Bodh", BPIT's International Journal of Technology & Management
ISSN: 2454-8421, Volume 7, Issue 2, July- Dec 2021, Page 19-25

A Multi-Hierarchical Clustering on Topic Modeling using Latent Dirichlet Allocation

Anusha Chhabra[1], Monika Arora[2]
[1]*Research Scholar, Delhi Technological University,*
Delhi, India
[2]*Department of Information Technology, Bhagwan Parshuram Institute of Technology, Delhi, India,*

[1]anusha.chhabra@gmail.com, [2]monikaarora@bpitindia.com

Abstract- With the economic crisis, it is expected that credit and down payment growth will average later on. Credit score growth will be led by spending on the infrastructure while retail credit will display an average growth. Margin demands due to lag effect of quantity cuts between attention quantity on deposits and advances, reduced treasury gains and core fee earnings and improving in conditions for NPAs is likely to put pressure on the main point here of the financial organizations. In the light of above aspects the present document tends to analyze sectoral distribution of credit in Native India Financial industry. The document studies the styles in credit growth with a perspective to project upcoming courses of growth in bank credit.

Key words: Unsupervised learning, K-Means Clustering, MALLET, WEKA

I. INTRODUCTION

Topic modeling is the technique utilized in the area of text analysis. It is used for finding the abstract topics occurring in an assortment of documents, which provides us with strategies to arrange, comprehend, and sum up enormous assortments of text based data. It is unique in relation to rule-based text analysis approaches that utilize regular expressions or word reference based keyword-finding methods. It additionally helps in finding hidden topical patterns available across the assortment of data. Topic models have numerous applications in natural processing dialects. Many papers have been published using topic modeling approaches in different subjects like online social media stages, software engineering, Cognitive science and many more. Latent Dirichlet Allocation (LDA) [1], a generative probabilistic topic model is an amazing asset of topic modeling that permits a set of perceptions to be clarified by unobserved groups. The instinct behind LDA is that data archives exhibit numerous topics. Each record shows the topics to various extent; each word in each archive is extracted from one of the topics, where the chosen topic is taken from the per-report dissemination over topics [1], [2].

The unsupervised learning via probabilistic topic model [3] has been successfully developed for document categorization[3] [1], image analysis [4], text segmentation [5], speech recognition [6], information retrieval [7], document summarization [8], [9], and many other applications. Using topic models, latent semantic topics are learned from a bag of words to capture the salient aspects embedded in data collection. If the number of clusters is considered as a number of topics and the probabilities as the extent of cluster participation, then LDA is used as a way of soft-clustering. Cluster analysis is an errand of grouping a set of objects so that objects of the same group are more similar in any sense to one another than to those in other groups. In LDA, the order of archives doesn't make any difference except hyper-parameters which rely upon whether Dirichlet distributions are assumed to be symmetric or asymmetric. For the symmetric distribution considering on alpha value, a high alpha-value is considered means that each corpus is probably going to contain a combination of maximum number of topics, not any single one explicitly [10].

On the other hand, low alpha value puts less such constraints on corpus with the meaning where it is more likely that a corpus contains the combinations of a few or even one topic. Whereas If considering beta values, a higher beta-value represents that each and every topic is likely to have a combination of maximum number of words, not any word explicitly, while a low beta value represents that a topic contains a combination of just a few of the words [1].

Table I. Parameters of LDA and their relevance with respect to documents and topics

Hyper-parameters	**Relevance**	**> values**	**< values**
Alpha	Document-topic density	Documents composed of large number of topics	Documents contain fewer number of topics
Beta	Topic-word density	Topics composed of a large amount of words in the document	Topics composed of few words

With the reference of parameters of LDA as in Table I, LDA model searches for rehashing term designs in the whole Document-Term framework while K-Means grouping relies a lot upon the determination of starting point of convergence [11]. So, in contrast with LDA, K-Means belongs to one cluster for each entity. So, to perform Clustering on the document containing the topics probabilities present in the corpus we are using WEKA. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA contains devices for information pre-handling, classification, regression, grouping, affiliation rules, and representation. Presently, Applying K-means clustering on output generated from LDA for example, the text record which is showing the breakdown, by rate, of every topic inside every unique text document we imported or we can say the points that form our archives. The next section begins with a review of current studies related to LDA. In Section2, discusses the Literature review, Section 3 discusses the proposed methodology. Section 4 elucidates the experimental setups and Results and Analysis has been presented in Section5. In section 6 we conclude the paper along with suggestions for future research opportunities.

