Journal of Theoretical and Applied Information Technology
15th May 2020. Vol.98. No 09
© 2005 – ongoing JATIT & LLS
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A HYBRID WORD EMBEDDING MODEL BASED ON ADMIXTURE OF POISSON-GAMMA LATENT DIRICHLET ALLOCATION MODEL AND DISTRIBUTED WORD-DOCUMENT-TOPIC REPRESENTATION

[1,2]IBRAHIM BAKARI BALA, [2]MOHD ZAINURI SARINGAT, [2]AIDA MUSTAPHA

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

E-mail: balagombi@yahoo.co.uk, {zainuri, aidam}@uthm.edu.my

ABSTRACT

This paper proposes a hybrid Poisson-Gamma Latent Dirichlet Allocation (PGLDA) model designed for modelling word dependencies to accommodate the semantic representation of words. The new model simultaneously overcomes the shortcomings of complexity by using LDA as the baseline model as well as adequately capturing the words contextual correlation. The Poisson document length distribution was replaced with the admixture of Poisson-Gamma for words correlation modelling when there is a hub word that connects words and topics. Furthermore, the distributed representation of documents (Doc2Vec) and topics (Topic2Vec) vectors are then averaged to form new vectors of words representation to be combined with topics with largest likelihood from PGLDA. Model estimation was achieved by combining the Laplacian approximation of log-likelihood for PGLDA and Feed-Forward Neural Network (FFN) approaches of Doc2Vec and Topic2Vec. The proposed hybrid method was evaluated for precision, recall, and F_1 score based on 20 Newsgroups and AG's News datasets. Comparative analysis of F_1 score showed that the proposed hybrid model outperformed other methods.

Keywords: *Poisson-Gamma Distribution, Topic Model, LDA, Word2vec, Doc2vec, Topic2Vec*

1. INTRODUCTION

The distributed representation of words also referred to as word embedding is an alternative way of modelling words or topics with a specific interest in the semantic correlation between adjacent words. One of the fundamental issues with the Bags-of-Words (BoW) model is that they do not accommodate for semantic correlation that may occur between words. Thus, this implies incorporation of words semantic in modelling word or topic is expected to improve the performance of the word embedding model.

In word embeddings, words are often represented as fixed vectors, hence the term embeddings. There are several different models useful for the construction of embeddings, but they are all based on the distributional hypothesis widely known as "a word is characterized by the company it keeps". The goal of a word embeddings technique is to accommodate inherent semantic and syntactic features in language even when dealing with large sets of documents. Words that occur in the same context should be represented by vectors close to each other.

Most of the recent Natural Language Processing (NLP) techniques categorize words as a single entity with the belief that there is no interaction between words that are later defined as indices in a vocabulary. This type of technique has many advantages including simplicity, robustness and specifically the fact that simple approach applied to complex data is better than complex methods on simple data. A typical example is the popular statistical language model called *N*-gram which has been used on several datasets of varying dimensions [1]. However, there is a limitation to the capability of simple techniques. This exemplifies the situation where simple techniques will break down and hence the quest for more advanced techniques.

Mimno and Blei [2], Mimno and Mccallum [3] among others provide the basis for modelling words dependences in text mining. The authors claimed a more realistic model can only be achieved if word dependencies are taking into cognizance. Their idea

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

motivated the paper by Inouye et al. [4, 5] where the Poisson Markov Random Fields (PMRFs) was used to model the chain of words in topic modelling. Inouye et al. [4] introduced the idea behind PMRF and showed that it is collapsible to the LDA with Poisson document length under mild regularity conditions. The model presented a new dimension to dependent word modelling, but its complex structure makes it difficult to estimate.

Consequently, [5] presented a fast algorithm for estimating PMRF but the complex nature issue is inherent. The idea is complex because their approach treats each rate parameter as a multivariate distribution and thus introducing an estimation problem on top of solving the modelling issue.

As clearly spelt out in the previous paragraph, most of the existing topic models do not have the capability of modelling simultaneously word dependencies and words contextual correlation such as the prediction of next word in the sentence *"the temperature of Johor is 36°C"*. Word dependency refers to the use of the hub word which in the example is *"temperature"* as it has the capability of being a word as well as topic in the example. Similarly, in terms of contextual correlation, the word *temperature"* is synonymous to place or thing and thus we will expect a place such as the word *"Johor"* to be the next word within the sentence or paragraph. The relevance of the current work is in this parlance as the two issues often occur in topic modeling.

To address this gap, this paper proposes a hybrid method for modelling word dependencies with Poisson-Gamma Latent Dirichlet Allocation (PGLDA) model. PGLDA extends LDA by simultaneously overcoming the complexity of LDA in capturing contextual words correlation. The LDA-based models lack the capability to incorporate words context and thus the essence of the proposition of word embeddings by [1, 5, 6]. The hybrid word embeddings approach is hoped to accommodate words context by using semantic representation of words.

The remainder of this paper proceeds as follows. Section 2 presents the preliminary models used in this study. Section 3 presents the proposed the hybrid word embeddings method. Section 4 presents the experimental setup, Section 5 discusses the results and finally Section 6 concludes the paper with some indication of future work.

