defined by (Inouye *et al*., 2014a).

In recent times, researchers have used the LDA in the field of sentiment analysis and information retrieval in general (Santosh *et al*., 2016). Ren and Hong (2017) used extracted topics from online travel reviews to perform Topic-based sentiment analysis. The objective of the research was to determine the most important to the tourist from topics and emotions. The performance of the LDA-based feature selection approach was investigated by (Onan *et al*., 2016) in the area of text classification. Sentiment classification was done via optimal latent topic that was obtained from the combination of machine learning-based classifiers and LDA to obtain the optimal number of latent topics. Sentiments analysis of Twitter expression was performed using encoded information of topics by word embedding (Ren *et al*., 2016). Tweets were first generated using LDA before the incorporation of the topic function. The system performance recorded about 4% improvement when the topics were integrated into word embedding. Hong *et al*. (2017) presented an LDA-based learning system for updating the civil aviation domain system. The representation content was enriched by the system making the information to provide better support for the management of the emergency system. A similar study by (Ko *et al*., 2017) used the LDA-based procedure for product opportunities. In their work, customers' needs changes were monitored by identifying the product opportunity preference. However, the topics that were generated by the LDA-based techniques returned topics with irrelevant words. Also, as observed from another similar study, the LDA-based approach fails to capture the semantic correlation between adjacent words. Therefore, (Zhang *et al*., 2019) suggested the use of a preliminary feature representation method with LDA for the identification of a topic.

Santosh *et al*. (2016) presented a new performance improvement approach for LDA. They first used an ontology approach to identify appropriate features after clustering and showed that the accuracy of the feature extraction largely improved. Ali *et al*. (2017) presented an ontology-based, feature-level sentiment analysis for describing the relationships between concepts in a specific domain.

Another class of BoW model is the Restricted Boltzmann Machines (RBMs) which is a probabilistic graphical model used for modeling of topics (Gupta *et al*., 2019). The RBM has been proven to be good in the representations of distributed latent on the input data and performed exemplarily well in clustering and information retrieval tasks. However, it was found to be inapplicable in modeling documents of different lengths and thus making model training hard and unstable. For undirected models, like RBM, marginalizing over latent variables is generally an intractable operation, which makes modeling far more difficult (Gupta *et al*., 2019). Larochelle and Murray (2011) solved the problem by introducing a feed-forward neural network called Neural

Autoregressive Distribution Estimator (NADE), which was inspired by RBM, but it is asymmetrical in structure. As a further extension, (Larochelle and Lauly, 2012; Gupta *et al*., 2019) proposed an extension of the NADE model and named it Document Neural Autoregressive Distribution Estimator (DocNADE) and which can learn interpretable representations of texts in a document collection using an unsupervised learning approach. Like NADE, the model architecture is also based on a feed-forward procedure that learns the probability distribution of the bag-of-words representation of documents. It uses the same autoregressive connections for the visible softmax as well as hidden layers.

DocNADE (Larochelle and Lauly, 2012) learns word occurrences across the whole document, i.e., coarse granularity (in the sense that the regeneration probability of a word in a document strongly depends on the context of the previous word). However, since DocNADE is based on the BoW assumptions, the contextual meaning of words is also ignored. To tackle this problem of missing contextual meaning in topic models, (Gupta *et al*., 2019) incorporated language structure information using Long Short-Term Memory (LSTM) based Language Model (LSTM-LM), thereby accounting for word order (semantics) and language concepts (syntax). This allows for the combined use of global context, i.e., coarse granularity, from DocNADE model, without word order information and local context, i.e., fine granularity, from LSTM-LM. Gupta *et al*. (2019) named the model as contextualized-Document Neural Autoregressive Distribution Estimator (ctx-DocNADE). ctx-DocNADE helps in learning complementary semantics by combining language and latent topic learning in a unified neural autoregressive framework.

Furthermore, to tackle sparsity of words and improving the meaningfulness of predictions from ctx-DocNADE, (Gupta *et al*., 2019) combined ctx-DocNADE with embeddings model Global Vectors (glove) (Pennington *et al*., 2014) to form a new model called ctx-DocNADEe where the letter “e” denotes embeddings. The performance of ctx-DocNADEe was found to be slightly better than ctx-DocNADE and moderately better than DocNADE. In the same vein, the GPLDA proposed by (Bala and Saringat, 2019) was developed to tackle the lack of inherent word dependencies (word correlation) structure in LDA as shown in Fig. 3. In GPLDA, word dependency was conjectured as a problem that resulted from having unequal document lengths across documents, which is the area of strength of Generalized Poisson in terms of modeling over or under dispersed data. Over or under dispersed Poisson process implies that the events are not independent, this in this case is the document. However, it was later observed that there are still some interpretability issues in terms of the GPLDA predicted word coherence score. Although, the applicability aspect yet yielded significant improvement over LDA but couldn’t compete with the recent models such as DocNADE and its variants- (ctx-DocNADE and ctx-DocNADEe). This led to the proposition of two distributions (Poisson and Gamma) in this study, which maintained the structure of LDA and as well captured the word dependencies. The Poisson distribution in the Poisson-Gamma mixture behaves similarly to the standard Poisson in LDA while the Gamma distribution captures the word dependencies in terms of correlation. This mixture distribution was found to be more valid than Generalized Poisson in terms of applicability and interpretability, hence the new for the current model presented in this study. Figure 3 shows the word correlation with using the 20 Newsgroups dataset.