II. LITERATURE REVIEW

This section summarizes a few scholarly works proposed about extracting the topics using Latent Dirichlet Allocation (LDA) and various types of clustering used to cluster similar data points. Related work can also be found in the above mentioned applications in multi-domain data. Authors in a research paper [2], proposed an inference algorithm for LDA results that leads to identify documents and the relationships between the documents. [12], applied initial clustering center for K-Means as preprocessing which further makes the latent clustering center more focus on a certain topic.[10], proposed multi-grain clustering topic model which integrates topic modeling and document clustering together. [13], tried to combine LDA with document-specific word distributions for organizing large documents archives. [14], evaluated LDA with the perspective of classifying and identifying the noteworthy topics further applied to filtered collection of Twitter. [1], proposed a generative statistical model, which projects a document into topic space, and each topic contains multiple words. [15], focuses on calculating the temporal weights to reveal the importance of all the topics extracted from the probabilistic approach. [16], introduced an improved K-Means algorithm solving the limitation of traditional K-Means algorithm in terms of dependency on selecting the initial focal points by combining the largest minimum distance algorithm and the traditional K-Means algorithm. Authors in [17] conducted a comprehensive survey on already existing works. Predominantly, they concentrated on event feature learning that has ability on translating social media data into computer friendly numerical form. In [18], Different methods of document clustering and topic modeling on social media content offers a technique of categorizing, interpreting and understanding the large volume of user generated content. Authors have implemented hybrid models to cluster large document sets. Based on contextual keywords, they optimized the clustering similarity index to find the essential key document clusters. Their experimental results show that the clustering based document classification models have better performance [19].

III. METHODOLOGY

This section depicts the proposed methodology for finding the probabilities of topic distribution after implementing via Single-Link Type, Complete-Link Type and Average-Link Type Hierarchical clustering. The Process diagram for implementation is in Fig. 1.

A. **Single Linkage Type:** In this approach, we calculate the similarity for the closest pair. The limitation for this method is that close pair groups merge faster than the optimal, regardless of whether there is dissimilarity for those groups.

B. **Complete Linkage Type:** This approach is based on the phenomenon for calculating similarity of distant pairs along with the limitation of having outliers which causes less merging than optimal one.

C. **Average Linkage, or group linkage Type:** In this, grouping of objects are considered for calculating the similarity, rather than individual objects.

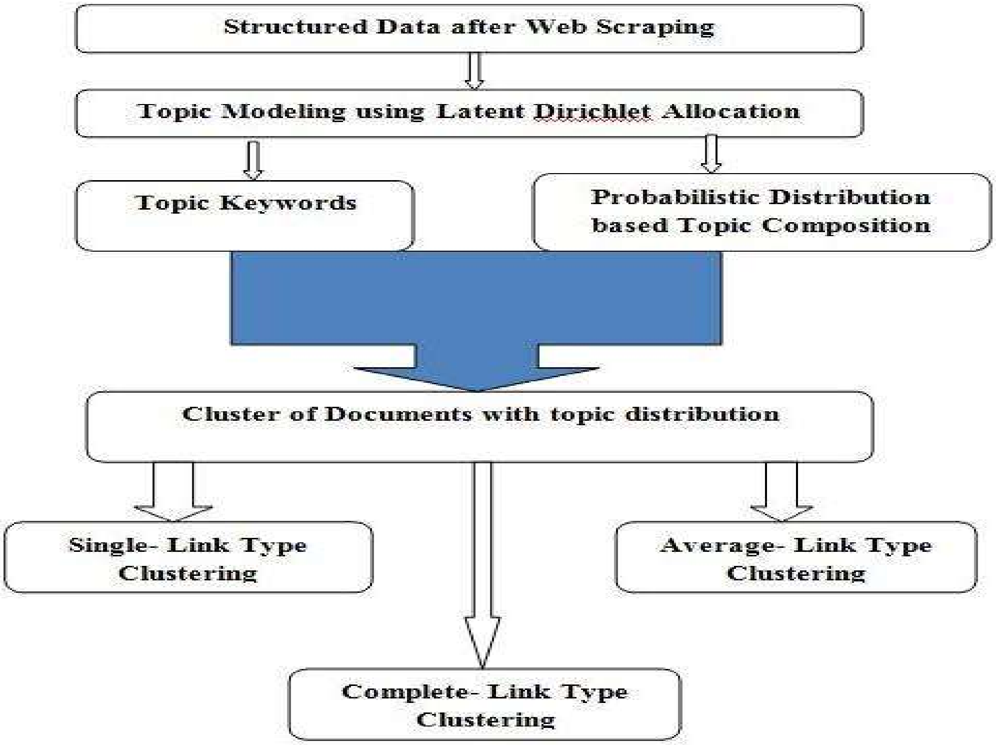
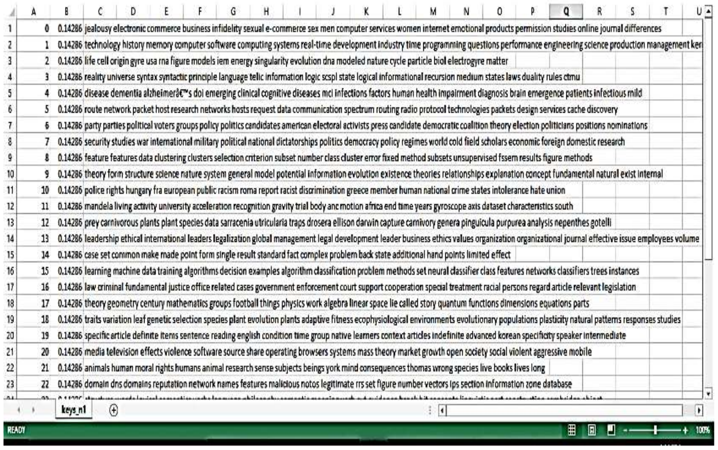