2. PRELIMINARIES

To assess the performance of the proposed hybrid Poisson-Gamma Latent Dirichlet Allocation (PGLDA) method, the results will be compared against various methods such as Latent Dirichlet Allocation (LDA), Distributed Representation of Word Vectors, Global Vectors for Word Representation (Glove2Vec), Distributed Representation of Documents (Doc2Vec), Topic2Vec, and LDA2Vec [8].

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) developed by Blei [22] requires the use of two conditional probability distributions, $P(z|d)$ and $P(w|z)$, which assumes the multinomial distributions whereby topics in the whole documents have the same Dirichlet prior distribution $P(\alpha)$ likewise the conditional distributions of words on topics possess the same Dirichlet prior $P(\beta)$ [8].

After choosing the most suitable prior hyper-parameters α and β in document d, a parameter θ is created from the conditional distribution of K topics which is supposed to be multinomially distributed from the Dirichlet distribution $Dir(\theta|\alpha)$. Additionally, for a specific topic k, a conditional distribution of V words are created, which is assumed to be multinomially distributed from the Dirichlet distribution $Mult(w|z,\beta)$. The Dirichlet prior distribution is selected as a result of the unification property existing between the multinomial and Dirichlet distribution, this fact simplifies the LDA statistical inference as depicted in Figure 1.

Algorithm 1: Latent Dirichlet Allocation (LDA)

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

Figure 1: LDA Algorithm.

In recent times, LDA have been globally accepted in the domain of Information Retrieval and Sentiment Analysis [9-10]. Santosh et al. [9] presented a new performance improvement approach for LDA. They first used an ontology approach to identify the most suitable features after clustering and this method depicts that, the accuracy of feature extraction largely improved. Meanwhile, Ren and Hong [11] used extracted

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

topics from online travel review to perform Topic-based SA to determine the most important to the tourist from topics and emotions. Tweets were first generated using LDA before the incorporation of topic function. The system performance recorded about 4% improvement when the topics were integrated into word embedding.

The performance of the LDA-based feature selection approach was also investigated by [12] in the area of text classification. Sentiment classification was done via optimal latent topic obtained from the combination of machine learning-based classifiers and LDA in order to obtain the optimal number of latent topics. Hong et al. [13] presented an LDA-based learning system for updating civil aviation domain system. The representation content was enriched by the system making the information to provide better support for management of the emergency system. A similar study by [14] used the LDA-based procedure to product opportunities. In their work, changes of customers' needs were monitored by identifying the product opportunity preference.

However, the topics that were generated by LDA-based techniques returned topics with irrelevant words. In addition, as observed from other similar studies LDA-based approach fails to capture the semantic correlation between adjacent words. Therefore, [10] suggested the use of a preliminary feature representation method with LDA for identification of a topic. In [15], Ali et al. presented an ontology-based, feature-level sentiment analysis for describing the relationships between concepts in a specific domain. The previous works are based on traditional LDA approaches which do not incorporate word dependencies. We propose a Poisson-Gamma LDA-based topic modelling method for documents classification.

2.2 Distributed Representation of Word Vectors

Word2Vec is one of the recent extensions of distributed representation of word vectors proposed originally by [1]. The approach follows from the extension of the works by [16-17]. Word2Vec approach focused on the preliminary step of the methods where the word vectors are learned using a simple model. The model estimation was done using different model structure trained on various corpora. The resulting word vectors are then in turn used for future prediction comparison.

Mikolov et al. [1] discussed the computational complexity of the word2vec model especially those not involving certain version of log-bilinear model where diagonal weight matrices are used. Mikolov et al. [1] also reported that LDA and Latent Semantic Analysis (LSA) do not only performed significantly poor on large dataset but are also computationally expensive. Thus, this places word2vec above LSA and LDA for large datasets in terms of predictive performance.

There are two major approaches for modelling Word2Vec; Continuous Bag-of-Words approach (CBOW) and Continuous Skip-gram approach. The two approaches try to minimize the overall model computational complexity [18]. The two estimation techniques NNLM and RNNLM showed that the complexity is caused by the non-linearity at the hidden layer. Although, the strength of neural network approaches is in its ability to capture non-linearity, to minimize computation cost, most word2vec fitting techniques adopt simpler models with the ability to train the data effectively.

Mathematically, given a sequence of training words $w_1, w_2, w_3, \dots, w_T$, the objective of the word vector model is to maximize the average log probability is given as follows.

$$\frac{1}{M}\sum_{t=k}^{M-k} \log p(w_t|w_{t-k}, \dots, w_{t+k})$$

The final prediction is carried out using a softmax classifier as the following equation,

$$p(w_t|w_{t-k}, \dots, w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_{w_i}}}$$

where y_i is un-normalized log-probability for each ith word and it is computed by:

$$y = b + Uh(w_{t-k}, \dots, w_{t+k}; W)$$

and b is the model bias parameter, U is the hidden layer parameters for the words and h is a concatenation function which is the average of all word vectors from W. The above procedure is the neural network methodology of Word2Vec which is the bedrock of Doc2Vec. Figure 2 shows the framework of Word2Vec in learning vectors of words ("the," "cat," and "sat") which is used to predict the fourth word ("on"). The input words are mapped to columns of the matrix W to predict the output word [6].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

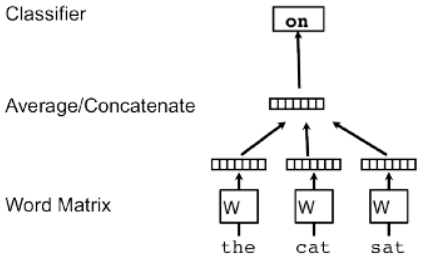

Figure 2: Word2Vec Framework.

