International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Analyzing Educational Tweets using LDA Model

Sundravadivelu Kamatchi *[1], Dr. Thangaraj Muthuraman[2]

Submitted: 10/09/2022 **Accepted:** 20/12/2022

Abstract: For the purpose of generating best educational reforms, knowledge discovery of educational tweet analysis is more important. Today the social media content on the Internet is rigorously increasing hour by hour. Hence analyzing this textual content is a vital task to solve problems in education. In this study Latent Dirichlet Allocation (LDA) is used to analyze the text content which finds the relationships among documents in the corpus. This proposed work shows that the LDA provide better result to extract topic with accurate coherence & prevalence score. This work also infers that the LDA performs best than Latent Semantic Analysis (LSA).

Keywords: Topic modeling, Latent Dirichlet Allocation, Latent Semantic Analysis, Educational tweets and etc.

1. Introduction

Due to the vast amount of information generation in social media, require effective searching & management & analysis of text data is one of the major interest in researchers [1]. Topic modeling is under the branch of unsupervised machine learning [2]. It has an ability to scan a set of documents from a corpus, analyze the word and phrases the similar word groups into clusters [3].

Traditional methods like k-means are well opted for clustering. But in social media most of the data are in textual format [4]. During the process of clustering text documents, there is an overlapping arises among documents [5]. Therefore this framework uses topic modeling. Topic modeling consists of group of algorithms which are developed to identify the hidden topics with in a document [6].

A document is a collection of topic. Topic model are used to find the hidden themes from the collection then annotate the word phrases with respect to the themes [7]. Each and every word phrase is represented by these topics. This process generates a distribution of topics with document coverage which is mainly utilized to explore investigate the data which correlation to model [8]. This is applied to drawn the group of latent topics among document in a corpus [9]. One of the leadings sources of information is captured from twitter in the form of tweets which is highly suitable to analyze the sentiments [10].

There are two basic types of topic modeling one is linear modeling and another one is Probabilistic modeling [11]. Linear modeling uses inverse document frequency to analyze the word phrases [12]. One example of linear modeling is Latent Semantic Analysis (LSA) which is applied to draw similar topics but it needs large corpus to produce precise results. It is because of their inefficient representation [13]. Probabilistic model is developed to solve the issues of LSA by the use of applying probabilistic functions [14]. LDA is the Bayesian version of probabilistic model [15].

[1]*Research Scholar, Department of Computer Science, School of Information Technology, Madurai Kamaraj University,Madurai,Tamil Nadu, India-625 021.*

[2] *Professor, Department of Computer Science, School of Information Technology, Madurai Kamaraj University,Madurai, India-625 021.*

**Corresponding Author Email: svadiveluk2021@gmail.com*

2. Literature Study

This research study [16] collects tweets from Surabaya citizen. Then analyze the tweets with respect to two algorithms LDA and LSA. Finally this study concludes that the LDA provides better results than LSA. But this paper compares LDA and LSA corresponds to coherence value only. This work is devoted to government of Surabaya and their media canter.

This paper [17] focuses the challenging events that occur in Kenya. Therefore this work collects tweets from the peoples of Kenya. Then apply the LDA algorithm to evaluate the analysis of Normalized mutual information (NMI) and coherence are used to choose the accurate model and the work concludes that the LDA performs better.

This framework [18] analyzes the educational tweets by the use of deep learning techniques. The educational tweets are collected from student's feedback. This work assures that the MLP returns best results than CNN in terms concerned dataset.

This work [19] developed a live micro blog search engine for twitter. Similar tweets are identified through the use of LDA. Topic modeling is utilized to retrieve and rank the tweets. They do not provide the evaluation results. This research [20] analyzes the tweets regarding the reactions about COVID-19 in Canada. LDA is used for topic modeling and the ABSA is used for sentimental analysis. Health experts are also devoted to this work for interpreting the results.