Structured Data after Web Scraping
Topic Modeling using Latent Dirichlet Allocation
Topic Keywords
Probabilistic Distribution based Topic Composition
Cluster of Documents with topic distribution
Single- Link Type Clustering
Average- Link Type Clustering
Complete- Link Type Clustering

Fig. 1. Process Flow Diagram

The Process flow diagram as in Fig.1 starts with the structured data generated after web scraping. The extracted HTML code from web scraping, we stored the data in the database. The generated dataset is then normalized and LDA is applied to extract the topic keywords and probabilistic distribution-based topic composition. The documents obtained after implementing LDA, passed through soft clustering techniques which results in clustering of documents with topic distribution classified against the types of hierarchical clustering as Single-Link type, Complete-Link type and Average- Link type clustering. The experiments have been carried out on WEKA. In this paper, the dataset has been generated by using the BeautifulSoup4 Python tool of web scraping; the python script has been employed to generate the dataset. The generated dataset in the proposed method has been stored in two text files that comprises top keywords for each topic in one file and weight distribution of each topic in another file.

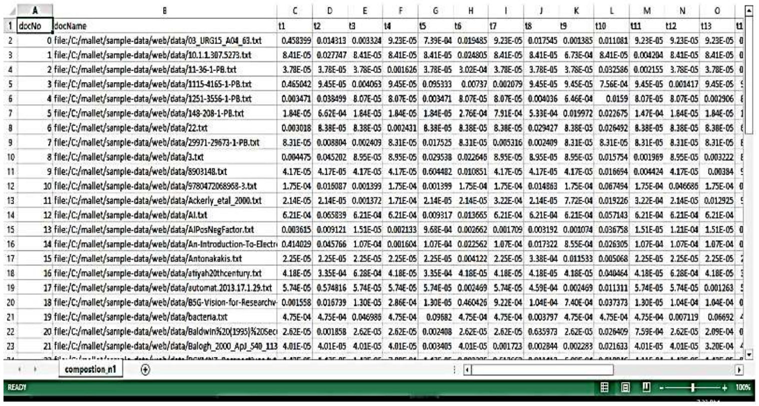
IV. RESULTS AND ANALYSIS

The output using the commands in MALLET is stored in 2 files. 1st file- keys_n1.txt is representing what all the top keywords are for each topic as in Fig.2 and the 2nd file- composition_n1.txt indicates the breakdown, by weight distribution, of every topic extracted from original textual file as in Fig.3.