2.3 Global Vectors for Word Representation

An alternative distributive word representation focused on reducing the computational complexity of Word2Vec is Glove2Vec [18]. The model introduced as Glove was developed by [19]. It is an unsupervised learning algorithm that gives vector representations of words. The algorithm combined the advantages of the two major models used in learning word vectors: global matrix factorization technique such as Latent Semantic Analysis (LSA) and local context window technique such as the Skip-gram model. Glove2Vec make use of co-occurrence probabilities ratios, instead of the co-occurrence probabilities themselves. This follows from the fact that the ratios between the probabilities of surrounding words carry more information than individual probabilities.

2.4 Distributed Representation of Documents

The Doc2Vec [20] is a method based on the earlier formulation of learning from paragraph vectors through word vectors. The word vectors are believed to contribute a significant amount to the prediction of next word in the paragraph. The word vectors random initialization does not in any way influences prediction task since they can capture the semantics via indirect mechanism. Le and Mikolov [20] used the idea to develop the Doc2Vec methodology. The process ends by making the paragraph vectors contribute to the prediction task of the next word given many contexts sampled from each paragraph. For Doc2Vec, the additional feature *D* captures the document properties as introduced into the objective function so that it becomes:

$$\frac{1}{M}\sum_{t=k}^{M-k}\log p(w_t|w_{t-k},\dots,w_{t+k};D_{t-k},\dots,D_{t+k})$$

The final prediction is carried out using a softmax classifier as shown as follows,

$$p(w_t|w_{t-k},\dots,w_{t+k};D_{t-k},\dots,D_{t+k})=\frac{e^{y_{w_t}}}{\sum_i e^{y_{w_i}}}$$

where y_i is un-normalized log-probability for each ith word and it is computed as:

$$y=b+Uh(w_{t-k},\dots,w_{t+k};D_{t-k},\dots,D_{t+k};W;D)$$

In the Doc2Vec algorithm, each paragraph is mapped into a unique vector using a column in matrix *D* and likewise each word is mapped to a unique vector using column a column in matrix W. The paragraph vector and word vector are then averaged using same concatenation function *h*. Figure 2 shows the framework of Doc2Vec. Based on this figure, in Doc2Vec, each document is connected to a unique vector represented by a column in matrix *D* and as well each word is also connected to a unique vector, represented by a column in matrix *W*.

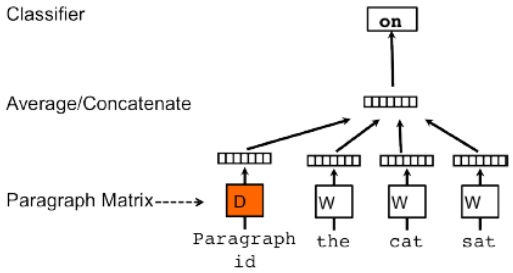

Figure 3: Doc2Vec Framework.

The document vector and word vectors are averaged or concatenated to predict the next word in a context. Le and Mikolov [20] used concatenation as the method to combine the vectors. The model is only different from the Word2Vec model in terms of the way the hidden layer is constructed. In Doc2Vec, the hidden layer is constructed from the word W and document *D*. Furthermore, in Doc2Vec, the paragraph is tokenized as another word such that it acts as a memory for a missing current context or the overall topic of a document.

2.5 Topic2Vec

Topic2Vec model learns topic representations that are in the same semantic vector space with words [15]. Word2Vec is divided into two models: CBOW, which predicts both a word and topic vector using the context words and Skip-gram that predicts the context given the current word and its assigned topic by LDA. Figure 4 shows the architecture of Topic2Vec where (*wt*−2, *wt*−1,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

wt+1, *wt*+2) are context words and *wt* is the current word paired with a topic *zt* [21].

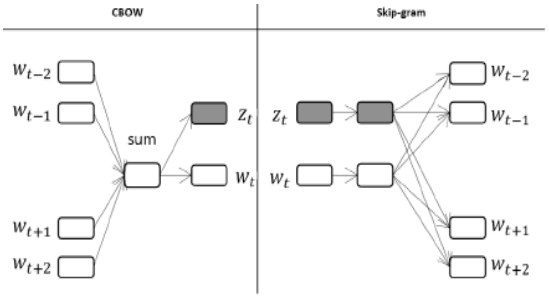

Figure 4: Topic2Vec Framework.

2.6 LDA2Vec

LDA2Vec by Moody [21] is a combination of two models: the LDA and word embedding (Word2Vec). LDA presented by Blei [22] generates document (and topic) representations that can easily be decoded by human but lacks flexibility while Word2Vec [1] is a model that generates vector representations of words by correlating the semantic relatedness between words but the vectors generated cannot be interpreted by humans. The limitations of both model lead to the development of LDA2Vec which solves the setbacks by embedding words and document vectors together into the same space and concurrently trains both representations. Based on Mikolov [6], if word vectors are added together the combination of both words can form a semantically meaningful combination.