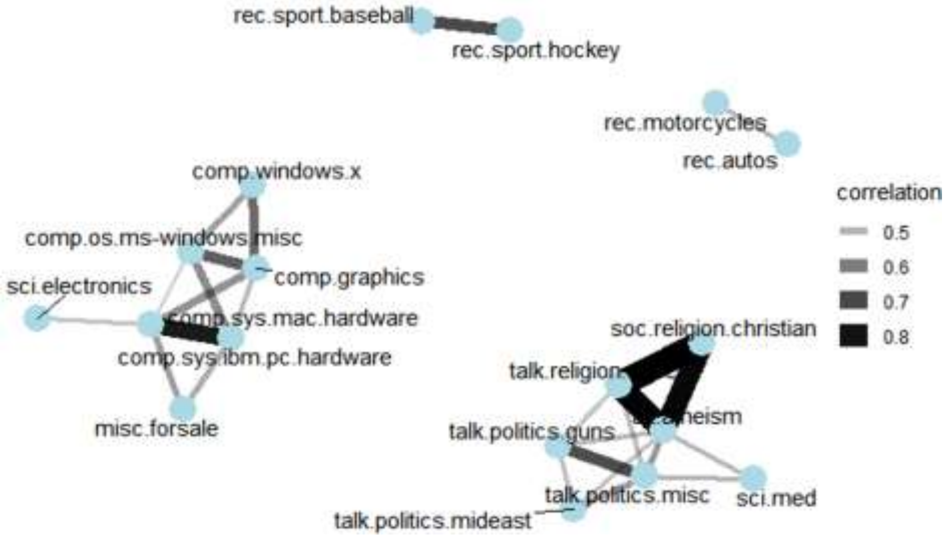


Fig. 3: Word correlation for topics with a correlation greater than 0.4

In summary, the previous works are based on traditional LDA approaches that do not incorporate word dependencies (word correlation). Thus, in this study, we propose a Poisson-Gamma LDA-based topic modeling method for document classification.

Poisson-Gamma Mixture

The original LDA procedure by (Blei *et al*., 2003) strongly relies on the following assumptions:

i. The number of topics k, which is the dimension of the Dirichlet distribution is fixed
ii. The word probabilities parameterized by $k \times V$ matrix β with elements defined as; $\beta_{ij} = P(w^j = 1|z^j = 1)$
iii. The document length N is independent of all other data generating process

Given N documents, following from LDA with the assumed probability of n document at a specific time interval distributed as Poisson, the probability of N assuming n is:

$$P(N=n\,|\,\xi)=\frac{\exp(-\xi)\xi^n}{n!},n=0,1,2,...$$

Under assumption (iii), the Poisson parameter ξ (the rate of documents at a specific time) is assumed to be fixed and unrelated to other model parameters such as words or topics. The Poisson-Gamma Mixture case ξ is assumed to be a latent random variable and follows a Gamma distribution with parameters (b, a). Thus, the probability density function can be defined as;

$$P(\xi\setminus a,b)=\frac{a^b\exp(-\xi a)\xi^{b-1}}{\Gamma(b)},\xi,a,b>0$$

where, a, b are the latent parameter that captures the interdependence (correlation) between documents lengths and topics or words. Thus, the joint probability of N assuming n and the latent variable is:

$$P(N=n,\xi\,|\,a,b)=P(N=n\,|\,\xi)\times P(\xi\,|\,a,b)$$

$$P(N=n,\xi\,|\,a,b)=\frac{\exp(-\xi)\xi^n}{n!}\times\frac{a^b\exp(-\xi a)\xi^{b-1}}{\Gamma(b)}$$

$$P(N=n)=\frac{a^b}{n!\Gamma(b)}\frac{\Gamma(n+b)}{(a+1)^{n+b}}$$

$$P(N=n\,|\,a,b)=\frac{\Gamma(n+b)}{\Gamma(n+1)\Gamma(b)}\left(\frac{a}{a+1}\right)^b\left(\frac{1}{a+1}\right)^n$$

Extended LDA (PGLDA)

The PGLDA follows from the earlier derived Poisson-Gamma mixture by replacing the Poisson sampling distribution of N documents in LDA with the Poisson-Gamma mixture. The structure of PGLDA is given in Fig. 3.

Figure 3 differs from Fig. 2 of LDA in terms of the distribution of document length. The generating process of PGLDA goes thus:

1) Sample ξ from gamma distribution $G(b,a)$
2) Sample N from Poisson-Gamma Mixture $P(N = n|a,b)$
3) Sample θ from dirichlet distribution $Dir(\alpha)$
4) For each N words w_n:
   a) Sample topic z_n from *multinomial*(θ)
   b) Sample a word w_n from the conditional distribution of topic and latent distribution of β

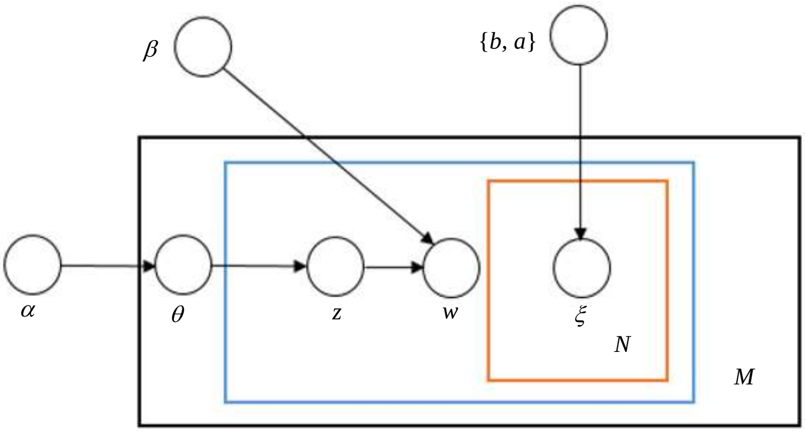


Fig. 3: Graphical model representation of PGLDA. The boxes are "plates" representing replicates. The first layer plate (*black lines*) represents documents, the second layer plate (*blue lines*) represents the repeated choice of topics and words within a document and the third layer plate (*brown lines*) represents the latent layer of document length that describes hidden relation of document length N and z or w

PGLDA relaxes assumption (iii) of LDA and carries over the first two assumptions as; (i) the number of topics k, which is the dimension of the Dirichlet distribution is fixed and (ii) The word probabilities parameterized by $k \times V$ matrix β with elements defined as; $\beta_{ij} = P(w^j = 1|z^j = 1)$. The relationship between ξ and topic or word parameter is not direct but exists and it is captured in the extraneous latent parameters (b,a) of the Gamma distribution. PGLDA is more flexible and realistic when compared to LDA.