3. Proposed Work

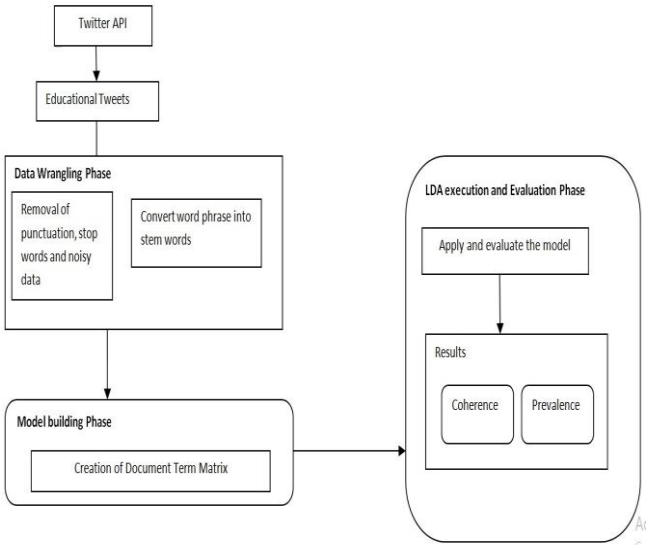
This research work comprises three components,

3.1 Data wrangling phase

3.2 Model building phase

3.3 LDA Execution and evaluation phase.

The architectural framework of the model is depicted in **Fig.1.**



These three phases are expressed in detail in the following sections.

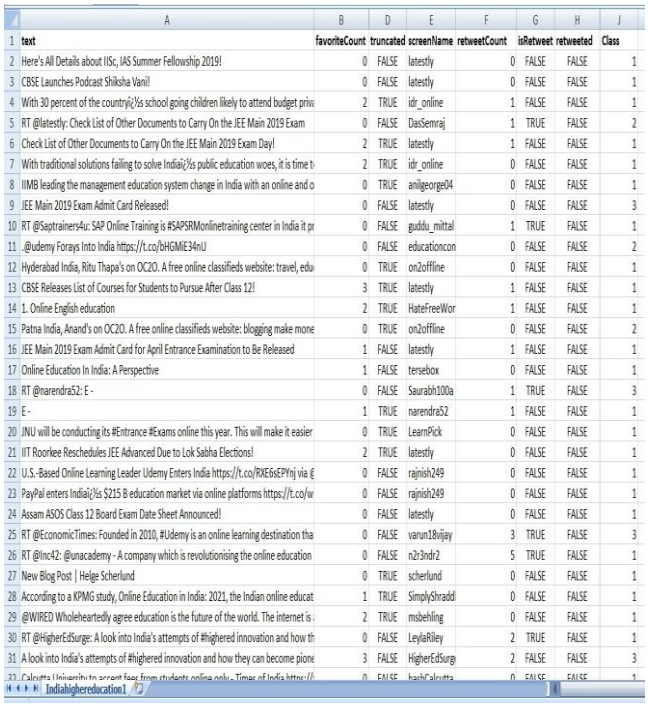
3.1 Data wrangling phase

The tweets related to higher education are retrieved from twitter API V2.0. Data wrangling is the process of cleaning and managing the huge volume of data which is necessary for the analysis of textual data. The collected tweets contain much irrelevant information such as punctuation, stop words, twitter handles and hash tags. These are cleaned, after cleaning the dataset consists of three classes (positive, Negative and Neutral) with respect to the ranks of polarity. Then convert the word phrase into stem words which is the process of transforming the word phrase with corresponds to its base form. After data wrangling the parameters of the higher education dataset are described in table I.

Table I: Parameters of the Higher Education Dataset

Parameter name	Data type
Id	Number
Text	Nominal
Favorite count	Number 0/1/2
Screen name	Nominal
Retweet count	Number
Is retweet	Boolean
retweeted	Boolean
Class	Positive/Negative/Neutral

The following Fig.2 shows the visualization of the higher education dataset in CSV format. It consists of 5000 instances with eight attributes.