3. HYBRID PGLDA AND DISTRIBUTED REPRESENTATION OF WORD-DOCUMENT-TOPIC

One of the drawbacks of word2vec is its inability to capture word order and document properties such as the one captured using PGLDA. [20] proposed the doc2vec also referred to as paragraph to vector. The approach was inspired by the fact that the word vectors contribute to a prediction task about the next word in the sentence. The Doc2Vec approach is an update over the Word2Vec by introducing the paragraph or document features in the word embedding. The general approach of distributed vector of words involves the prediction of a word based on other words in the context. In the framework every word is mapped to a unique vector, represented by a column in a matrix W. The column is indexed by the position of the word in the vocabulary. The concatenation or sum of the vectors is then used as features for prediction of the next word in a sentence.

Suppose there are *N* documents defined as in the case of LDA with the assumed probability of n document at a specific time interval distributed as Poisson, the probability of *N* assuming *n* as:

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

Under regularity assumption, 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. For 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 follows.

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

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

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

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

The unconditional distribution of *N* assuming *n* given ξ is as follows.

$$P(N = n) = \int_0^\infty \frac{a^b \exp[-(a+1)\xi]\xi^{n+b-1}}{n!\,\Gamma(b)} d\xi$$

Therefore, the following shows the admixture of Poisson-Gamma used in PGLDA.

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

PGLDA relaxes assumption of independence used by LDA and carry over the other 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 × 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.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PGLDA is more flexible and realistic when compared to LDA. Mathematically, the joint distribution of topic z, word w and topic mixture θ is defined as:

$$P(\theta, z, w|\alpha, \beta, b, a) = P(\theta|\alpha)\prod_{n=1}^{N} P(z_n|\theta)\, P(w_n|z_n, \beta)P(N = n|a, b)$$

$$P(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\dots\theta_k^{\alpha_k-1}$$

$$P(z_n|\theta) = \prod_{n=1}^{N}\frac{\Gamma\left(\sum_{i=1}^{k} z_{ni}+1\right)}{\prod_{i=1}^{k}\Gamma(z_{ni}+1)}\prod_{i=1}^{k}\theta_i^{z_{ni}}$$

$$P(w_n|z_n, \beta) = \prod_{n}^{N}\prod_{v=1}^{V}\frac{\Gamma\left(\sum_{i=1}^{k} w_{ni}+1\right)}{\prod_{i=1}^{k}\Gamma(w_{ni}+1)}\prod_{i=1}^{k}\beta_{iv}^{w_{ni}}$$

$$P(\theta, z, w|\alpha, \beta, b, a) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\dots\theta_k^{\alpha_k-1} \times \prod_{n=1}^{N}\left\{\left[\prod_{n=1}^{N}\frac{\Gamma\left(\sum_{i=1}^{k} z_{ni}+1\right)}{\prod_{i=1}^{k}\Gamma(z_{ni}+1)}\prod_{i=1}^{k}\theta_i^{z_{ni}}\right] \times \left[\prod_{n}^{N}\prod_{v=1}^{V}\frac{\Gamma\left(\sum_{i=1}^{k} w_{ni}+1\right)}{\prod_{i=1}^{k}\Gamma(w_{ni}+1)}\prod_{i=1}^{k}\beta_{iv}^{w_{ni}}\right] \times \left[\frac{\Gamma(n+b)}{\Gamma(n+1)\Gamma(b)}\left(\frac{a}{a+1}\right)^{b}\left(\frac{1}{a+1}\right)^{n}\right]\right\}$$

The marginal distribution of document D can be obtained by marginalizing the joint distribution $P(\theta, z, w|\alpha, \beta, b, a)$ as follows.

$$P(D|\theta_d, z, w, \alpha, \beta, b, a) = \prod_{d=1}^{M}\int\left(\frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_{d1}^{\alpha_1-1}\dots\theta_{dk}^{\alpha_k-1} \times \prod_{n=1}^{N}\left\{\left[\prod_{n=1}^{N}\frac{\Gamma\left(\sum_{i=1}^{k} z_{ni}+1\right)}{\prod_{i=1}^{k}\Gamma(z_{ni}+1)}\prod_{i=1}^{k}\theta_{di}^{z_{ni}}\right] \times \left[\prod_{n}^{N}\prod_{v=1}^{V}\frac{\Gamma\left(\sum_{i=1}^{k} w_{ni}+1\right)}{\prod_{i=1}^{k}\Gamma(w_{ni}+1)}\prod_{i=1}^{k}\beta_{iv}^{w_{ni}}\right] \times \left[\frac{\Gamma(n+b)}{\Gamma(n+1)\Gamma(b)}\left(\frac{a}{a+1}\right)^{b}\left(\frac{1}{a+1}\right)^{n}\right]\right\}\right)d\theta_d$$

In the hybrid PGLDA, Doc2Vec and Topic2Vec, the modified distribution of word length from PGLDA are supplied into the objective function to improve the accuracy in the prediction of words. That is; formally, the objective function can be defined in terms of CBOW and Skip-gram approach.

