GPLDA: A Generalized Poisson Latent Dirichlet Topic Model

Ibrahim Bakari Bala[1], Mohd Zainuri Saringat[2]
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
Johor, Malaysia

***Abstract*—The earliest modification of Latent Dirichlet Allocation (LDA) in terms of words or document attributes is by relaxing its exchangeability assumption via the Bag-of-word (BoW) matrix. Several authors have proposed many modifications of the original LDA by focusing on model that assumes the current topic depends on the words from previous topic. Most of the earlier work ignored the document length distribution since it is assumed that it will fizzle out at the modelling stage. Thus, in this paper, the Poisson document length distribution of LDA model is replaced with Generalized Poisson (GP) distribution which has the strength of capturing complex structures. The main strengths of GP are in capturing overdispersed (variance larger than mean) and under dispersed (variance smaller than mean) count data. The Poisson distribution used by LDA strongly relies on the assumption that the mean and variance of document lengths are equal. This assumption is often unrealistic with most real-life text data where the variance of document length may be greater than or less than their mean. Approximate estimate of the GPLDA model parameters was achieved using Newton-Raphson approximation technique of log-likelihood. Performance and comparative analysis of GPLDA with LDA using accuracy and F_1 showed improved results.**

Keywords*—Bag-of-word; generalized Poisson distribution; topic model; latent Dirichlet allocation

I. Introduction

In recent years, a stochastic generative model that has been used widely in the field of computer science with the focus on text mining and information retrieval is referred to as a topic model. Since the early proposition of the model, it has been used by many researchers in several fields such as text mining [2], computer vision [1], population genetics, and social networks [3].

Topic modelling can be traced to latent semantic indexing (LSI) by [4]. It is the basis of the developing topic models. However, LSI is not a probabilistic model. Hence uncertainty is not quantifiable. After the era of LSI, towards the search for a realistic probabilistic model, probabilistic latent semantic analysis (PLSA) by [5] was developed and served as the basis of modern topic models. As a further earlier extension of PLSA, [6] proposed latent Dirichlet allocation (LDA). The model was referred to as a complete generative stochastic model. Nowadays, there is a growing number of probabilistic models that are based on LDA via combination with particular tasks.

Since the introduction of topic models, researchers have introduced this approach into the fields of text mining. Because of its superiority in the analysis of large-scale document collections, better results have been obtained in such fields as text mining [7] and clinical informatics [8-9]. On the other hand, most of these studies follow the classic text-mining method of a topic model.

In LDA, we let d denotes the document indicator, z for the topic, w for word and consequently N is the number of words in a specific document d. Also, we define $P(z|d)$ as the conditional distribution of topic z in the document d and $P(w|z)$ as the conditional distribution of words w in topic z. The two conditional probability distributions, $\boldsymbol{P(z|d)}$ and $\boldsymbol{P(w|z)}$, are presumed to follow multinomial distributions such that the topics in the entire documents have common Dirichlet prior distribution $P(\boldsymbol{\alpha})$ and the word conditional distributions on topics have common Dirichlet prior $P(\boldsymbol{\beta})$ [10].

Algorithm 1: Pseudocode of LDA Algorithm

1. Sample N from Poisson $\mathbf{P(N = n|\lambda)}$
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 $\mathbf{\theta_d} \sim \mathbf{Dir(\theta_d|\alpha)}$
5. **for** each word w $\in \mathbf{d} \in \{1, 2, 3, \dots, N\}$:
6. Simulate $\mathbf{z_{dn}} \sim \mathbf{Mult(z_{dn}|\theta_d)}$
7. Simulate $\mathbf{w_{dn}} \sim \mathbf{Mult(w_{dn}|z_{dn}, \beta)}$
8. **end for w**
9. **end for d**
10. **end for k**

After the selection of appropriate prior hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for a document d, a conditional distribution of K topics with parameter θ is formed and it is assumed to be multinomially distributed from the Dirichlet distribution $Dir(\theta|\alpha)$. Also, for a specific topic k, a conditional distribution of V words are formed, and it is assumed to be multinomially distributed from the Dirichlet distribution $Mult(w|z, \beta)$. The Dirichlet prior distribution is choosing because of the conjugacy property between the multinomial and Dirichlet distribution which thus makes the statistical inference of LDA easy.