Parameter Estimation of PGLDA

The Laplace approximation technique of approximating the posterior distribution in Bayesian inference is employed here. The technique involves obtaining the log-likelihood of the distribution.

The Laplace procedure starts by determining the first and second partial derivatives with respect to the parameters using the $\log[P(D|\theta, z, w, \alpha, \beta, b, a)]$. The derivatives are intractable and thus the iterative approximation solution was used. The Laplace approximation technique used here is summarized as follows:

(i). Calculate the log-likelihood for the marginal distribution of D using the $\log[P(D|\theta, z, w, \alpha, \beta, b, a)]$

(ii). Determine the first derivative with respect to each parameter in the parameter space $\Omega = \{\theta, \alpha, \beta, b, a\}$ and find the iterative estimate of parameter Ω using:

$$\Omega_{t+1} = \Omega_t - \frac{\log\left[P\left(D \mid z,w,\Omega\right)\right]}{\partial \log\left[P\left(D \mid z,w,\Omega\right)\right] / \partial \Omega}$$

The process continues until $|\Omega_{t+1}-\Omega_t| \leq \varepsilon$ where $\varepsilon \rightarrow 0$.

Algorithm 3: Pseudocode of PGLDA Algorithm

```
Sample $\xi$ from Gamma distribution $G(b,a)$
Sample $N$ from Poisson-Gamma Mixture $P(N = n|a,b)$
For each topic $k \in \{1,2,3,\ldots,K\}$:
For each document $d \in \{1,2,3,\ldots,N\}$:
Simulate $\theta_d \sim \mathrm{Dir}(\theta_d \mid \alpha)$
For each word $w \in d \in \{1,2,3,\ldots,N\}$:
Simulate $z_{dn} \sim Mult(z_{dn}|\theta_d)$
Simulate $w_{dn} \sim Mult(w_{dn}|z_{dn},\beta)$
```

Experiment and Evaluation

The convergence of the PGLDA is observed by simulating several Poisson-Gamma mixture with varying parameter $b = \{1,2,3, \ldots, 10\}$ for the case with $b \geq 1$ and $b = \{0.1,0.2,0.3, \ldots, 1\}$ for the case with $b \leq 1$ and fixed-parameter a and rate parameter $\xi = 5$. The Poisson-Gamma mixture was generated by first sampling ξ from the Gamma distribution with parameter $\{b, a\}$ and then sampling document N from Poisson distribution using ξ initially sampled using $G(b, a)$. Figure 4 shows the convergence results. All analyses were achieved using the *R* package version 3.6.1 (2019-07-05) on a 64 bit system with CPU @ 1.60 GHz and 8GB RAM.

Figure 4 shows that the higher the value of b the closer the density of Poisson-Gamma mixture to Poisson and likewise PGLDA to LDA. However, as shown in Fig. 5, it will be highly inaccurate to assume a Poisson distribution as there exists large disparity in the density curve. The density curve of Poisson appears to flatten out and more centered on the average document length $\xi = 5$, however for $b \leq 1$; the density is more centered on 0 even when the average used in the simulation was 5.

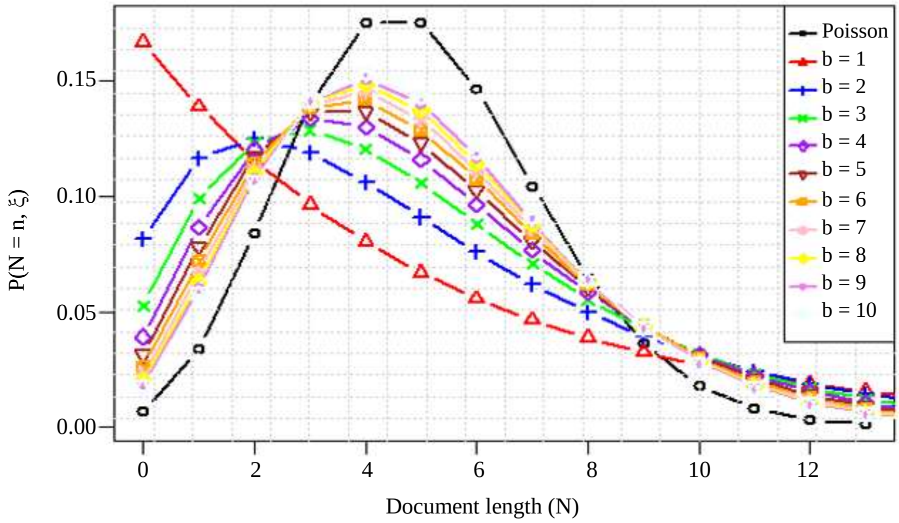


Fig. 4: Document length distribution at various mixing parameter $b = \{1,2,3, \ldots, 10\}$. The plot confirms that as $b \rightarrow \infty$ the Poisson-Gamma mixtures collapse to Poisson distribution and consequently the PGLDA will subsume to LDA