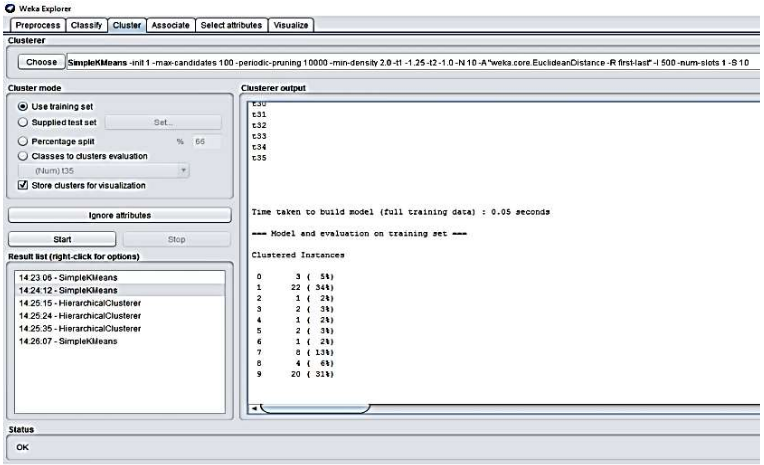
	A	B	C
1	0	0.14286	jealousy electronic commerce business infidelity sexual e-commerce sex men computer services women internet emotional products permission studies online journal differences
2	1	0.14286	technology history memory computer software computing systems real-time development industry time programming questions performance engineering science production management ker
3	2	0.14286	life cell origin gyre usa rna figure models iem energy singularity evolution dna modeled nature cycle particle biol electrogyre matter
4	3	0.14286	reality universe syntax syntactic principle language telic information logic scspl state logical informational recursion medium states laws duality rules ctmu
5	4	0.14286	disease dementia alzheimerâ€™s doi emerging clinical cognitive diseases mci infections factors human health impairment diagnosis brain emergence patients infectious mild
6	5	0.14286	route network packet host research networks hosts request data communication spectrum routing radio protocol technologies packets design services cache discovery
7	6	0.14286	party parties political voters groups policy politics candidates american electoral activists press candidate democratic coalition theory election politicians positions nominations
8	7	0.14286	security studies war international military political national dictatorships politics democracy policy regimes world cold field scholars economic foreign domestic research
9	8	0.14286	feature features data clustering clusters selection criterion subset number class cluster error fixed method subsets unsupervised fssem results figure methods
10	9	0.14286	theory form structure science nature system general model potential information evolution existence theories relationships explanation concept fundamental natural exist internal
11	10	0.14286	police rights hungary fra european public racism roma report racist discrimination greece member human national crime states intolerance hate union
12	11	0.14286	mandela living activity university acceleration recognition gravity trial body anc motion africa end time years gyroscope axis dataset characteristics south
13	12	0.14286	prey carnivorous plants plant species data sarracenia utricularia traps drosera ellison darwin capture carnivory genera pinguicula purpurea analysis nepenthes gotelli
14	13	0.14286	leadership ethical international leaders legalization global management legal development leader business ethics values organization organizational journal effective issue employees volume
15	14	0.14286	case set common make made point form single result standard fact complex problem back state additional hand points limited effect
16	15	0.14286	learning machine data training algorithms decision examples algorithm classification problem methods set neural classifier class features networks classifiers trees instances
17	16	0.14286	law criminal fundamental justice office related cases government enforcement court support cooperation special treatment racial persons regard article relevant legislation
18	17	0.14286	theory geometry century mathematics groups football things physics work algebra linear space lie called story quantum functions dimensions equations parts
19	18	0.14286	traits variation leaf genetic selection species plant evolution plants adaptive fitness ecophysiological environments evolutionary populations plasticity natural patterns responses studies
20	19	0.14286	specific article definite items sentence reading english condition time group native learners context articles indefinite advanced korean specificity speaker intermediate
21	20	0.14286	media television effects violence software source share operating browsers systems mass theory market growth open society social violent aggressive mobile
22	21	0.14286	animals human moral rights humans animal research sense subjects beings york mind consequences thomas wrong species live books lives long
23	22	0.14286	domain dns domains reputation network names features malicious notos legitimate rrs set figure number vectors ips section information zone database