	A	B	D	E	F	G	H	J
1	text	favoriteCount	truncated	screenName	retweetCount	isRetweet	retweeted	Class
2	Here's All Details about IISc, IAS Summer Fellowship 2019!	0	FALSE	latestly	0	FALSE	FALSE	1
3	CBSE Launches Podcast Shiksha Vani!	0	FALSE	latestly	0	FALSE	FALSE	1
4	With 30 percent of the countryï¿½s school going children likely to attend budget privat	2	TRUE	idr_online	1	FALSE	FALSE	1
5	RT @latestly: Check List of Other Documents to Carry On the JEE Main 2019 Exam	0	FALSE	DasSemraj	1	TRUE	FALSE	2
6	Check List of Other Documents to Carry On the JEE Main 2019 Exam Day!	2	TRUE	latestly	1	FALSE	FALSE	1
7	With traditional solutions failing to solve Indiaï¿½s public education woes, it is time t	2	TRUE	idr_online	0	FALSE	FALSE	1
8	IIMB leading the management education system change in India with an online and o	0	TRUE	anilgeorge04	0	FALSE	FALSE	1
9	JEE Main 2019 Exam Admit Card Released!	0	FALSE	latestly	0	FALSE	FALSE	3
10	RT @Saptrainers4u: SAP Online Training is #SAPSRMonlinetraining center in India it pr	0	FALSE	guddu_mittal	1	TRUE	FALSE	1
11	.@udemy Forays Into India https://t.co/bHGMiE34nU	0	FALSE	educationcon	0	FALSE	FALSE	2
12	Hyderabad India, Ritu Thapa's on OC20. A free online classifieds website: travel, edu	0	TRUE	on2offline	0	FALSE	FALSE	1
13	CBSE Releases List of Courses for Students to Pursue After Class 12!	3	TRUE	latestly	1	FALSE	FALSE	1
14	1. Online English education	2	TRUE	HateFreeWor	1	FALSE	FALSE	1
15	Patna India, Anand's on OC20. A free online classifieds website: blogging make mone	0	TRUE	on2offline	0	FALSE	FALSE	2
16	JEE Main 2019 Exam Admit Card for April Entrance Examination to Be Released	1	FALSE	latestly	1	FALSE	FALSE	1
17	Online Education In India: A Perspective	1	FALSE	tersebox	0	FALSE	FALSE	1
18	RT @narendra52: E -	0	FALSE	Saurabh100a	1	TRUE	FALSE	3
19	E -	1	TRUE	narendra52	1	FALSE	FALSE	1
20	JNU will be conducting its #Entrance #Exams online this year. This will make it easier	0	TRUE	LearnPick	0	FALSE	FALSE	1
21	IIT Roorkee Reschedules JEE Advanced Due to Lok Sabha Elections!	2	TRUE	latestly	0	FALSE	FALSE	1
22	U.S.-Based Online Learning Leader Udemy Enters India https://t.co/RXE6sEPYnj via @	0	FALSE	rajnish249	0	FALSE	FALSE	1
23	PayPal enters Indiaï¿½s $215 B education market via online platforms https://t.co/w	0	FALSE	rajnish249	0	FALSE	FALSE	1
24	Assam ASOS Class 12 Board Exam Date Sheet Announced!	0	FALSE	latestly	0	FALSE	FALSE	1
25	RT @EconomicTimes: Founded in 2010, #Udemy is an online learning destination tha	0	FALSE	varun18vijay	3	TRUE	FALSE	3
26	RT @Inc42: @unacademy - A company which is revolutionising the online education	0	FALSE	n2r3ndr2	5	TRUE	FALSE	1
27	New Blog Post | Helge Scherlund	0	TRUE	scherlund	0	FALSE	FALSE	1
28	According to a KPMG study, Online Education in India: 2021, the Indian online educat	1	TRUE	SimplyShraddl	0	FALSE	FALSE	1
29	@WIRED Wholeheartedly agree education is the future of the world. The internet is	2	TRUE	msbehling	0	FALSE	FALSE	1
30	RT @HigherEdSurge: A look into India's attempts of #highered innovation and how th	0	FALSE	LeylaRiley	2	TRUE	FALSE	1
31	A look into India's attempts of #highered innovation and how they can become pione	3	FALSE	HigherEdSurge	2	FALSE	FALSE	3
32	Calcutta University to accept fees from students online only - Times of India https://t	0	FALSE	bashCalcutta	0	FALSE	FALSE	1

IndiahighereducationT

Fig.2. visualization of the higher education dataset in CSV format.

3.2 Model building phase

The first step of model building is to construct a word cloud which displays the most frequent words. The next step is to build the Document Term Matrix (DTM). It is a matrix with terms is in columns and tweets are in rows. If the word phrase appears in the document then its value is indicated as 1 otherwise 0.

In order to evaluate the model performance, the log-likelihood of the model is calculated. The result assures that the log-likelihood per word is inspected to be good. The log-likelihood of the model in 500 Iterations are presented in Fig.3. It shows the likelihood of topics in iteration wise.

