International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-6, August 2020

Comprehensive Analysis of Variants of TF-IDF Applied on LDA and LSA Topic Modelling

S. Sai Manasa Bala, Santoshi Kumari

Abstract— Present generation is fully connected virtually through many sources of social media. In social media, opinions of people for any post, news or about any product through comments or emoticon designed to express the satisfactory note. Market standards improve on this basis. There are different online markets like Amazon, Flipkart, Myntra improve their businesses using these reviews passed. Analyzing large scale opinion or feedback of individual's helps to identify hidden insights and work towards customer satisfaction. This paper proposes for applying different weighting scheme of TF-IDF (Term Frequency-Inverse Document Frequency) for topic modeling methods LSA and LDA to cluster the topics of discussion from large scale reviews related to booming online market 'Amazon'. The main focus of the paper is to observe the changes in the topic modeling by applying different weighting schemes of TF-IDF. In this work topic-based models like LDA (Latent Dirichlet Allocation) and LSA (Latent Semantic Allocation) applied to various weighting schemes of TF-IDF and observed the changes of weights leads to variation of term frequency of different topics with respect to its documents. Results also show that the variation of term weights results changes in topic modeling. Visualization results of topic modeling clusters with different TF-IDF weighting schemes are presented.

***Keywords*: Data Analysis, LDA, LSA, TF-IDF Weights, Topic Modeling**

I. INTRODUCTION

Word representations are one of the critical tasks in the field of natural language processing. It is an obvious representation of words as indices of vocabulary, but this procedure fails to extract the abundant relational structure of the lexicon. Vector-space model perform better in this prospect, which encodes continuous similarities between words as measure of distance or as a measure of angle between different words for a high dimensional space. There are very general approaches that have proved helpful for the jobs like named entity recognition, POS tagging, word sense disambiguation and document retrieval [1]. Text classification has been intensively researched from the last 15 years.[6] Numerous of classifiers have been developed to extract better results in classification of data[13]. Different models of documents have been proposed for the betterment of representing the documents. This project we study different topic based modelling like LDA (Latent Dirichlet Allocation) and LSA (Latent Semantic Allocation) also named as LSI (Latent Semantic Indexing)[11],[14]. The most popular methodology used to represent each document as the weighting vector in real numbers is the TF-IDF weighting scheme, that is applied to each words in the document. The TF is term frequency that counts the number of times the particular word has repeated in the total number of documents where as IDF is Inverse document frequency measure for importance of document in the corpus. The actual motive in LDA is to present each document as a mixture of topics, and learn these topics and words which are produced by each topic for each document. This method can be applied when a large corpus is handled. Also, LSA works in the similar fashion but it learns the topics by performing a matrix decomposition like SVD and it is much faster that LDA. The main target of this paper is to vary different weights of TF-IDF on the corpus and is applied to LDA and LSA topic models. The variations of weighting schemes is proved to affect the results in the visualization of LDA models as well as in the LSA models. Also, a keen observation towards the relevance term shows the variations in the probabilities of the term in that topic will also vary.

LDA is an unsupervised learning and a probabilistic document model that assumes each document as a mixture of latent topics. For a single latent topic T, the model learns the conditional distribution $p(w|T)$ means the probability that the word w occurs in the topic T. We can represent i dimensional vector representation of words by first training a i- topic model and then fill the matrix with the values of $p(w|T)$. The corpus which has the complete set of documents is verified for multiple times and then gives a end result with consistent topics [14]. In the following sections: literature survey is discussed in section 2, proposed system is explained in section 3 and section 4 discusses on Experimental results and analysis and conclusion in the last part.

II. LITERATURE SURVEY

Sentiment analysis is the most important area of NLP where there is the highest possibility of improvisation and scope of learning is enhanced. Many research scholars have worked on different parts segments and different combinations of methodologies to make the machine understand the human sentiment [2].

Revised Manuscript Received on August 15, 2020.
* Correspondence Author
S. Sai Manasa Bala, M.Sc in Applied Mathematics, M S Ramaiah University of Applied Sciences, Bangalore, Inida. Email: manasa.bala369@gmail.com
Santoshi Kumari*, Department of Computer Science and Engineering, M S Ramaiah University of Applied Sciences, Bangalore, Inida. Email: santoshik29@gmail.com

This paper also works on the same perspective, but tries to perform a comparative study over different models and its combinations to analyze the effectiveness of the performance level. Text classification has been intensively researched from the last 15 years [3].

Numerous of classifiers have been developed to extract better results in classification of data [4].Different models of documents have been proposed for the betterment of representing the documents.

