Name: Ahmad Raza **Registration Number:** 2101194

Tuning: In [1], taking TREC Genomic protocol 2005 dataset, authors used Normalized Absolute Coherence and Normalized Absolute Perplexity along with 5-fold cross validation. They found that optimal number of topics for range [2,500] calculated by NAC and NAP is 43 and 2 respectively. Authors in [2] chose a range of [40,150] no. of topics and found that 90 topics describe their dataset. They used Rate of Perplexity Change and 5-fold cross validation on a dataset containing geo-referenced tweets from three USA East Coast cities. In [3], authors analyzed a dataset containing customer reviews from Amazon, eBay, and Best Buy using 10-fold cross validation and coherence to find ideal number of topics to be 8.

Varying Topics: Authors in [4] analyzed dataset containing abstracts from journals of Informatics in China. They computed perplexity on different number of topics upto 100 and found that 35 topics were ideal. Using 28,154 abstracts published in PNAS from 1991 to 2001 and varying the number of topics from 50 to 1000, authors in [5] found that 300 topics ideally describe their dataset. In [6], authors experimented on a range [10,100] number of topics using data from mail gate incident and found that optimal number of topics calculated by using perplexity and KM-SSVW(keyword matching and subjective statistical value word comparison) is 65 and 30 respectively.

Bayesian non-parametric: With values γ Gamma (1, 0.1) and α_0 Gamma(1, 1) for HDP model, using 10 fold cross-validated accuracy, [7] utilizes a MIMIC II dataset to learn optimal number of topics to be [39,44]. Using coherence metric, authors in [8] found that 16 and 92 topics best describe the 20 News Group and NIPS dataset respectively. They used sHDP model.

References:

[1] Hasan, Mahedi & Rahman, Anichur & Razaul, Md & Khan, Md & Islam, Md. (2021). 'Normalized Approach to Find Optimal Number of Topics in Latent Dirichlet Allocation (LDA)'. 10.1007/978-981-33-4673-4_27.

[2] Cheng Fu, Grant McKenzie, Vanessa Frias-Martinez, Kathleen Stewart, 'Identifying spatiotemporal urban activities through linguistic signatures, Computers, Environment and Urban Systems', Volume 72, 2018, Pages 25-37, ISSN 0198-971.

[3] Joung, Junegak, and Harrison M. Kim. 'An lda-based approach for product attribute identification from online customer reviews.' (2020).

[4] Zhu, Maoran, Xiaopeng Zhang, and Hongwei Wang. 'A LDA based model for topic evolution: Evidence from information science journals.' *2016 international conference on modeling, simulation and optimization technologies and applications.*

[5] T. L. Griffiths and M. Steyvers. 'Finding scientific topics'. Proceedings of the National academy of Sciences, 101(suppl 1):5228–5235, 2004.

[6] Gong, Hechen, Fucheng You, and Shuren Lai. 'Research on Evaluation Method of LDA Topic Model in Mail Classification'. *Journal of Physics: Conference Series*.

[7] Lehman, Li-wei, et al. 'Risk stratification of ICU patients using topic models inferred from unstructured progress notes.' *AMIA annual symposium proceedings*.

[8] Batmanghelich K, Saeedi A, Narasimhan K, Gershman S. 'Nonparametric Spherical Topic Modeling with Word Embeddings'. Proc Conf Assoc Comput Linguist Meet. 2016 Aug;2016:537-542.

1.Dataset: Restaurant Reviews data from 'Yelp' for year 2016-17 has been used.
2.Data Preparation: **a)** Only Reviews with text length (50,200) filtered to use. **b)** Star ratings(1/2/3/4/5) column converted to binary values(0/1) to be used as a target column to classify positive and negative reviews using text from text column.
3.Text-Preprocessing: **a)** Along with nltk stopwords, an extended list of 18 stopwords removed. **b)** New lines replaced with spaces whereas, **c)** multiple spaces replaced with a single space. **d)** Gensim's simple_preprocess() function used for tokenization and punctuations removal. **e)** Bigrams are formed and finally **f)** lemmatization is done allowing only noun and adverb postags.
4.Model Codes: Gensim has been used for LDA topic modeling.
5.Experiment Setting: As dataset contains labels, therefore binary Classification setting used with F1 as evaluation metric.
6.Hyper-parameters: Optimized values obtained through tuning approach for iterations[20,50,80]=**50**, alpha[0.005,0.05,0.5]=**0.5** and eta[0.005,0.05,0.5]=**0.5**
7.Description of methodology used to achieve optimal topics: a) Reviews for year 2016 used as training set(108356) whereas reviews for year 2017 used as test set(121393). **b)** Both sets were pre-processed using steps mentioned in section 3 above. **c)** After final pre-processing of training set, a corpus in the form of sparse matrix containing unique id of words and their occurrence frequency was obtained using Gensim's Dictionary class and doc2bow function to be fed to Gensim's LDA model **d)** To determine optimal value of number of topics, LDA model was run multiple times for number of topics [10,20,30,70,100,150,250]. Every time document by topic matrix was obtained from the LDA model and fed to the RandomForestClassifier. Obtained feature matrix and target column was divided into 5 folds where data in 4 folds was used to train the classifier whereas remaining unseen data was used as a validation set to compute f1 score. Scores for all folds were then averaged to get a mean f1 score and saved against number of topics **e)** LDA model giving the highest f1 score was saved to be used for test data. **f)** Lemmatized test data and word mappings from train data were used to obtain sparse matrix for test data to be fed to the saved optimized lda model. **g)** Feature Matrix obtained with test data was fed to the RandomForestClassifier to compute the F1 score for test data to verify the generalization of our optimized model.

Training(number topics)	**Tuning (f1 score)**	**Test (f1 score)**
10	0.8861	
20	0.9183	
30	**0.9307**	**0.9521**
70	0.9115	
100	0.8522	
150	0.8231	
250	0.7909	