Fig. 2. Top Keywords



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	docNo	docName	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13
2	0	file:/C:/mallet/sample-data/web/data/03_URG15_A04_63.txt	0.458399	0.014313	0.003324	9.23E-05	7.39E-04	0.019485	9.23E-05	0.017545	0.001385	0.011081	9.23E-05	9.23E-05	9.23E-05
3	1	file:/C:/mallet/sample-data/web/data/10.1.1.307.5273.txt	8.41E-05	0.027747	8.41E-05	8.41E-05	8.41E-05	0.024805	8.41E-05	8.41E-05	6.73E-04	8.41E-05	0.004204	8.41E-05	8.41E-05
4	2	file:/C:/mallet/sample-data/web/data/11-36-1-PB.txt	3.78E-05	3.78E-05	3.78E-05	0.001626	3.78E-05	3.02E-04	3.78E-05	3.78E-05	3.78E-05	0.032586	0.002155	3.78E-05	3.78E-05
5	3	file:/C:/mallet/sample-data/web/data/1115-4165-1-PB.txt	0.465042	9.45E-05	0.004063	9.45E-05	0.095333	0.00737	0.002079	9.45E-05	9.45E-05	7.56E-04	9.45E-05	0.001417	9.45E-05
6	4	file:/C:/mallet/sample-data/web/data/1251-3556-1-PB.txt	0.003471	0.038499	8.07E-05	8.07E-05	0.003471	8.07E-05	8.07E-05	0.004036	6.46E-04	0.0159	8.07E-05	8.07E-05	0.002906
7	5	file:/C:/mallet/sample-data/web/data/148-208-1-PB.txt	1.84E-05	6.62E-04	1.84E-05	1.84E-05	1.84E-05	2.76E-04	7.91E-04	5.33E-04	0.019972	0.022675	1.47E-04	1.84E-05	1.84E-05
8	6	file:/C:/mallet/sample-data/web/data/22.txt	0.003018	8.38E-05	8.38E-05	0.002431	8.38E-05	8.38E-05	8.38E-05	0.029427	8.38E-05	0.026492	8.38E-05	8.38E-05	8.38E-05
9	7	file:/C:/mallet/sample-data/web/data/29971-29673-1-PB.txt	8.31E-05	0.008804	0.002409	8.31E-05	0.017525	8.31E-05	0.005316	0.002409	8.31E-05	8.31E-05	8.31E-05	8.31E-05	8.31E-05
10	8	file:/C:/mallet/sample-data/web/data/3.txt	0.004475	0.045202	8.95E-05	8.95E-05	0.029538	0.022646	8.95E-05	8.95E-05	8.95E-05	0.015754	0.001969	8.95E-05	0.003222
11	9	file:/C:/mallet/sample-data/web/data/8903148.txt	4.17E-05	4.17E-05	4.17E-05	4.17E-05	0.604482	0.010851	4.17E-05	4.17E-05	4.17E-05	0.016694	0.004424	4.17E-05	0.00384
12	10	file:/C:/mallet/sample-data/web/data/9780472068968-3.txt	1.75E-04	0.016087	0.001399	1.75E-04	0.001399	1.75E-04	1.75E-04	0.014863	1.75E-04	0.067494	1.75E-04	0.046686	1.75E-04
13	11	file:/C:/mallet/sample-data/web/data/Ackerly_etal_2000.txt	2.14E-05	2.14E-05	0.001372	1.71E-04	2.14E-05	2.14E-05	3.22E-04	2.14E-05	7.72E-04	0.019226	3.22E-04	2.14E-05	0.012925
14	12	file:/C:/mallet/sample-data/web/data/AI.txt	6.21E-04	0.065839	6.21E-04	6.21E-04	0.009317	0.013665	6.21E-04	6.21E-04	6.21E-04	0.057143	6.21E-04	6.21E-04	6.21E-04
15	13	file:/C:/mallet/sample-data/web/data/AIPosNegFactor.txt	0.003615	0.009121	1.51E-05	0.002133	9.68E-04	0.002662	0.001709	0.003192	0.001074	0.036758	1.51E-05	1.21E-04	1.51E-05
16	14	file:/C:/mallet/sample-data/web/data/An-Introduction-To-Electr	0.414029	0.045766	1.07E-04	0.001604	1.07E-04	0.022562	1.07E-04	0.017322	8.55E-04	0.026305	1.07E-04	1.07E-04	1.07E-04
17	15	file:/C:/mallet/sample-data/web/data/Antonakakis.txt	2.25E-05	2.25E-05	2.25E-05	2.25E-05	2.25E-05	0.004122	2.25E-05	3.38E-04	0.011533	0.005068	2.25E-05	2.25E-05	2.25E-05
18	16	file:/C:/mallet/sample-data/web/data/atiyah20thcentury.txt	4.18E-05	3.35E-04	6.28E-04	4.18E-05	3.35E-04	4.18E-05	4.18E-05	4.18E-05	4.18E-05	0.040464	4.18E-05	6.28E-04	4.18E-05
19	17	file:/C:/mallet/sample-data/web/data/automat.2013.17.1.29.txt	5.74E-05	0.574816	5.74E-05	5.74E-05	5.74E-05	0.002469	5.74E-05	4.59E-04	0.002469	0.011311	5.74E-05	5.74E-05	0.001263
20	18	file:/C:/mallet/sample-data/web/data/B5G-Vision-for-Researchv	0.001558	0.016739	1.30E-05	2.86E-04	1.30E-05	0.460426	9.22E-04	1.04E-04	7.40E-04	0.037373	1.30E-05	1.04E-04	1.04E-04
21	19	file:/C:/mallet/sample-data/web/data/bacteria.txt	4.75E-04	4.75E-04	0.046986	4.75E-04	0.09682	4.75E-04	4.75E-04	0.003797	4.75E-04	4.75E-04	4.75E-04	0.007119	0.06692
22	20	file:/C:/mallet/sample-data/web/data/Baldwin%20(1995)%20Sec	2.62E-05	0.001858	2.62E-05	2.62E-05	0.002408	2.62E-05	2.62E-05	0.635973	2.62E-05	0.026409	7.59E-04	2.62E-05	2.09E-04
23	21	file:/C:/mallet/sample-data/web/data/Balogh_2000_ApJ_540_113	4.01E-05	4.01E-05	4.01E-05	4.01E-05	0.003405	4.01E-05	0.001723	0.002844	0.002283	0.021633	4.01E-05	4.01E-05	3.20E-04