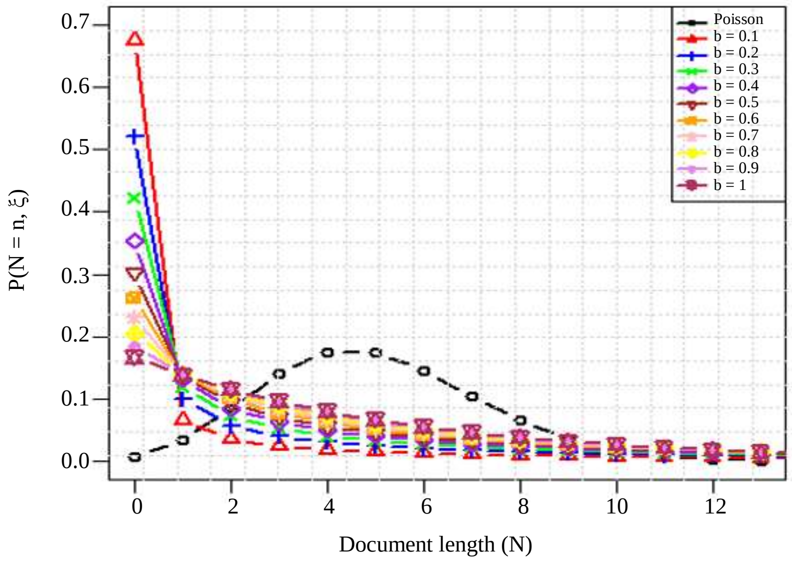


Fig. 5: Document length distribution at various mixing parameter *b* = {0.1,0.2,0.3, …, 1}

Real-Life Data Experiment

The 20 Newsgroups and the AG's News datasets (Albishre *et al*., 2015; Del Corso *et al*., 2005) were used because they are commonly employed to evaluate the performance of text categorization and text clustering algorithms. The 20 Newsgroup dataset contains 18,846 documents, covering 20 different categories. The topics in the classes are very diverse, including sports, politics and religion. For validation, 11,314 documents from a total of 18,846 documents were used for training and the remaining 7,532 documents were used for testing. Also, the AG's News dataset was constructed using the four largest classes from the original corpus. Each of the classes contains 30,000 training samples and 1,900 testing samples corresponding to a total of 120,000 training documents and 7,600 test documents. The four categories are world, business, science and technology and sports news. Performance evaluation, as in (Jamil *et al*., 2017), was employed to analyze the efficiency of the algorithm.

We have used the interpretability (topic coherence) and applicability (Information Retrieval and classification) performance measures for the evaluation of PGLDA and existing models.

Interpretability: Topic Coherence Score

Topic models help to understand, summarize and organize a large collection of documents by finding some latent features called topics. Topics are a collection of words based on co-occurrence statistics in the document corpus. While it is important to evaluate topics models on the criterion of generalization and applicability, it is equally important to have a quantifiable evaluation of these latent topics learned by the model to distinguish good topics from bad topics. Therefore, we use topic coherence as the criterion to assess the meaningfulness of the underlying topics captured by the model. We use the coherence measure used by (Gupta *et al*., 2019), which identifies context features for each topic word using a sliding window over the reference corpus. The topics with high scores imply more coherency. The coherence score is calculated using the *Normalized Point-wise Mutual Information* (NPMI). The formula is given as:

$$NPMI\left(\omega_i,\omega_j\right)=\left\{-\frac{\log\dfrac{P\left(\omega_i,\omega_j\right)}{P\left(\omega_i\right)\cdot P\left(\omega_j\right)}+\varepsilon}{\log\left[P\left(\omega_i,\omega_j\right)+\varepsilon\right]}\right\}^{\gamma}$$

where, ω_i, ω_j are *i*th and *j*th words respectively whose similarity is to be obtained, γ is the shape parameter that determines the increase or decrease in coherence score ε is the random error that ensures $\log[P(\omega_i, \omega_j)]$ is computable when $P(\omega_i, \omega_j) = 0$. The coherence score is calculated over the top 10 words for each of the optimized numbers of topics generated by PGLDA.

Applicability

In this section, I considered two performance metrics which are information retrieval and classification accuracy.

Information Retrieval (IR)

When it comes to the practicality of topic modeling, document retrieval is a critical evaluation. Suppose we have a query document, document retrieval is defined as finding the most semantically related documents in a given document corpus. Document retrieval is a form of information retrieval where a higher-level document representation, i.e., latent vector representation, is used for retrieval tasks. For the topic models, this higher-level representation of a document is, generally, a topic mixture representation, i.e., a vector with mixture coefficients for all latent topics learned by the model. Therefore, it is important to learn the vector representation of the two most semantically related documents in such a way that the similarity distance between the vector representation of the two documents is very less as compared to other semantically unrelated documents. The similarity distance can be either cosine similarity or Euclidean distance. Hence, a topic model needs to learn all the different types of semantics present in a document corpus. For our proposed models, we call these vector representations contextualized representations. The criterion used is Precision (P) (also referred to as positive predictive value) is the proportion of relevant cases among the retrieved cases. At the same time, Recall (R) (also referred to as sensitivity) is the proportion of the total amount of relevant cases that were retrieved. The detailed description is found below under classification.