II. Related Work

The earliest modification of LDA in terms of words attributes is relaxing the exchangeability assumption of LDA via the BoW matrix by [11]. Wallach proposed a model that assumes that the current topic depends on the words from the previous topic. The method involves using a hierarchical procedure by combining the *n*-grams statistics procedure and latent topic models. Specifically, Wallach [11] extended the unigram topic model to include the properties of a hierarchical Dirichlet bigrams model. The author reported that the hybrid model is better than either of the unigram topic model or the Dirichlet bigram model. The results were inferred from two datasets consisting of 150 documents each. The model was supported by [12] with the claim that it is unrealistic to impose the exchangeability of words as orders of words matters when dealing with words contexts. Hu et al. [8] countered the class of models that either supported the exchangeability assumptions or relaxes it. The authors claimed the models are not interactive but rather employ several *apriori* fixes that are unrealistic. In addition, Inouye et al. [13] also exemplify that these class of models do not incorporate word dependencies within a topic but rather incorporates inter-topic word correlation which is the major strength of models such as Bigrams language model by [11].

Reisinger et al. [14] modified the word absences drawback features of LDA. The algorithm specifically improved the accuracy of LDA in terms of increasing the possibility of modelling rare words. The procedure addresses the use of multinomial draws by proposing the Von-Mises Fishers distribution for topics.

Most of the existing modifications targeted one or the other loopholes in LDA, but none has considered the overdispersed or under dispersed drawback that is inherent in text data. Thus, in this paper, the Poisson document length distribution of the Latent Dirichlet Allocation (LDA) model is replaced with Generalized Poisson (GP) distribution which has the strength of capturing complex structures. The new model referred to as GPLDA was tested on the 20-newsgroup dataset to facilitate comparison with the LDA.

III. Generalized Poisson Distribution

Suppose we have *N* documents that assumed Poisson distribution with rate λ, the probability mass function of having *n* realizations of *N* is given by [15]:

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

The Generalized Poisson (GP) [15-18] which is the extension of (1) can be defined in terms of additional dispersion parameter η as:

$$P(N = n|\,\lambda\,,\eta) = \lambda(\lambda + \eta n)^{n-1}\frac{\exp(-\lambda-\eta n)}{n!}, n = 0,1,2 \qquad (2)$$

It is obvious from (2) that GP can be reduced to Poisson when $\eta = 0$. The behaviour of the dispersion parameter η tell about the direction of disparity. If $\eta < 0$, underdispersion is suspected and $\eta > 0$ overdispersion is suspected.

A. Generalized Poisson Latent Dirichlet Allocation Model (GPLDA)

The GPLDA assumes the same structure as LDA except for the change in document length distribution. Mathematically, the joint distribution of document *N*, topics z, word w and topic mixture θ is defined as:

$$P(\theta, z, w|\alpha, \beta, \lambda\,, \eta)$$

$$= P(\theta|\alpha)\prod_{n=1}^{N} P(z_n|\theta)\,P(w_n|z_n, \beta) \times P(N = n|\,\lambda\,, \eta)$$

where;

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

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

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

Therefore,

$$P(\theta, z, w|\alpha, \beta, \lambda\,, \eta)$$

$$= \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\ldots\theta_k^{\alpha_k-1} \times \prod_{n=1}^{N}\left\{\left[\prod_{n=1}^{N}\frac{\Gamma\left(\sum_{i=1}^{k} z_{ni}+1\right)}{\prod_{i=1}^{k}\Gamma(z_{ni}+1)}\prod_{i=1}^{k}\theta_i^{z_{ni}}\right] \times \left[\prod_{n}^{N}\prod_{v=1}^{V}\frac{\Gamma\left(\sum_{i=1}^{k} w_{ni}+1\right)}{\prod_{i=1}^{k}\Gamma(w_{ni}+1)}\prod_{i=1}^{k}\beta_{iv}^{w_{ni}}\right] \times \left[\lambda(\lambda+\eta n)^{n-1}\frac{\exp(-\lambda-\eta n)}{n!}\right]\right\}$$