Fig.3. Topic Composition



Clusterer

Choose SimpleKMeans -init 1 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 10 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

- ● Use training set
- ○ Supplied test set Set...
- ○ Percentage split % 66
- ○ Classes to clusters evaluation (Num) t35
- ☑ Store clusters for visualization

Result list (right-click for options)

- 14:23:06 - SimpleKMeans
- 14:24:12 - SimpleKMeans
- 14:25:15 - HierarchicalClusterer
- 14:25:24 - HierarchicalClusterer
- 14:25:35 - HierarchicalClusterer
- 14:26:07 - SimpleKMeans

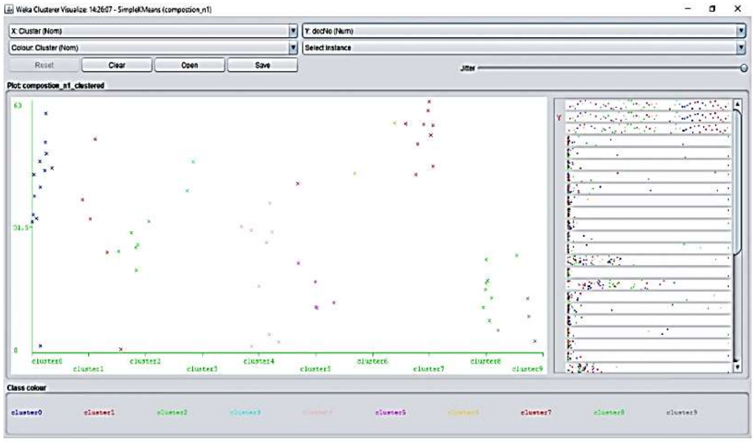
Clusterer output

```
t30
t31
t32
t33
t34
t35



Time taken to build model (full training data) : 0.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

 0        3 (  5%)
 1       22 ( 34%)
 2        1 (  2%)
 3        2 (  3%)
 4        1 (  2%)
 5        2 (  3%)
 6        1 (  2%)
 7        8 ( 13%)
 8        4 (  6%)
 9       20 ( 31%)
```

Status

OK

Fig. 4. K-Means Clustering



Weka Clusterer Visualize: 14:26:07 - SimpleKMeans (compostion_n1)

X: Cluster (Nom)

Y: docNo (Num)

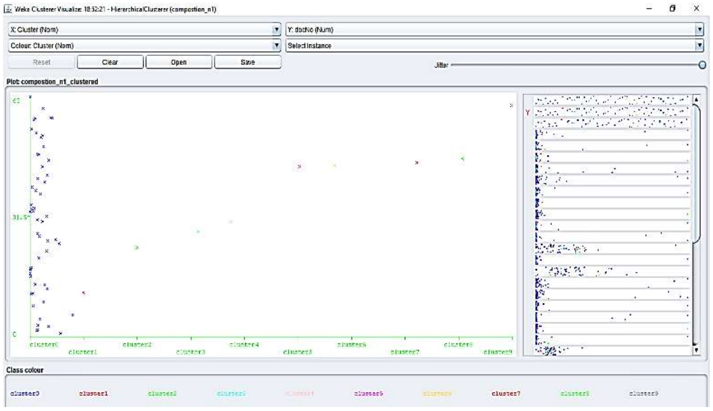
Colour: Cluster (Nom)