Classification (F_1 Score)

To get a list of most related text documents for a given query is a very important task, but equally important is the classification of text documents into a predefined set of different categories, i.e., text categorization. Text categorization does not require the presence of a query document, but it is done on an absolute scale. It gives one or more tag(s) to each document based on its semantic information which eventually put each document in different categories, hence reducing cluttering and facilitate easy search and navigation of the user. For example, action, adventure, thriller, romantic, etc. are different tags that can be given to each movie plot (text) which will categorize them into different genres. In topic modeling, text categorization can be done in two ways. First, during training, the label information can be leveraged to perform supervised classification along with unsupervised regeneration of documents. Second, after learning latent document topic representations in an unsupervised fashion, use those representations as static input data to perform supervised classification. We adopt the second method to perform text categorization using contextualized representations of our proposed models and document representations of all other baselines models we have used.

The performance of the proposed models will be compared with the existing models using the confusion matrix-based scores such as accuracy, recall rate and F_1 score.

Table 1: Confusion matrix

	True class		
Predicted class	Relevant	Not relevant	Total
Retrieved	TP	FP	P
Not retrieved	FN	TN	N
Total	P*	N*	T

where, TN represents True Negative, FP is the False Positive, FN represents False Negative and TP is the True Positive. Also, N* is the total predicted negative and P* represents the total predicted positive. Similarly, N is the total actual negative, while P is the total actual positive. T represents the total number of observation equivalent to:

$$T = TN + FP + TP + FN$$

$$Accuracy\left(A\right) = \frac{TN + TP}{T}$$

$$Recall\,Rate\left(R\right) = \frac{TP}{TP + FN}$$

$$Precision\,Rate\left(R\right) = \frac{TP}{TP + FP}$$

The F_1 is a measure of the accuracy of the test dataset and is defined as follows:

$$F_1\ Score = \frac{2 \times R \times P}{R + P}.$$

The final F_1 Score termed as **micro** F_1 for a specific dataset is simply the average of the score over the number of classes or topics in a dataset. It is used here as the performance measure for the relative comparison of various methods. Formally the average F_1 Score $\left(\bar{F}_1\right)$ is calculated using:

$$\bar{F}_1 = t^{-1}\sum_{i=1}^{t} F_{1i}.$$

We used *R* package version 3.6.1 (2019-07-05) on a 64 bit system with CPU @ 1.60 GHz and 8 GB RAM for data extraction, pre-processing, partitioning and model building. Figure 6 presents the analysis flow.

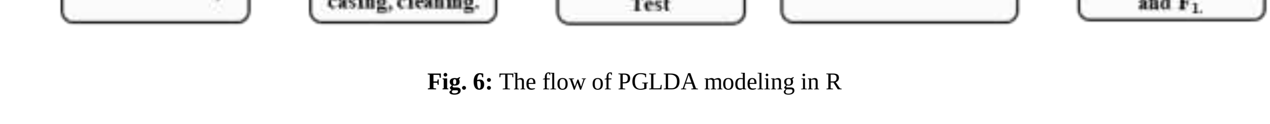

Data extraction: (20Newsgroups or AG's News)
Preprocessing: Stop words, stemming, lower casing, cleaning.
Data partition: Train, Validation and Test
Model building: PG + LDA = PGLDA
Performance Evaluation: Recall, Precision and F1.

Fig. 6: The flow of PGLDA modeling in R

Results

Table 2 presents the parameter estimation results for the datasets used. The Poisson-Gamma model was first fitted on the document length (number of words per document). The main parameter of concern that determines the validity of PGLDA over LDA is b. It has been earlier shown that large b implies there is no difference in the information that is retrieved by either PGLDA or LDA and vice-versa. The results in Table 2 show that the estimated values of b for the datasets are less than one and it implies that the Poisson distribution is not accurate for modeling document length for the two datasets. As a confirmation, if the Poisson distribution is accurate for modeling document length, it is expected that the mean words per document (ξ) and variance of words per document should be equal. Meanwhile, the result of $\overline{\xi}$ in Table 1 shows that the variances of words per document for the various datasets are largely greater than their means, about 5.5 (22.613 Vs 4.045); 7920.7 (8641.7 Vs 10.910) times higher for 20 Newsgroups and AG's News respectively. This shows that the assumption of equidispersion in Poisson and carried over to LDA, leading to the independence of words assumption, is largely violated especially in AG's News dataset.

To corroborate the findings in Table 2, the empirical density plots for the datasets shown in Fig. 7 reveal a close resemblance with the simulated plot when $b \leq 1$. This implies that the most appropriate model for the document distribution is the Poisson-Gamma distribution which hence confirms its validity.

Interpretability: Topic Coherence

Table 3 shows the average coherence score over different optimal number of topics for the top 10 words in each topic. It can be noted that PGLDA achieved a higher average score than the baseline models- DocNADE, ctx-DocNADE and ctx-DocNADEe. This shows that the PGLDA capability of modeling words arising from different structures as well as hub words helped in generating more coherent topics.

Table 4 shows the top 10 words and the respective coherence score for the AGNews dataset using PGLDA. The higher the score, the more meaningful the topic is predicted by the model. For class "World" for example, all the top 10 words are general and represent diverse themes such as world topic and this corresponds to the high coherence score of 0.955. Similar behavior was observed in Table 5 for the 20Newsgroups dataset.

Having established the validity of Poisson-Gamma, the next is to use it in the model building step of PGLDA for information retrieval, as shown in Fig. 8 and 9. Figure 9 showed that the prediction using the PGLDA method is excellent for most of the 20 newsgroups except talk.religion, soc.religion.christian, sci.med and rec.motorcycles. This implies the 20 newsgroups; the correct prediction was achieved in 16 newsgroups while poor performance was observed in 4 newsgroups. The results revealed notable improvement over the performance of LDA reported in (Xue, 2019) where it was observed that of the 20 newsgroups only eight newsgroups achieved an F_1 of at least 80%. Similarly, for AG's News dataset, the class-specific performance of PGLDA in terms of F_1 scores presented in Fig. 7 is high for the entire four classes which indicate, PGLDA is suitable for the different newsgroups in AG's News dataset.