The marginal distribution of document *D* can be obtained by marginalizing the joint distribution $P(\theta, z, w|\alpha, \beta, \lambda\,, \eta)$ as follows:

Algorithm 2: Pseudocode of GPLDA Algorithm

1. Sample N from Generalized Poisson $\mathbf{P(N = n|\lambda\,, \eta)}$
2. **for** each topic $\mathbf{k} \in \{1, 2, 3, \ldots, K\}$:
3. **for** each document $\mathbf{d} \in \{1, 2, 3, \ldots, N\}$:
4. Simulate $\boldsymbol{\theta_d} \sim \mathbf{Dir}(\boldsymbol{\theta_d}|\boldsymbol{\alpha})$
5. **for** each word w $\in \mathbf{d} \in \{1, 2, 3, \ldots, N\}$:
6. Simulate $\mathbf{z_{dn}} \sim \mathbf{Mult}(\mathbf{z_{dn}}|\boldsymbol{\theta_d})$
7. Simulate $\mathbf{w_{dn}} \sim \mathbf{Mult}(\mathbf{w_{dn}}|\mathbf{z_{dn}}, \boldsymbol{\beta})$
8. **end for w**
9. **end for d**
10. **end for k**

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

B. Parameter Estimation of GPLDA

In this section, we present an approximate procedure for estimating the parameters of the GPLDA model. The Newton-Raphson approximation technique is employed by obtaining the log-likelihood of the distribution. The log-likelihood of the distribution of corpus of words in document *D* is:

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

The procedure involves obtaining the first and second partial derivatives which are intractable from the $\log[P(D|\theta, z, w, \alpha, \beta, b, a)]$. Thus, the Newton-Raphson approximation technique is used. The procedure involves finding the approximate derivatives of the $\log[P(D|\theta, z, w, \alpha, \beta, b, a)]$. The Newton-Raphson procedure for obtaining the parameters of GPLDA can be summarized below as:

1) Obtain the log-likelihood of the marginal distribution that is $\log[P(D|\theta_d, z, w, \alpha, \beta, b, a)]$.

2) Find the first derivative w.r.t to each parameter in the parameter space $\omega = \{\theta, \alpha, \beta, b, a\}$ and obtain the iterative estimate of parameter Ω using;

$$\omega_{t+1} = \omega_t - \frac{\log[P(D|z, w, \omega)]}{\partial \log[P(D|z, w, \Omega)]/\partial \omega}$$

The process continues until $|\omega_{t+1} - \omega_t| \leq \varepsilon$ where $\varepsilon \to 0$.

IV. Simulation Experiment

The behaviour of the GPLDA is observed by simulating several Generalized Poisson distributed variates with varying parameter η = -0.5, -0.4, -0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4, 0.5 and fixing rate parameter $\lambda = 5$. Fig. 1 shows the behavioural patterns for both under dispersed and overdispersed scenarios. All analyses were carried using the R package.

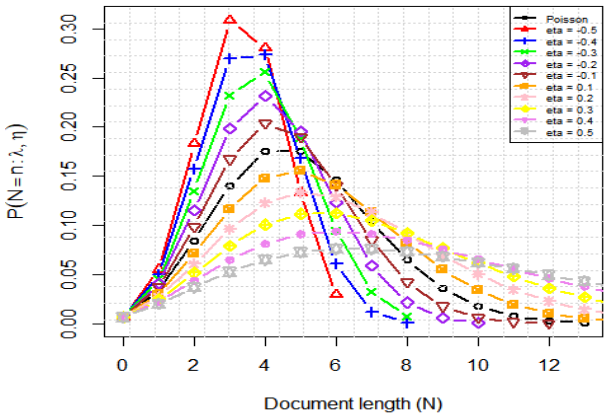

Fig. 1. Document Length Distribution at Various Dispersion Parameter η = -0.5, -0.4, -0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4, 0.5.