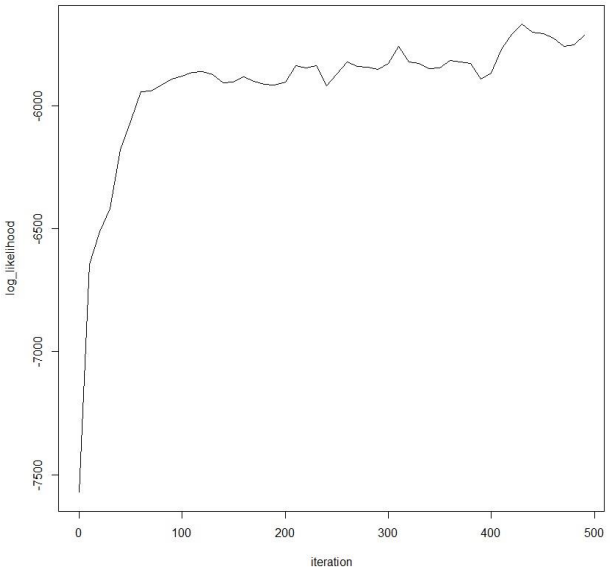
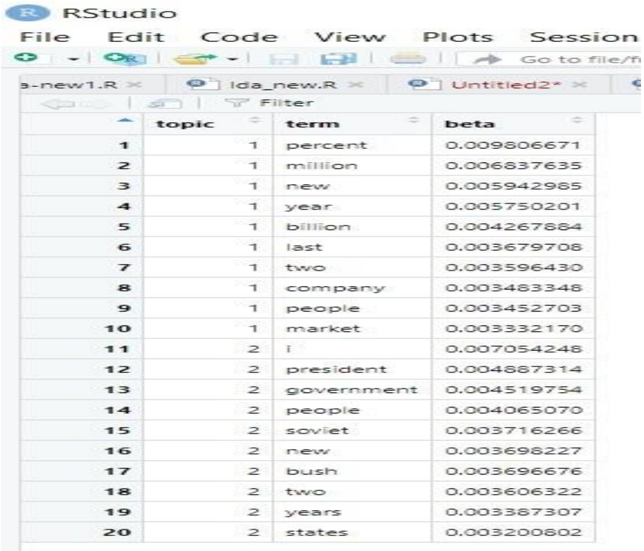


Fig.3. Log-likelihood of the model in 500 iterations.

3.3 LDA Execution and Evaluation phase

The main future of LDA is to understand the relationship between word phrases by the use of topics. The term topic is used to relating a word with a definition. For example if the machine reads: cow is white, the LDA first tokenizes this into two topics cow which is an animal and white is a color. The steps of LDA algorithm is given below,

Step 1: Determine the number of words in the document.
Step 2: Topic mixture is created regarding to the collection of topics.
Step 3: Based on the document's multinomial distribution, select the topics from topic mixture.
Step 4: Picked the corresponding words with respect to topics.
Step 5: Calculate the probability of the topic t in the document d and probability of the word appeared on the topic.
Step 6: Based on the probability value, topics are updated.
The top 20 terms with its beta value are given in Fig.4. Beta value corresponds to the density of the topic word. High beta value indicates that the most of the topics in the corpus are represented correctly.