Select Instance

Plot: compostion_n1_clustered

Fig. 5. Cluster of Documents with topic distribution present in the document corpus

We can also visualize the clusters that are formed. The clusters formed after applying K- means clustering on the topics composing the documents as in Fig. 5.



Weka Clusterer Visualize: 18:52:21 - HierarchicalClusterer (compostion_n1)

X: Cluster (Nom)

Y: docNo (Num)

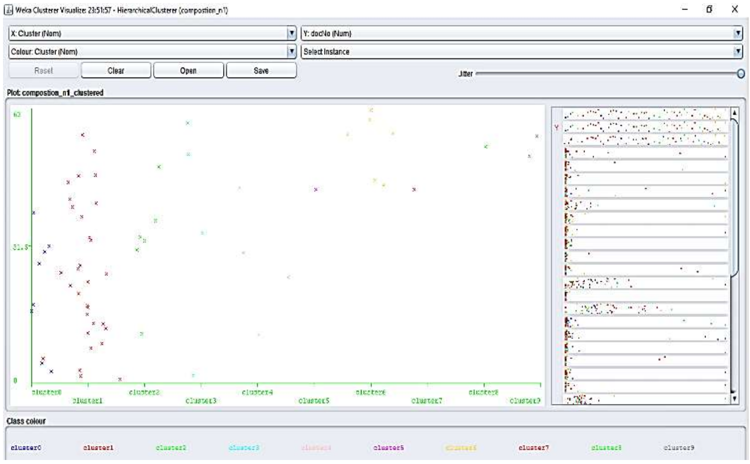
Colour: Cluster (Nom)

Select Instance

Plot: compostion_n1_clustered

Fig. 6. Single-link type Hierarchical clustering

The results of the Single link, complete link and Average-link of the Hierarchical clustering are as in Fig 6, Fig.7 and Fig. 8.



Weka Clusterer Visualize: 23:51:57 - HierarchicalClusterer (compostion_n1)

X: Cluster (Nom)

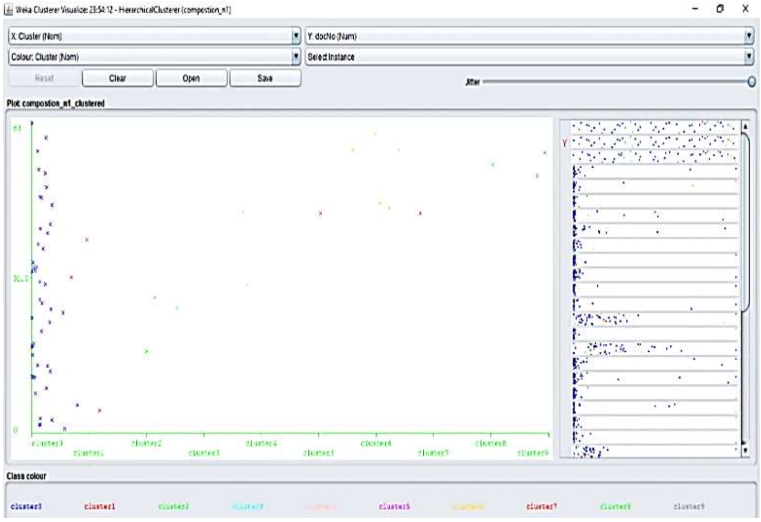
Y: docNo (Num)

Colour: Cluster (Nom)

Select Instance

Plot: compostion_n1_clustered

Fig. 7. Complete-link type Hierarchical clustering


Weka Clusterer Visualize: 23:54:12 - HierarchicalClusterer (compostion_n1)

X: Cluster (Nom)

Y: docNo (Num)

Colour: Cluster (Nom)