V. Performance Evaluation using 20-Newgroup Dataset

Performance evaluation of the GPLDA algorithm was achieved using the 20-Newsgroup dataset [19-22]. There are 18846 documents in the dataset, and it cut across 20 different topics categories. The topics in the classes include sports, politics, religion etc., which is diverse enough. The **Precision (P)** was used as class-specific index while **Recall (R)** (also known as sensitivity) is the proportion of the total amount of relevant cases that were actually retrieved [23-29]. The F_1 is a measure of the accuracy of the test dataset and it is defined as:

$$F_1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

VI. Results and Discussion

The plot in Fig. 1 confirms that when $\eta = 0$ the Generalized Poisson distribution reduces to Poisson distribution and consequently the GPLDA will reduce to LDA. The underdispersed situation yields observations with high probability of having values close to zero than Poisson while the overdispersed situation yields observation with low probability of having values close to zero than Poisson. The graph also confirms that the Poisson distribution only assumes the midpoint position by averaging the scenarios, this may be true but not in all cases.

TABLE. I. Performance Comparison for LDA and the Proposed GPLDA using 20-New Group Dataset

Performance	LDA	GPLDA	Relative Increase (%)
Accuracy	0.42	0.77	83.3
Micro F_1	0.60	0.87	45.0
Macro F_1	0.48	0.84	75.0

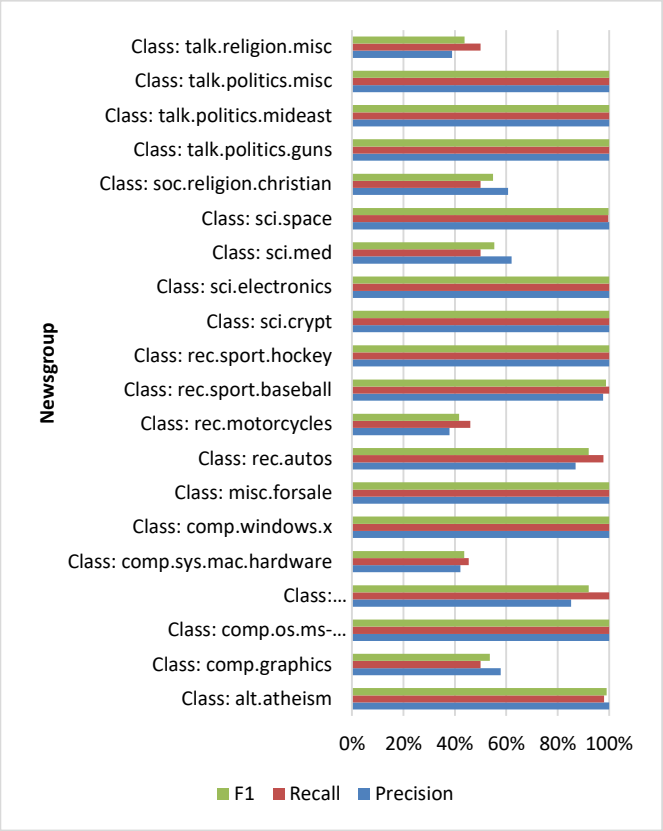

Fig. 2. Precision, Recall and F_1 Score in Percentage for Each News Group Class using GPLDA.

The predictive classification results performance analysis in Fig. 2 showed that the GPLDA algorithm results are high in terms of precision, recall and F_1 scores in 14 of the 20 classes but average on the other 6 topics/class. Performance comparison with LDA in Table I shows that the algorithm showed significant improvement over LDA. For Accuracy, GPLDA has about 83.3% percentage increased from the LDA result likewise for Micro F_1 45% increase and 75% increase for Macro F_1.