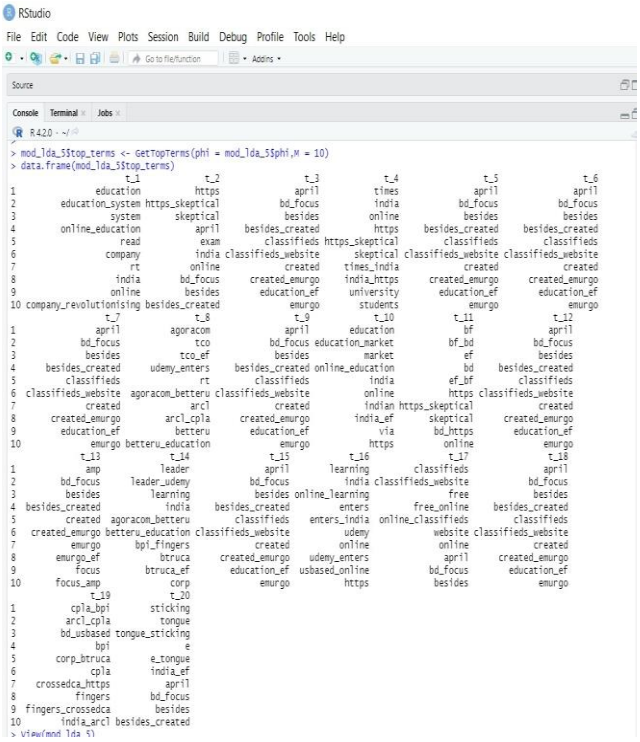


	topic	term	beta
1	1	percent	0.009806671
2	1	million	0.006837635
3	1	new	0.005942985
4	1	year	0.005750201
5	1	billion	0.004267884
6	1	last	0.003679708
7	1	two	0.003596430
8	1	company	0.003483348
9	1	people	0.003452703
10	1	market	0.003332170
11	2	i	0.007054248
12	2	president	0.004887314
13	2	government	0.004519754
14	2	people	0.004065070
15	2	soviet	0.003716266
16	2	new	0.003698227
17	2	bush	0.003696676
18	2	two	0.003606322
19	2	years	0.003387307
20	2	states	0.003200802

Fig.4. Top 20 terms with its beta value.

The screenshot of 20 topics with their terms are shown in fig.5. The following fig.6 displays the topic modeling with rating 5 selected from the top terms. By measuring the percentage of semantic similarity of top terms, the coherence score is calculated. How the words are related on a topic is described by the coherence which is represented as below,

P (b|a) – P (b) here {a,b} are the pair of words. The best topics with its coherence score is presented in fig.7.



```
> mod_lda_5$top_terms <- GetTopTerms(phi = mod_lda_5$phi,M = 10)
> data.frame(mod_lda_5$top_terms)
                        t_1             t_2               t_3             t_4               t_5               t_6
1                 education           https             april           times             april             april
2          education_system https_skeptical          bd_focus           india          bd_focus          bd_focus
3                    system       skeptical           besides          online           besides           besides
4          online_education           april   besides_created           https   besides_created   besides_created
5                      read            exam       classifieds https_skeptical       classifieds       classifieds
6                   company           india classifieds_website      skeptical classifieds_website classifieds_website
7                        rt          online           created     times_india           created           created
8                     india        bd_focus    created_emurgo     india_https    created_emurgo    created_emurgo
9                    online         besides      education_ef      university      education_ef      education_ef
10 company_revolutionising besides_created            emurgo        students            emurgo            emurgo
                  t_7               t_8                 t_9             t_10            t_11                t_12
1               april          agoracom               april        education              bf               april
2            bd_focus               tco            bd_focus education_market           bf_bd            bd_focus
3             besides            tco_ef             besides           market              ef             besides
4     besides_created      udemy_enters     besides_created online_education              bd     besides_created
5         classifieds                rt         classifieds            india           ef_bf         classifieds
6 classifieds_website  agoracom_betteru classifieds_website           online           https classifieds_website
7             created              arcl             created           indian https_skeptical             created
8      created_emurgo         arcl_cpla      created_emurgo         india_ef       skeptical      created_emurgo
9        education_ef           betteru        education_ef              via       bd_https        education_ef
10             emurgo betteru_education              emurgo            https          online              emurgo
              t_13              t_14                t_15            t_16                t_17                t_18
1              amp            leader               april        learning         classifieds               april
2         bd_focus      leader_udemy            bd_focus           india classifieds_website            bd_focus
3          besides          learning             besides online_learning                free             besides
4  besides_created             india     besides_created          enters         free_online     besides_created
5          created  agoracom_betteru         classifieds    enters_india online_classifieds         classifieds
6   created_emurgo betteru_education classifieds_website           udemy             website classifieds_website
7           emurgo       bpi_fingers             created          online              online             created
8        emurgo_ef            btruca      created_emurgo    udemy_enters               april      created_emurgo
9            focus         btruca_ef        education_ef  usbased_online            bd_focus        education_ef
10       focus_amp              corp              emurgo           https             besides              emurgo
                t_19            t_20
1           cpla_bpi        sticking
2          arcl_cpla          tongue
3         bd_usbased tongue_sticking
4                bpi               e
5        corp_btruca        e_tongue
6               cpla        india_ef
7      crossedca_https           april
8            fingers        bd_focus
9  fingers_crossedca         besides
10        india_arcl besides_created
> View(mod_lda_5)
```

Fig.5. Top 20 topics with their terms.