Performance comparisons using (*Recall, Precision, Accuracy and F_1 score*) with recent BoW models such as DocNADE ctx-DocNADE and ctx-DocNADE (Gupta *et al*., 2019) are presented in Table 6. The results reveal that the proposed PGLDA model is better than all the competing methods in terms of applicability in information retrieval via the F_1 score. The baseline comparison with LDA shows a significant improvement over LDA for the two datasets. Precisely, there is a gain of about 24.3% (.906 Vs .729) and 21.6% (.995 Vs .818) for 20 Newsgroups and AG's News datasets respectively.

Table 2: Parameter estimate of the PGLDA model for the datasets

	Parameter		
Dataset	$\hat{b}$ (SE)	$\hat{a}$ (SE)	$\hat{\xi}$ (SE)
20 Newsgroups	0.881 (0.00321)	0.218 (0.00101)	4.045 (4.75540)
AG's News	0.385 (0.00138)	0.021 (0.00013)	10.910 (293.96380)

SE: Standard Error; $\hat{\xi} = \hat{b} / \hat{a}$.

Table 3: Coherence score for various models across datasets

	Dataset	
Model	20News	AGNews
DocNADE (Gupta *et al*., 2019)	0.606	0.731
ctx-DocNADE (Gupta *et al*., 2019)	0.615	0.739
ctx-DocNADEe (Gupta *et al*., 2019)	0.631	0.746
PGLDA	0.671	0.757

Note: ***The results of DocNADE, ctx-DocNADE and ctx-DocNADEe were culled from (Gupta *et al*., 2019)

Table 4: Top 10 words for all topics in AGNews and their respective coherence score using PGLDA

Class	Top 10 words	Coherence score
World	reuters, president, iraq, minister, ap, people, government, killed, prime, quot	0.955
Business	reuters, oil, company, gt, lt, york, stocks, prices, percent, corp	0.951
Sci/Tech	gt, lt, software, microsoft, internet, company, quot, ap, computer, reuters	0.895
Sports	game, ap, night, season, team, world, win, victory, league, sunday	0.225
Average		0.757

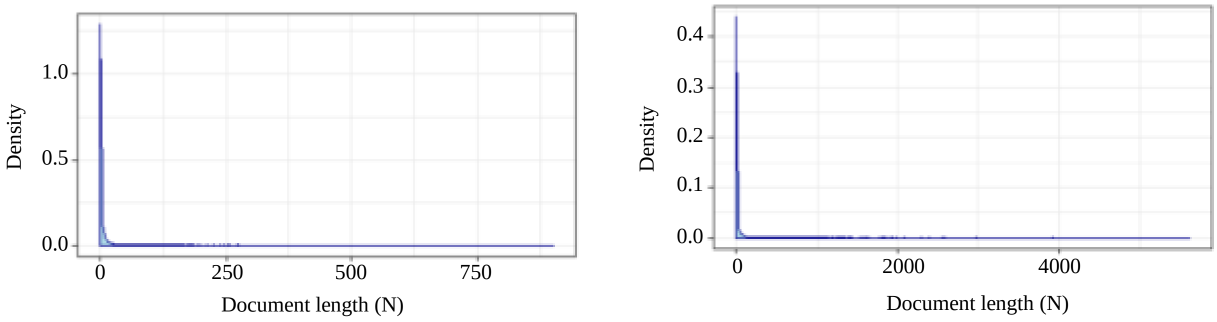
Table 5: Top 10 words for all topics in 20 News and their respective coherence score using PGLDA

Newsgroups	Top 10 words	Coherence score
soc.religion.christian	god, Jesus, people, time, church, faith, Christians, Christ, bible, life	0.545
alt.atheism	people, god, Jesus, time, im, evidence, Islam, religion, bible, argument	0.617
sci.space	space, nasa, orbit, data, time, launch, earth, lunar, shuttle, moon	0.618
comp.sys.ibm.pc.hardware	god, lord, people, church, im, Jesus, found, love, Christian, accept	0.600
comp.sys.mac.hardware	scsi, drive, card, system, mac, bit, im, mb, disk, apple	0.584
comp.os.ms-windows.misc	windows, file, dos, files, card, program, version, driver, drivers, run	0.604
rec.motorcycles	bike, im, dod, time, ive, ride, people, bikes, riding, helmet	0.737
comp.graphics	graphics, image, program, im, email, software, bit, files, computer, file	0.743
talk.religion.misc	god, Jesus, people, bible, Christian, life, time, law, ra, word	0.811
rec.autos	car, cars, engine, time, im, speed, drive, oil, people, dealer	0.793
sci.electronics	ground, wire, power, circuit, wiring, current, im, time, voltage, output	0.769
rec.sport.hockey	game, team, hockey, season, play, games, players, nhl, teams, time	0.629
talk.politics.mideast	people, Armenian, Turkish, Armenians, Israel, Jews, Israeli, Turks, Armenia, turkey	0.635
comp.windows.x	file, window, program, entry, server, motif, output, code, email, set	0.564
talk.politics.guns	people, gun, guns, law, government, weapons, firearms, time, fire, weapon	0.793
sci.med	people, time, im, msg, food, pain, patients, doctor, water, day	0.760
misc.forsale	sale, email, offer, dos, shipping, price, condition, drive, excellent, sell	0.809
talk.politics.misc	people, president, government, Stephanopoulos, time, im, jobs, tax, money, American	0.563
sci.crypt	key, db, encryption, chip, government, clipper, people, keys, system, security	0.444
rec.sport.baseball	game, team, games, players, runs, hit, baseball, time, league, season	0.791
Average		0.671