$$f_{CBOW}(S) = \frac{1}{M}\sum_{i=1}^{M}\{\log[p(w_i|w_{cxt})] + \log[p(z_i|w_{cxt})] + \log[p(D_t|D_{t-k}, \dots, D_{t+k})]\}$$

$$f_{skip-gram}(S) = \frac{1}{M}\sum_{i=1}^{M}\{\log[p(w_{i+c}|w_i)] + \log[p(w_{i+c}|z_i)] + \log[p(D_t|D_{t-k}, \dots, D_{t+k})]\}$$

where $p(w_t|w_{t-k}, \dots, w_{t+k}; D_{t-k}, \dots, D_{t+k})$ and

$$p(D_t|D_{t-k}, \dots, D_{t+k}) = P(D|\theta_d, z, w, \alpha, \beta, b, a) = \prod_{d=1}^{M}\int\left(\frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_{d1}^{\alpha_1-1}\dots\theta_{dk}^{\alpha_k-1} \times \prod_{n=1}^{N}\left\{\left[\prod_{n=1}^{N}\frac{\Gamma\left(\sum_{i=1}^{k} z_{ni}+1\right)}{\prod_{i=1}^{k}\Gamma(z_{ni}+1)}\prod_{i=1}^{k}\theta_{di}^{z_{ni}}\right] \times \left[\prod_{n}^{N}\prod_{v=1}^{V}\frac{\Gamma\left(\sum_{i=1}^{k} w_{ni}+1\right)}{\prod_{i=1}^{k}\Gamma(w_{ni}+1)}\prod_{i=1}^{k}\beta_{iv}^{w_{ni}}\right] \times \left[\frac{\Gamma(n+b)}{\Gamma(n+1)\Gamma(b)}\left(\frac{a}{a+1}\right)^{b}\left(\frac{1}{a+1}\right)^{n}\right]\right\}\right)d\theta_d$$

Figure 5 shows the algorithm for PGLDA.

Algorithm 2: Poisson-Gamma Latent Dirichlet Allocation (PGLDA)

1) Sample $\boldsymbol{\xi}$ from Gamma distribution $\mathbf{G(b, a)}$
2) Sample N from Poisson-Gamma Mixture $\mathbf{P(N = n|a, b)}$
3) For each topic $\mathbf{k} \in \{1, 2, 3, \dots, K\}$:
4) For each document $\mathbf{d} \in \{1, 2, 3, \dots, N\}$:
5) Simulate $\boldsymbol{\theta_d} \sim \mathbf{Dir(\theta_d|\alpha)}$
6) For each word w $\in \mathbf{d} \in \{1, 2, 3, \dots, N\}$:
7) Simulate $\mathbf{z_{dn}} \sim \mathbf{Mult(z_{dn}|\theta_d)}$
8) Simulate $\mathbf{w_{dn}} \sim \mathbf{Mult(w_{dn}|z_{dn}, \beta)}$

Figure 5: PGDLA Algorithm.

Next, Figure 6 shows Framework of hybrid PGLDA, Doc2Vec and Topic2Vec. It involves the addition of context from the PGLDA model, Doc2Vec and TopicVec features. The most probable context of word is chosen from the PGLDA topic model in addition to the document probability distribution which are in turned combined with word vector and document vector for the prediction of next words from PGLDA and Doc2Vec. The next words predicted are then summed to predict topics and words in the Topic2Vec model.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The procedure involves three stages. In the first stage (PGLDA phase) the latent topics and latent document distribution that are likely to be associated with the training document set $D = \{d_1, \dots, d_n\}$ are initialized. The trained documents then yield the context distribution of likely topics. The topics with maximum probability are then chosen for onward further processing in the doc2vec process. This ends the first stage of the PGLDA2Vec. In the second stage, the document vectors are created vis-à-vis word vector. The combined procedure of context, document vector and word vector are then used to predict the final output "on".

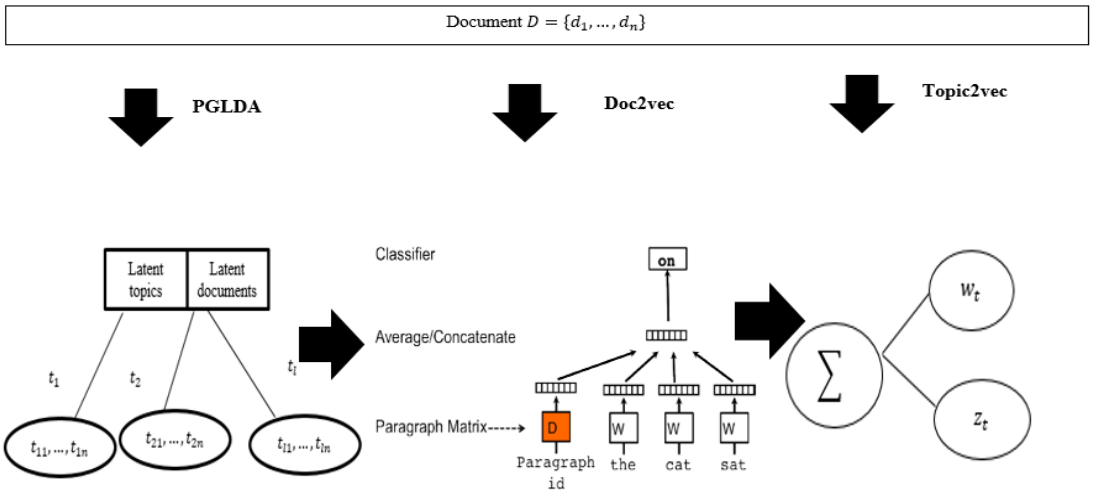

Figure 6: Hybrid Framework of PGLDA, Doc2Vec and Topic2Vec.