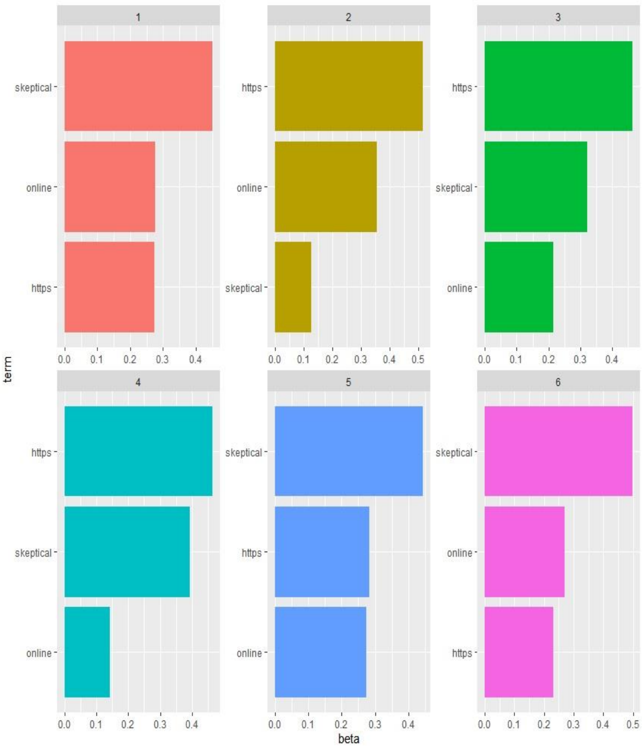


Fig.6. Topic modeling with rating 5.

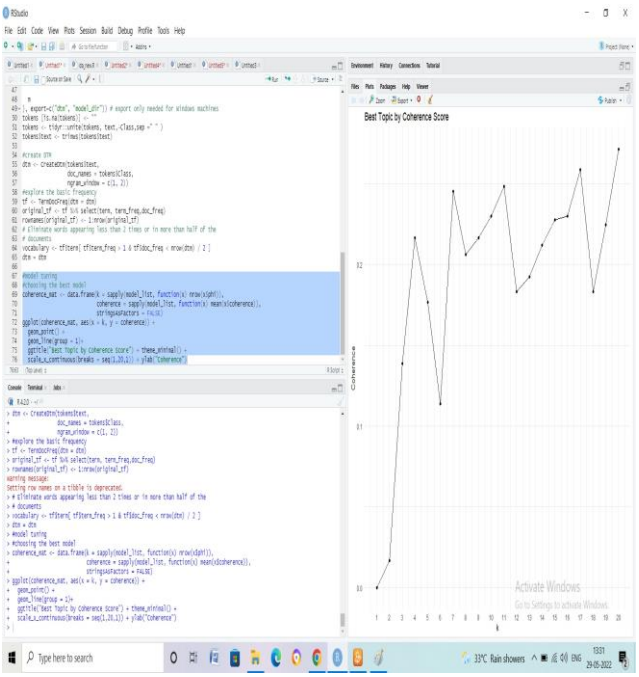


Fig.7. Best topics with its coherence score.

It shows that the topic 17 has the high coherence value that means the words in the topic are highly associated to each other. It is best way to group the topics using dendogram that shows how the topics are closely related. The result of the dendogram for this model is presented in fig.8.

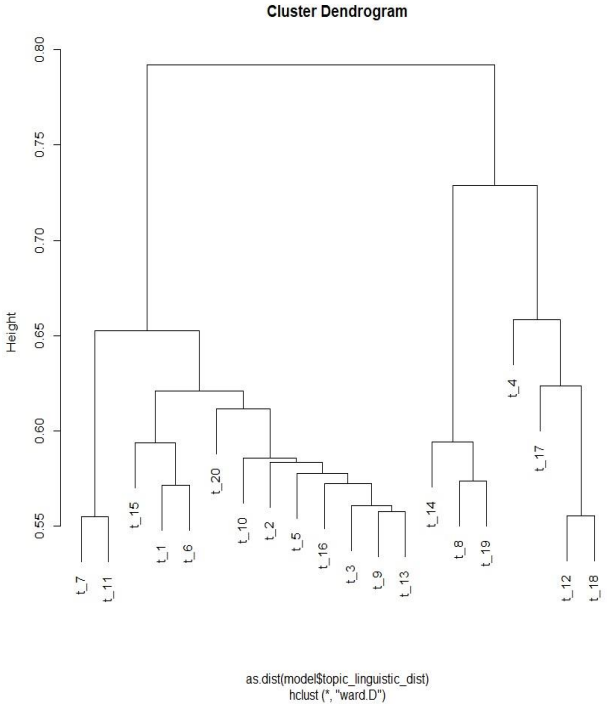


Fig.8. Dendogram of the model.