Table 6: Recall, precision, accuracy and F_1 score for various models across datasets

		Models				
Dataset	Performance metrics	LDA (Gupta *et al*., 2019)	DocNADE (Gupta *et al*., 2019)	ctx-DocNADE (Gupta *et al*., 2019)	ctx-DocNADEe (Gupta *et al*., 2019)	PGLDA
20 Newsgroups	Recall	0.732	0.730	0.738	0.750	0.911
	Precision	0.726	0.724	0.726	0.740	0.901
	Accuracy	0.872	0.872	0.874	0.878	0.952
	F_1	0.729	0.727	0.732	0.745	0.906
AG's News	Recall	0.823	0.897	0.895	0.903	1.000
	Precision	0.812	0.879	0.885	0.885	0.984
	Accuracy	0.912	0.944	0.945	0.946	0.997
	F_1	0.818	0.888	0.890	0.894	0.995

Note: ***The results of DocNADE, ctx-DocNADE and ctx-DocNADEe were culled from (Gupta *et al*., 2019)



(a)

(b)

Fig. 7: Empirical density plots against document length (N) for PGLDA; (a): Density plot: 20 Newsgroups. (b): Density plot: AG's News

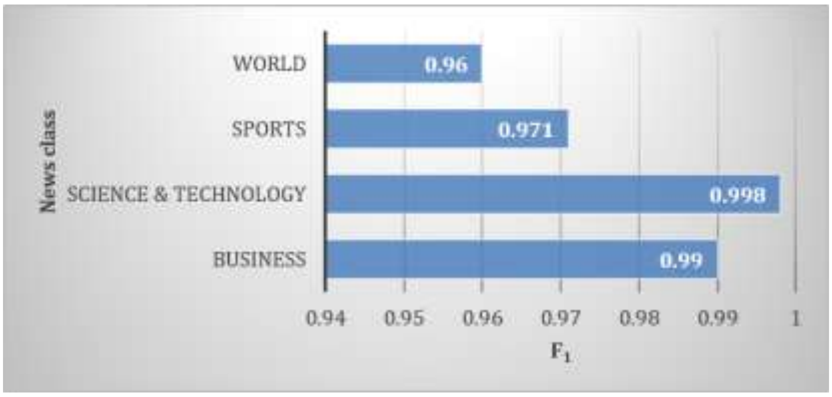


Fig. 8: *F*1 score for each newsgroup class in AG's News dataset using PGLDA

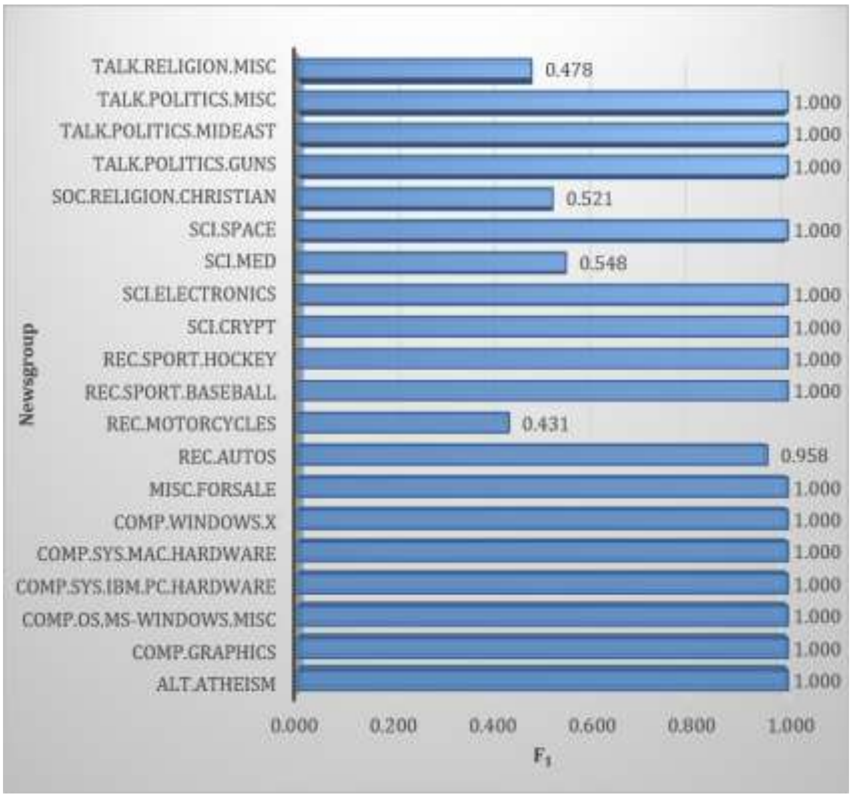


Fig. 9: *F*1 score for each newsgroup class in 20 Newsgroups dataset using PGLDA

Similar results as observed with the F_1 score were observed using Recall, Precision and Accuracy.

Conclusion

In this study, an extended LDA model tagged PGLDA was developed for modeling word dependencies in text data. Specifically, the Poisson document length was extended to capture the word/topic correlation features inherent in most text data. The algorithmic approach of PGLDA follows from LDA with modification of document length distribution with the Poisson-Gamma mixture. The performance comparison of the model with LDA showed that the PGLDA model fits better than the standard LDA. However, other flaws in LDA such as not capturing semantic correlation and word order are still carried over to PGLDA. Thus, in the future, we intend to combine PGLDA with the word embeddings model to solve these issues.

Funding Information

This research is self-sponsored by its corresponding authors.