The final stage of the model developed in this paper is based on the position that it is better to sacrifice algorithmic time to accuracy of prediction. Simultaneous prediction of word and topic is expected to yield more accurate prediction than prediction of words alone especially when posed with documents with higher likelihood of word similarities and presence of hub or pivot words. The topic2vec model yields a reasonable accuracy in words and topics prediction, thus improving its accuracy with document features from PGLDA2vec is expected to yield a better performance. The complete pseudocode of the algorithm is presented in Algorithm 3 (Figure 7).

Algorithm 3: Hybrid PGLDA2Vec

1) Sample $\boldsymbol{\xi}$ from Gamma distribution $\mathbf{G(b, a)}$
2) Sample N from Poisson-Gamma Mixture $\mathbf{P(N = n|a, b)}$
3) For each topic $\mathbf{k} \in \{1, 2, 3, \dots, K\}$:
4) For each document $\mathbf{d} \in \{1, 2, 3, \dots, N\}$:
5) Simulate $\boldsymbol{\theta_d} \sim \mathbf{Dir(\theta_d|\alpha)}$
6) For each word w $\in \mathbf{d} \in \{1, 2, 3, \dots, N\}$:
7) Simulate $\mathbf{z_{dn}} \sim \mathbf{Mult(z_{dn}|\theta_d)}$
8) Simulate $\mathbf{w_{dn}} \sim \mathbf{Mult(w_{dn}|z_{dn}, \beta)}$
9) Construct the topic vector $\mathbf{v(z_{dn})}$ using topics from $\mathbf{Mult(z_{dn}|\theta_d)}$
10) Construct the document vector $\mathbf{v(d_i)}$ using word vector $\mathbf{v(w_{dn})}$.
11) Finally determine the topic-document distance $\mathbf{Dis[v(d_i), v(t_i)] = |v(d_i) - v(t_i)|}$
12) Predict the target topic and word using the topic-word-document distance.

Figure 7: PGDLA Algorithm.

4. EXPERIMENTAL SETUP

The proposed hybrid method PGLDA2Vec was tested on 20 Newsgroup and the AG's News dataset [24, 25] in a multi-class classification experiment. 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 the total 18,846 documents were used for training and the remaining 7,527 documents were used for testing. Table 1 detail out the split of training and training newsgroup datasets for the experiments.

On the other hand, the AG's News dataset was constructed based on 4 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 & Technology and Sports News.

Class-specific performance analysis was measured based on precision, recall, and F measure.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Precision (P) (also referred to as positive predictive value) is the proportion of relevant cases among the retrieved cases, while Recall (R) (also referred to as sensitivity) is the proportion of the total amount of relevant cases that were actually retrieved [26]. The F_1 is a measure of accuracy of the test dataset and is defined as follows.

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

The micro F_1 Score 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 micro F_1 Score is measured based on:

$$\text{Micro } F_1 = t^{-1} \sum_{i=1}^{t} F_{1t}.$$

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

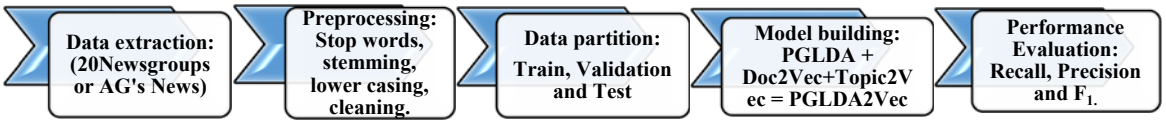

Figure 8: Flow of PGLDA2Vec modelling in R

Table 1: 20 Newsgroup Documents/Datasets.

Class	Training Documents	Testing Documents	Total Documents
alt.atheism	480	319	799
comp.graphics	584	389	973
comp.os.ms-windows.misc	591	394	985
comp.sys.ibm.pc.hardware	590	392	982
comp.sys.mac.hardware	578	385	963
comp.windows.x	593	395	988
misc.forsale	585	390	975
rec.autos	594	396	990
rec.motorcycles	598	398	996
rec.sport.baseball	597	397	994
rec.sport.hockey	600	399	999
sci.crypt	595	396	991
sci.electronics	591	393	984
sci.med	594	396	990
sci.space	593	394	987
soc.religion.christian	599	398	997
talk.politics.guns	465	310	775
talk.politics.mideast	546	364	910
talk.politics.misc	564	376	940
talk.religion.misc	377	251	628
Total	11,314	7,527	18,846

5. RESULTS AND DISCUSSION

Table 2 shows the class-specific performance analysis for PGLDA2Vec. The results from Table 2 showed that prediction using the new hybrid PGLDA, Doc2Vec and Topic2Vec method achieved excellent results for most newsgroups except *comp.sys.mac.hardware* and *misc.forsale*. This implies that from the total of 20 newsgroups; correct prediction was achieved in 18 newsgroups while poor performance was observed in 2 newsgroups. The results revealed notable improvement over the performance of LDA reported in [8] where it was observed that only 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

newsgroups achieved an F1 of at least 80% from the total of 20 newsgroups.