It shows that the topic t7 and t11 have high similarity to each other.

Prevalence is a score which explains the most frequent topics that is, probability of topics in the entire document. Fig. 9 illustrates both the coherence and prevalence score of the top topics. Based on coherence topic 17 has high value but in prevalence topic 11 has high score and topic 17 has the low value, which means that the tough words inside the topic 17 are also supporting each other.

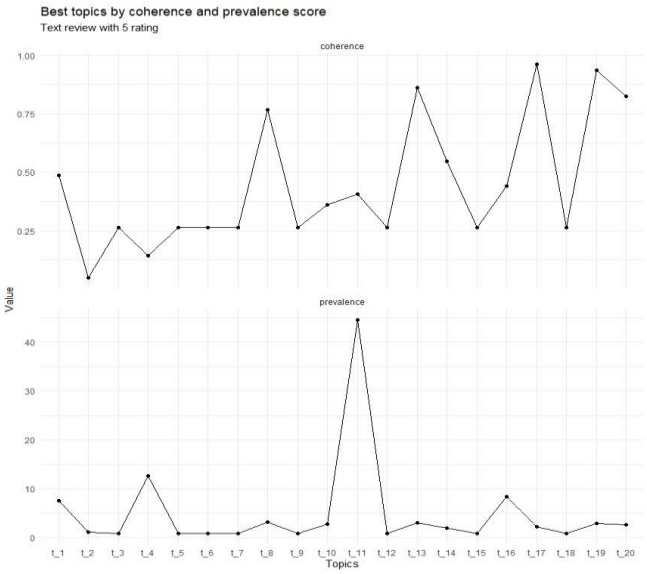


Fig.9. Coherence and prevalence of the top 20 terms.

4. Experimental Result

In order to compare LDA and LSA the first 10 topic coherence that are derived by both modeling is calculated which is indicated in fig.10. It clearly defines that the number of best topic is 6 for LDA and 2 for LSA. Coherence score greater than 0.75 is considered to be high. It concludes LDA bring off good results than LSA.

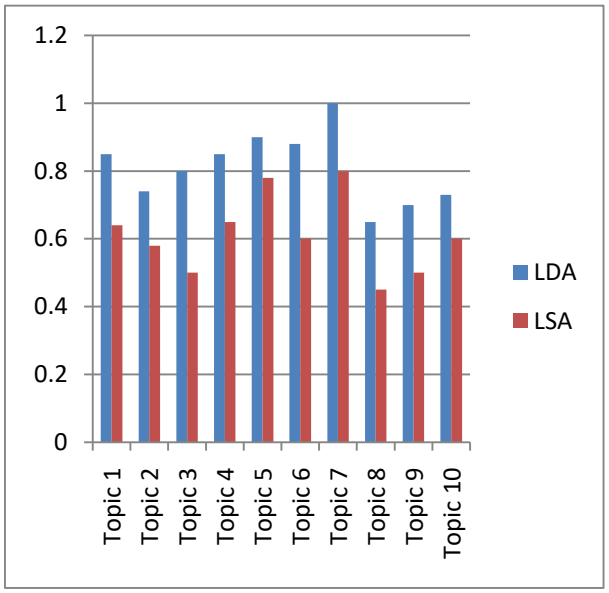


Fig.10. Comparison of LDA and LSA using 10 Topics.

5. Conclusion and Future Work

After implementing the framework LDA works effectively to modeling topics in a corpus as well as topic modeling with respect to tweets. LDA construct topics that are both coherent and consistent with regarding to the determined clusters that are thus far presented in the tweet data. In future the proposed framework would be enhanced to analyze the relation between user feedback

and coherence of the topics drawn from the framework which creates more understanding among the topics.

6. References

[1]. Aletras, N., Baldwin and et.al,"Evaluating Topic Representations for Exploring Document Collections", Journal of the Association for Information Science and Technology, 2015.