Select Instance

Plot: compostion_n1_clustered

Fig. 8. Average-link type Hierarchical clustering

V. CONCLUSION

From the experiments, it has been concluded that the clusters are made on the basis of the distribution of topics over which the documents are close to each other in terms of their meanings and the probability value is same for Complete-Link Type and Average-Link Type clustering for each topic in each document. Less-fold cross validation might be the reason for the same values in the probability values. In contrast to this, Single- Link type has very low probability for topics. As Topic modeling is an emerging and current field, we can increase the cross validation on dataset to have more accuracy and we can also approach different methods to improve the performance of the models for managing large archives of information. The futuristic directions for further research can Fuzzy logics and the Dynamic data-set. We can perform a hybrid approach by combining LDA and K-Means clustering on a dynamic dataset instead of static. In the Fuzzy Logic approach the simple LDA topics can be considered in a crisp set. Hence, we can fuzzify the output in results; and further the recovered topics can be in the form of a fuzzy set.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, no. 4–5, pp. 993–1022, 2003, doi: 10.1016/b978-0-12-411519-4.00006-9.

[2] M. Griffiths, T.L.; Steyvers, “‘Finding scientific topics,’” in Proceedings of the National Academy of Sciences, 2004, pp. 5228--35.

[3] D. M. Blei, “Probabilistic Topic Models,” pp. 77–84.

[4] D. Blei, L. Carin, and D. Dunson, “and applications to document and image analysis] [A focus on graphical model design,” no. November, pp. 55–65, 2010.

[5] J. Chien, S. Member, and C. Chueh, “Topic-Based Hierarchical Segmentation,” vol. 20, no. 1, pp. 55–66, 2012.

[6] J. Chien, S. Member, C. Chueh, and S. Member, “Dirichlet Class Language Models for Speech Recognition,” vol. 19, no. 3, pp. 482–495, 2011.

[7] J. Chien and M. Wu, “Adaptive Bayesian Latent Semantic Analysis,” vol. 16, no. 1, pp. 198–207, 2008.

[8] Y. Chang and J. Chien, “LATENT DIRICHLET LEARNING FOR DOCUMENT SUMMARIZATION,” pp. 1689–1692, 2009.

[9] N. Chiao, “HIERARCHICAL THEME AND TOPIC MODEL FOR SUMMARIZATION Jen-Tzung Chien and Ying-Lan Chang Department of Electrical and Computer Engineering,” 2013.

[10] E. P. Xie, P.; Xing, “‘Integrating Document Clustering and Topic Modeling,’” 2013.

[11] W. H. Li Y., “A clustering Method Based on K- Means Algorithm,” in 2012 International Conference on Solid State Devices and Materials, Science Direct: Physics Procedia 25, pp. 1104–1109.

[12] Z. Guan, P.; Wang, M.; Chen, B.; Fu, “‘K-means Document Clustering Based on Latent Dirichlet Allocation,’” Forty-Fifth Annu. Meet. West. Decis. Sci. Inst., 2016.

[13] M. Chemudugunta C.; Smyth P.; Steyvers, “‘Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model,’” Adv. neural Inf. Process. Syst., 2006.

[14] O. D.A., “‘Using Latent Dirichlet Allocation for Topic Modeling in Twitter,’” in In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, CA, USA, 2015, pp. 978-1-4799-7935–6.

[15] L. J. Z. D. Chen H., Zhang G., “A Fuzzy Approach for measuring development of topics in patents using Latent Dirichlet Allocation,” in In: 2015 IEEE International Conference on Fuzzy Systems, 2015, p. ISBN: 978-1-4673-7428-6., doi: 10.1007/s10489-014-0595-0.

[16] Z. Yang, Q. Li, Z. Lu, Y. Ma, Z. Gong, and H. Pan, “Semi-supervised Multimodal Clustering Algorithm Integrating Label Signals for Social Event Detection,” Proc. - 2015 IEEE Int. Conf. Multimed. Big Data, BigMM 2015, pp. 32–39, 2015, doi: 10.1109/BigMM.2015.26.

[17] H. Zhou, H. Yin, H. Zheng, and Y. Li, “A survey on multi-modal social event detection,” Knowledge-Based Syst., vol. 195, p. 105695, 2020, doi: 10.1016/j.knosys.2020.105695.

[18] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, “An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit,” Inf. Process. Manag., vol. 57, no. 2, 2020, doi: 10.1016/j.ipm.2019.04.002.

[19] S. A. Devi and S. S. Kumar, “A hybrid document features extraction with clustering based classification framework on large document sets,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 7, pp. 364–374, 2020, doi: 10.14569/IJACSA.2020.0110748.