Table 2: Class Level Performance Analysis

Class	Prec. (%)	Recall (%)	F_1 (%)
alt.atheism	100.00	100.00	100.00
comp.graphics	100.00	100.00	100.00
comp.os.ms-windows.misc	100.00	100.00	100.00
comp.sys.ibm.pc. hardware	100.00	100.00	100.00
comp.sys.mac. hardware	45.54	50.00	47.66
comp.windows.x	100.00	100.00	100.00
misc.forsale	54.46	50.00	52.14
rec.autos	100.00	100.00	100.00
rec.motorcycles	100.00	100.00	100.00
rec.sport.baseball	100.00	100.00	100.00
rec.sport.hockey	100.00	100.00	100.00
sci.crypt	100.00	100.00	100.00
sci.electronics	100.00	100.00	100.00
sci.med	100.00	100.00	100.00
sci.space	100.00	100.00	100.00
soc.religion.christian	100.00	100.00	100.00
talk.politics.guns	100.00	100.00	100.00
talk.politics.mideast	100.00	100.00	100.00
talk.politics.misc	100.00	100.00	100.00
talk.religion.misc	100.00	100.00	100.00

Table 3 corroborate the findings in Table 1 with better average performance overall the 20 newsgroups for new hybrid PGLDA2Vec compared to TF-IDF, LDA, Word2Vec, LDA2Vec Glove2Vec, Doc2Vec, the most recent Document Neural Autoregressive Distribution Estimator (DocNADE) [27-28] and standard PGLDA.

The exciting results can be observed when compared to the most recent BoW models DoCNADE[28]. There is a gain of about 6.1% (0.962 vs 0.907) in using PGLDA2Vec for 20 Newsgroups document retrieval as against DocNADE.

Table 3: Micro F_1 Performance Comparison among Various Methods for 20 Newsgroup Dataset

Methods	Micro F_1
TF-IDF [8]	0.652
LDA [8]	0.729
Word2Vec [8]	0.803
LDA2Vec [8]	0.814
Glove2Vec [27]	0.627
Doc2Vec [27]	0.691
DocNADE [27]	0.907
PGLDA	**0.906**
Hybrid PGLDA2Vec	**0.962**

Similarly, class specific analysis was also computed for the AG's News dataset and results summarized in Table 4.

Table 4: Class Level Performance Analysis

Class	Prec. (%)	Recall (%)	F_1 (%)
Business	100.00	99.85	99.92
Science & technology	99.89	99.06	99.48
Sports	96.02	100.00	97.97
World	98.95	95.78	97.34

Table 4 shows the class-specific performance analysis for PGLDA2Vec. The results from Table 4 showed that prediction using the new hybrid PGLDA, Doc2Vec and Topic2Vec method achieved excellent results for all categories. This implies that from the total 4 categories; correct prediction was achieved in all news classifications while poor performance was observed in none. The results revealed significant notable improvement over the performance of LDA in [27] where it was found out that the average F_1 score is 0.818.

Table 5: Micro F_1 Performance Comparison among Various Methods for AGNews Dataset

Methods	Micro F_1
LDA [27]	0.818
Doc2Vec [27]	0.713
Glove2vec [27]	0.890
DocNADE [27]	0.888
PGLDA	**0.995**
Hybrid PGLDA2Vec	**0.997**

Table 5 corroborate the findings in Table 4 with better average performance overall the 4 categories of news for the new hybrid PGLDA2Vec compared to LDA, Glove2Vec, Doc2Vec, DocNADE and standard PGLDA.

Also, it can be observed that when compared to the most recent BoW models DoCNADE[28]. There is a gain of about 12.3% (0.997 vs 0.888) in using PGLDA2Vec for AG's News document retrieval as against DocNADE.

6. CONCLUSIONS

In this paper, the modified LDA model PGLDA was originally developed for modelling word dependencies and was extended to accommodate words contextual correlation by integrating probabilistic topic model (LDA) and word embeddings models (Doc2Vec and Topic2Vec). In particular, the document Poisson length distribution was replaced with an admixture of Poisson-Gamma to build new model called PGLDA. The most

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

probable topics from PGLDA is then combined with the average words and documents vectors from Doc2Vec and Topic2Vec to build new vectors useful for words weights. The new model is then calibrated on the 20 Newsgroup and AG's News datasets for model testing. The class-specific performance yielded high performance for the model in terms of high precision, recall and F1 score. In about 90% of the class/topics, 100% precision, as well as recall rates, were recorded in the two datasets which then confirms the adequacy of the method for information retrieval. Relative comparison with seven other methods for the 20 Newsgroup dataset shows that the new hybrid model largely outperformed the existing procedure. Similar, results were observed for the AG's News dataset when compared with four other methods.