[2]. Nugroho, R and D. Molla-Aliod et.al, "Incorporating Tweet Relationships into Topic Derivation", Proceedings of the 2015 Conference of the Pacific Association for Computational Linguistics, PACLING. 2015

[3]. Bettina Grun, kurt Hornik ,"topicmodels: An R Package for Fitting Topic Model", Journal of Statistical Software, 2011, 40(13).

[4]. Goel, Vikas, Amit Kr Gupta, and Narendra Kumar, "Sentiment Analysis of Multilingual Twitter Data using Natural Language Processing." 2018 8th International Conference on Communication Systems and Network Technologies (CSNT). IEEE, 2018.

[5]. L. Hong and B. D. Davison,"Empirical study of topic modelling in twitter", In Proceedings of the First Workshop on Social Media

[6]. Kumari, S. S. ., and K. S. . Rani. "Big Data Classification of Ultrasound Doppler Scan Images Using a Decision Tree Classifier Based on Maximally Stable Region Feature Points". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 8, Aug. 2022, pp. 76-87, doi:10.17762/ijritcc.v10i8.5679.

[7]. Ponweiser, M., Latent Dirichlet Allocation in R. Diploma Thesis Institute for Statistics and Mathematics, 1-138, (2012).

[8]. Qi Jing," Searching for Economic Effects of User Specified Event Based on Topic Modeling and Event Reference", Jordery School of Computer Science, Acadia University 2015.

[9]. David M.Blei , "Probabilistic Topic Models", Communications of the ACM, 2012, 55(4), 77-84.

[10]. David M.Blei and John D. Lafferty, "A Correlated Topic Model of Science", Annals of Applied Statistics 2006, 1(1), 17-35.

[11]. L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on Streaming document collections", In Proceedings of the 15th ACM SIGKDD international conference on KDD, KDD '09, 2009, 937–946, NY, USA, ACM.

[12]. M. Hoffman, D. Blei, and F. Bach,"Online learning for latent dirichlet allocation", Advances in Neural Information Processing Systems,2010, 23,856–864.

[13]. 12. D. M. Blei, A. Y. Ng, and M. I. Jordan,"Latent dirichlet allocation",JMLR, 2003, 3, 993–1022.

[14]. David Mimno and Hanna M Wallach et.al," Optimizing semantic coherence in topic models", In Proceedings of the Conference on Empirical Methods in Natural Language processing, 2011, 262–272. Association for Computational Linguistics.

[15]. Li-Qiang Niu and Xin-Yu Dai."Topic2vec: Learning distributed representations of topics" , 2015 Available at: arXiv preprint arXiv: 1506.08422

[16]. Vivek Kumar and Rangarajan Sridhar, "Unsupervised topic modeling for short texts using Distributed representations of words", In Proceedings of NAACL-HLT, 2015, 192–20

[17]. Siti Qomariyah, Nur Iriawan and Kartika Fithriasari, "Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis", AIP Conference Proceedings 2194, 020093 (2019); https://doi.org/10.1063/1.5139825 Published online: 18 December 2019.

[18]. Marina Sokolova1, Kanyi Huang1 and Stan Matwin and et.al," Topic Modeling and Event Identification from Twitter Textual Data",

[19]. G. Bala Krishna Priya 1], Dr. Jabeen Sultana 2. , Prof. M. Usha Rani 3," A Review to Classify Sentiments Using Some Machine Learning Techniques", International Journal of Computer Science Trends and Technology (IJCST), 2021, 9(4), 18-22.

[20]. Garg, D. K. . (2022). Understanding the Purpose of Object Detection, Models to Detect Objects, Application Use and Benefits. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(2), 01–04. https://doi.org/10.17762/ijfrcsce.v8i2.2066

[21]. Jabeen Sultana, M. Usha Rani and M.A.H. Farquad," Knowledge Discovery from Recommender Systems using Deep Learning", 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), 1074-1078. IEEE.

[22]. Kose, O., & Oktay, T. (2022). Hexarotor Yaw Flight Control with SPSA, PID Algorithm and Morphing. International Journal of Intelligent Systems and Applications in Engineering, 10(2), 216–221. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/1879

[23]. Christan Grant, Clint P. George, Chris Jenneisch," Online Topic Modeling for Real-time Twitter Search".