VII. Conclusion

This paper considered a new class of LDA using the Generalized Poisson distribution to model the length of a document. The Poisson distribution assumed by LDA has many stringent assumptions which are often violated in most real-life data. Thus, we propose the Generalized Poisson LDA (GPLDA) in order to provide a better fit. Estimation procedure was achieved using Newton-Raphson procedure and data calibration was done with the 20-Newsgroup dataset. The results from the simulation show that the Poisson distribution only assumes the midpoint position by averaging the scenarios which are not always correct. The results from the classification of 20-Newsgroup dataset showed that the GPLDA has an improved prediction over LDA. The results also established that the diversity in the Generalized Poisson over Poisson resulted in significant improvement. The GPLDA can be combined with the distributed learning system such as *word2vec* [10] to form a hybrid system like *lda2vec* by [30].

References

[1] W. Luo, B. Stenger, X. Zhao, and T.K., Kim. “Automatic topic discovery for multi-object tracking”, In Twenty-Ninth AAAI Conference on Artificial Intelligence. March 2015.

[2] L. C. Chen. “An effective LDA-based time topic model to improve blog search performance”, Information Processing & Management. Vol. 53, no. 6, pp. 1299-1319, November 2017.

[3] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei. “Author topic model-based collaborative filtering for personalized POI recommendations”, IEEE Trans Multimedia, vol. 17, no. 6, pp. 907–918, March 2015.

[4] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391

[5] Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42(1–2):177–196.

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

[7] Wang X, Zhu P, Liu T, Xu K (2016) BioTopic: a topic-driven biological literature mining system. Int J Data Mining Bioinform 14(4):373–386.

[8] Hu QV, He L, Li M, Huang JX, Haacke EM (2014) A semi-informative aware approach using topic model for medical search. 2014 IEEE international conference on bioinformatics and biomedicine (BIBM) 2014, pp 320–324.

[9] Huang Z, Dong W, Ji L, Gan C, Lu X et al (2014) Discovery of clinical pathway patterns from event logs using probabilistic topic models. J Biomed Inform 47:39–57.

[10] Xue, M. (2019). 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.

[11] Wallach HM (2006) Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on Machine learning, pp 977–984.

[12] 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.

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

[14] Reisinger, J., & Mooney, R. J. (2010). 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).

[15] Chandra, N. K., Roy, D., & Ghosh, T. (2013). A generalized Poisson distribution. Communications in Statistics-Theory and Methods, 42(15), 2786-2797.

[16] Consul, P. C. (1989). Generalized Poisson Distributions. New York: Dekker.

[17] Consul, P. C., & Famoye, F. (2006). Lagrangian probability distributions (pp. 21-49). Birkhäuser Boston.

[18] Joe, H., & Zhu, R. (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 47(2), 219-229.

[19] Albishre, K., Albathan, M., & Li, Y. (2015). 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.

[20] 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).

[21] 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.

[22] 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.

[23] 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). IOP Publishing.

[24] Olaniran, O. R., Olaniran, S. F., Yahya, W. B., Banjoko, A. W., Garba, M. K., Amusa, L. B., & Gatta, N. F. (2016). Improved Bayesian feature selection and classification methods using bootstrap prior techniques. Annals. Computer Science Series, 14(2), 46-52.

[25] Olaniran, O. R., & Yahya, W. B. (2017). Bayesian hypothesis testing of two normal samples using bootstrap prior technique. Journal of Modern Applied Statistical Methods, 16(2), 34.

[26] Olaniran, O. R., & Abdullah, M. A. A. (2017). Gene Selection for Colon Cancer Classification using Bayesian Model Averaging of Linear and Quadratic Discriminants. Journal of Science and Technology, 9(3).

[27] Olaniran, O. R., & Abdullah, M. A. A. (2019a). Bayesian Variable Selection for Multiclass Classification using Bootstrap Prior Technique. Austrian Journal of Statistics, 48(2), 63-72.

[28] Olaniran, O. R., & Abdullah, M. A. A. B. (2019b). BayesRandomForest: An R implementation of Bayesian Random Forest for Regression Analysis of High-dimensional Data. In Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017) (pp. 269-275). Springer, Singapore.

[29] Olaniran, O. R., & Abdullah, M. A. A. B. (2019c). Bayesian Random Forest for the Classification of High-Dimensional mRNA Cancer Samples. In Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017) (pp. 253-259). Springer, Singapore.

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