REFERENCES

[1] T. Mikolov, K. Chen, G. Corrado, J. Dean. "Efficient Estimation of Word Representations in Vector Space", 2013, arXiv: 1301.3781.

[2] D. Mimno, D. Blei, "Bayesian Checking for Topic Models", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 227-237.

[3] Mimno, A. McCallum, "Topic models conditioned on arbitrary features with dirichlet-multinomial regression", 2012, arXiv: 1206.2778.

[4] D.I. Inouye, P.K. Ravikumar, I.S. Dhillon, "Admixture of Poisson MRFs: A topic model with word dependencies", *Proceedings of the International Conference on Machine Learning*, 2014, pp. 683-691.

[5] D.I. Inouye, P.K. Ravikumar, I.S. Dhillon, "Capturing semantically meaningful word dependencies with an admixture of Poisson MRFs", *Advances in Neural Information Processing Systems*, 2014, pp. 3158-3166.

[6] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality", *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[7] T. Mikolov, W.T. Yih, G. Zweig. "Linguistic regularities in continuous space word representations", *Proceedings of HLT-NAACL*, 2013, pp. 746–751.

[8] M. Xue, "A Text Retrieval Algorithm Based on the Hybrid LDA and Word2Vec Model", *Proceedings of 2019 International Conference on Intelligent Transportation, Big Data & Smart City* (ICITBS), 2019, pp. 373-376.

[9] D.T. Santosh, K.S. Babu, S.D.V. Prasad, A. Vivekananda, "Opinion mining of online product reviews from traditional LDA Topic Clusters using Feature Ontology Tree and Sentiwordnet", *Int. J. Educ. Manag. Eng*, Vol. 6, 2016, pp. 34-44.

[10] Y. Zhang, J. Ma, Z. Wang, "Semi supervised classification of scientific and technical literature based on semi supervised hierarchical description of improved latent dirichlet allocation (LDA)", *Cluster Computing*, 2018, pp. 1-9.

[11] G. Ren, T. Hong, "Investigating Online destination images using a topic-based sentiment analysis approach", *Sustainability*, Vol. 9, No. 10, 2017, pp. 17-65.

[12] A. Onan, S. Korukoglu, H. Bulut, "LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis", *Int. J. Comput. Linguistics Appl.*, Vol. 7, No. 1, 2016, pp. 101-119.

[13] W. Hong, Z. Hao, J. Shi, "Research and Application on Domain Ontology Learning Method Based on LDA", *Journal of Software*, Vol. 12, No. 4, 2017, pp. 265-273.

[14] N. Ko, B. Jeong, S. Choi, J. Yoon, "Identifying product opportunities using social media mining: application of topic modeling and chance discovery theory", *IEEE Access*, Vol. 6, 2017, pp. 1680-1693.

[15] K. Ali, H. Dong, A. Bouguettaya, A. Erradi, R. Hadjidj, "Sentiment analysis as a service: a social media based sentiment analysis framework", *Proceedings of 2017 IEEE International Conference on Web Services* (ICWS), 2017, pp. 660-667.

[16] T. Mikolov, J. Kopecky, L. Burget, O. Glembek, "Neural network based language models for highly inflective languages", *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4725-4728.

[17] T. Mikolov, S. Kombrink, L. Burget, J. Černocký, S. Khudanpur, "Extensions of recurrent neural network language model", *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2011, pp. 5528-5531.

[18] K. Benlamine, N. Grozavu, Y. Bennani, R. Nicoleta, K. Haddadou, A. Amamou, "Domain Name Recommendation based on Neural Network", *Procedia Computer Science*, Vol. 144, 2018, pp. 60-70.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[19] J. Pennington, R. Socher, C. D. Manning, "Glove: Global vectors for word representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP), 2014, pp. 1527-1543.

[20] Q. Le, T. Mikolov, "Distributed representations of sentences and documents", *Proceedings of the 31st International Conference on Machine Learning* (ICML-14), 2014, pp. 1188–1196.

[21] L. Niu, X. Dai, J. Zhang, J. Chen, "Topic2vec: learning distributed representations of topics", *Proceedings of the* 2015 International Conference on *Asian Language Processing* (IALP), 2015, pp. 193-196.

[22] C.E. Moody, "Mixing dirichlet topic models and word embeddings to make lda2vec", 2016, arXiv:1605.02019.

[23] D.M. Blei, A.Y. Ng, M.I. Jordan, "Latent dirichlet allocation", *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993-1022.

[24] K. Albishre, M. Albathan, Y. Li, "Effective 20 newsgroups dataset cleaning", *Proceedings of the 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (WI-IAT), VOl. 3, 2015, pp. pp. 98-101.

[25] G.M. Del Corso, A. Gulli, F. Romani: Ranking a stream of news. In: Proceedings of the 14th International Conference on World Wide Web, pp. 97–106. ACM (2005).

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

[27] P. Gupta, Y . Chaudhary, F. Buettner, H. Schütze, "Texttovec: Deep contextualized neural autoregressive topic models of language with distributed compositional prior". arXiv preprint arXiv:1810.03947. 2018.

[28] P. Gupta, Y . Chaudhary, F. Buettner, H. Schütze, "Document informed neural autoregressive topic models with distributional prior". In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, 2019. URL https://arxiv.org/abs/1809. 06709