Author's Contributions

Mohd Zainuri Saringat: Reviewed and prepared the final paper manuscript. Contributed to analyzing all experiments.

Ibrahim Bakari Bala: Prepared the first draft of the manuscript, designed and developed the model, performed the experiments.

Ethics

The authors confirm that this work is an original research paper and there are no ethical issues behind the publication of this manuscript.

References

Albishre, K., Albathan, M., & Li, Y. (2015, December). Effective 20 newsgroups dataset cleaning. In 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (Vol. 3, pp. 98-101). IEEE.

Ali, K., Dong, H., Bouguettaya, A., Erradi, A., & Hadjidj, R. (2017, June). Sentiment analysis as a service: a social media based sentiment analysis framework. In 2017 IEEE International Conference on Web Services (ICWS) (pp. 660-667). IEEE.

Bala, I. B., & Saringat, M. Z. (2019). GPLDA: A Generalized Poisson Latent Dirichlet Topic Model.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Chen, L. C. (2017). An effective LDA-based time topic model to improve blog search performance. Information Processing & Management, 53(6), 1299-1319.

Del Corso, G. M., Gulli, A., & Romani, F. (2005, May). Ranking a stream of news. In Proceedings of the 14th international conference on World Wide Web (pp. 97-106).

Gruber, M., Gruber, S. B., Taube, W., Schubert, M., Beck, S. C., & Gollhofer, A. (2007). Differential effects of ballistic versus sensorimotor training on rate of force development and neural activation in humans. Journal of strength and conditioning research, 21(1), 274-282.

Gupta, P., Chaudhary, Y., Buettner, F., & Schütze, H. (2019, July). Document informed neural autoregressive topic models with distributional prior. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 6505-6512).

Hong, W., Hao, Z., & Shi, J. (2017). Research and Application on Domain Ontology Learning Method Based on LDA. JSW, 12(4), 265-273.

Hu, Q. V., He, L., Li, M., Huang, J. X., & Haacke, E. M. (2014, November). A semi-informative aware approach using topic model for medical search. In 2014 IEEE international conference on bioinformatics and biomedicine (BIBM) (pp. 320-324). IEEE.

Inouye, D., Ravikumar, P., & Dhillon, I. (2014a, January). Admixture of Poisson MRFs: A topic model with word dependencies. In International Conference on Machine Learning (pp. 683-691).

Inouye, D. I., Ravikumar, P. K., & Dhillon, I. S. (2014b). Capturing semantically meaningful word dependencies with an admixture of Poisson MRFs. In Advances in Neural Information Processing Systems (pp. 3158-3166).

Jamil, S. A. M., Abdullah, M. A. A., Kek, S. L., Olaniran, O. R., & Amran, S. E. (2017). Simulation of parametric model towards the fixed covariate of right censored lung cancer data. In Journal of Physics: Conference Series (Vol. 890, No. 1, p. 012172).

Jiang, B., Li, Z., Chen, H., & Cohn, A. G. (2018). Latent topic text representation learning on statistical manifolds. IEEE transactions on neural networks and learning systems, 29(11), 5643-5654.

Jiang, S., Qian, X., Shen, J., Fu, Y., & Mei, T. (2015). Author topic model-based collaborative filtering for personalized POI recommendations. IEEE transactions on multimedia, 17(6), 907-918.

Ko, N., Jeong, B., Choi, S., & Yoon, J. (2017). Identifying product opportunities using social media mining: application of topic modeling and chance discovery theory. IEEE Access, 6, 1680-1693.

Larochelle, H., & Lauly, S. (2012). A neural autoregressive topic model. In Advances in Neural Information Processing Systems (pp. 2708-2716).

Larochelle, H., & Murray, I. (2011, June). The neural autoregressive distribution estimator. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (pp. 29-37).

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. SpringerPlus, 5(1), 1608.

Onan, A., Korukoglu, S., & Bulut, H. (2016). LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis. Int. J. Comput. Linguistics Appl., 7(1), 101-119.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Reisinger, J., & Mooney, R. (2010, June). Multi-prototype vector-space models of word meaning. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 109-117).

Ren, G., & Hong, T. (2017). Investigating online destination images using a topic-based sentiment analysis approach. Sustainability, 9(10), 1765.

Ren, Y., Wang, R., & Ji, D. (2016). A topic-enhanced word embedding for Twitter sentiment classification. Information Sciences, 369, 188-198.

Santosh, D. T., Babu, K. S., Prasad, S. D. V., & Vivekananda, A. (2016). Opinion mining of online product reviews from traditional LDA Topic Clusters using Feature Ontology Tree and Sentiwordnet. International Journal of Education and Management Engineering, 6(6), 34-44.

Wallach, H. M. (2006, June). Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning (pp. 977-984).

Wang, X., Zhu, P., Liu, T., & Xu, K. (2016). BioTopic: a topic-driven biological literature mining system. International Journal of Data Mining and Bioinformatics, 14(4), 373-386.

Wang, Z., & Zhang, Y. (2016, June). A Text Information Retrieval Method by Integrating Global and Local Textual Information. In 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 504-505). IEEE.

Xue, M. (2019, January). A text retrieval algorithm based on the hybrid LDA and Word2Vec model. In 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) (pp. 373-376). IEEE.

Zhang, Y., Ma, J., & Wang, Z. (2019). Semi supervised classification of scientific and technical literature based on semi supervised hierarchical description of improved latent dirichlet allocation (LDA). Cluster Computing, 22(3), 6881-6889.

Zhao, J., Feng, Q., Wu, P., Warner, J. L., Denny, J. C., & Wei, W. Q. (2019). Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein (a)(LPA). PloS one, 14(2), e0212112.