Name: Heera Kesavan Ezhuthachan || **Student ID:** 2010488

Nonparametric Method: In this **paper [1],** the Authors develop a model that uses Nonparametric distributions for capturing Topics are varying with. The np-TOT model was evaluated with other models using Joint Likelihood and Perplexity of Document and time stamp. It was found the model np-TOT has best likelihood and Lower Perplexity on all three Datasets. In this **paper [2],** the authors used the NIPS Dataset the Hierarchical DP does not assume a finite fixed number of Topics. The fully Hierarchical model performs best on the Dataset with low perplexity.

Tuning Method: In this **paper [3**] the Authors used three different datasets and Models were build using n - fold Cross Validation. The stability was evaluated by comparing RPC Based method with Perplexity based method and Efficiency was evaluated by performing Cluster Analysis on output of LDA Models. the number of topics were 20,40, 50 for the Salmonella sequence dataset, the TCBB dataset, SIDER2 dataset, respectively. In this **Paper [4]** Authors have proposed a model “that integrates the class label information and the word order structure into the topic model itself” [4]. The Number of Topics is 80, 70,10,20 for validation set and F- measure of Test Set are 0.875, 0.639, 0.314, 0.419 of 20 Newsgroups, OHSUMED-23, TechTC300, Reuters-21578 Respectively.

Varying Latent Topics: In this **Paper [5],** the Authors have ‘implemented the Dynamic HDP Model using variational Bayesian Inference applied to the United States presidential State of the Union addresses from 1790 to 2008’ [5] the number of Topics vary with respect to time the topics are important. In this **paper [6],** the Authors have used Twitter Dataset with keyword ‘Domestic-Violence’. The Number of Topics was set to 20 by using Structural Units Bigrams. Out of 80868 Bigrams Top 20 words with highest percentage were chosen. When Topic distributions was calculated by date. There were changes in topics w.r.t time.

References

1. Dubey, Avinava, et al. "A nonparametric mixture model for topic modeling over time." Proceedings of the 2013 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2013.
2. Teh, Yee Whye, et al. "Hierarchical dirichlet processes." Journal of the american statistical association 101.476 (2006): 1566-1581.
3. Zhao, Weizhong, et al. "A heuristic approach to determine an appropriate number of topics in topic modeling." BMC bioinformatics. Vol. 16. No. 13. BioMed Central, 2015.
4. Jameel, Shoaib, Wai Lam, and Lidong Bing. "Supervised topic models with word order structure for document classification and retrieval learning." Information Retrieval Journal 18.4 (2015): 283-330.
5. Blei, David, Lawrence Carin, and David Dunson. "Probabilistic topic models." IEEE signal processing magazine 27.6 (2010): 55-65.
6. Xue, Jia, Junxiang Chen, and Richard Gelles. "Using data mining techniques to examine domestic violence topics on Twitter." Violence and gender 6.2 (2019): 105-114.

Dataset: 20 Newsgroup Dataset

Preprocessing Steps: Converted text corpus to lowercase , Removed punctuations and Non – ASCII characters, Replaced multiple spaces with single space, Removed Stopwords using NLTK Stopwords , Performed lemmatization using Wordnetlemmatizer , Tokenized words using NLTK word tokenize.

Model Codes: Gensim LDA model

Experimental setting: SVM Classification

Used classification as evaluation metric helps reach optimal number of topics also this dataset has labels which helps us in classification. Evaluation metric used is **F1 Score weigted average.**

Hyperparmeters: After Optimizing, alpha = 10

Optimized Alpha = [0.07454832, 0.18625298, 0.38022324, 0.8574413, 0.2655421 ,0 .6303619, 0.10972629, 0.01325522, 0.32956785, 0.04606212]

Topic model matrix used is Document topic Matrix

Experiments performed: The Dataset was split into Train-Test-Validation model

Train = 70%, Test = 15%, Validation = 15%

No.of iterations = 10

Validation dataset:

Number of Topics	F1 Score
10	0.63
20	0.59
30	0.56
40	0.55

Test Dataset:

Number of Topics	F1 Score
20	0.59