Different topic- based modelling like LDA (Latent Dirichlet Allocation) and LSA (Latent Semantic Allocation) also named as LSI (Latent Semantic Indexing [3]–[5] are introduced. As we know the most popular methodology used to represent each document as the weighting vector in real numbers is the TF-IDF [5] weighting scheme applied to each word in the document.

LDA is an unsupervised learning and a probabilistic document model that assumes each document as a mixture of latent topics. For a single latent topic *T*, the model learns the conditional distribution *p(w|T)* means the probability that the word w occurs in the topic *T*. We can represent *i* dimensional vector representation of words by first training a *i*- topic model and then fill the matrix with the values of *p(w|T)* The corpus which has the complete set of documents is verified for multiple times and then gives an end result with consistent topics[7].

Some other Vector space models (VSM) require to model the words directly [8]. Here, is the LSA popular for VSM, which explicitly learns semantic word vectors using SVD (singular value decomposition) to factor the term document co-occurrence matrix. To retrieve *i* dimensional representation for a given word, the entries corresponding to *i* largest singular values are only taken from the basis of word's in the form of factored matrix. Such matrix factorizations are highly successful and pave way to the researchers to make a number of choices in design like weighting, normalization, dimensionality reduction algorithm etc. [2] marks a point that by using the term frequency and inverse document frequency weighting schemes transforms the values of VSM and always increases the performance of retrieval and categorization [3] systems. Papers by "Martineau and Finin" show proof that the weighting helps in the sentiment classification, and "Paltoglou and Thelwall" [8] methodically scrutinize a numerous of weighting schemes in the context of sentiment analysis.

This paper, adopts these insights and apply the weighting schemes to different topic model like LDA and LSI [9]. Variants of weighting schemes are applied separately to the term frequency and inverse document frequency, taken in different combinations and are applied to these two topic models. The results have been showed different for different variants and are proved to be better than using LDA or LSI alone with basic TF-IDF weights [6].

III. PROPOSED SYSTEM

The proposed system focus is to develop, implement an automated topic modeling technique LDA and LSA in order to identify the topic of interest from the large-scale customer review data of India for the online trending market Amazon. Examining the posts of different people is a highly challenging task. So, making use of LDA and LSA like topic models can interpret the opinions of numerous people's opinions at once and especially the usage of pyLDAviz [10] in python helps to create a notebook which is user friendly interface that shows the number of topics containing that word and the probability of that topic containing in the documents that makes easy to interpret the classification of data according to the sentiment. The proposed system is implemented in the following steps:

A.Data Collection:

The data collected is the customer reviews on the products bought and the experience of shopping with the online market Amazon. The number of documents taken are the reviews provided by each customer in the form of text. It is a structured form of data [11]. The total number of customer reviews collected are 28333 each which 56666 data points are present.

B. Sentiment Analysis:

In this step an unsupervised lexicon approach is used for the classification of the reviews on the basis of polarity scores.

The sentiment or polarity scores are collected by a inbuilt library nltk (natural language tool kit) the function TextBlob [11] is applied. These interpretations can be represented graphically which makes an easy way to understand the data.

C.Create Term Document matrix TF-IDF

Consider a corpus and dictionary of terms that should contain all the words in corpus and a document term matrix is created. This contains a 2-D matrix whose rows are taken as documents and columns are terms in dictionary. In this for each entry of *(m, n)* the matrix represents the frequency of term as m for the respected document n. The TF-IDF [12] represents the irrelevance of the word for the document and relative document's importance among overall corpus.

D.Topic Modeling

***1)* Calculating LDA**

LDA is an unsupervised learning and a probabilistic document model that assumes each document as a mixture of latent topics. For a single latent topic $\mathbf{T}$, the model learns the conditional distribution $p(w|\mathbf{T})$ means the probability that the word w occurs in the topic $\mathbf{T}$. We can represent a i dimensional vector representation of words by first training a i- topic model and then fill the matrix with the values of $p(w|\mathbf{T})$. The corpus which has the complete set of documents is verified for multiple times and then gives a end result with consistent topics [13].

A document is a sequence of N words that is

$$d = (w_1, w_2, w_3, \dots w_n),$$

$$w_n \textit{ is the nth word of a document}$$

Now, a corpus is the collection of M documents of the form

$C = (d_1, d_2, d_3, \dots d_m)$, d_m is the m[th] document of the corpus

a) Choose a multinomial distribution ϕ_t , for a topic $t \textit{ where } (t \in \{1,2, \dots, T\}$ from a Dirichlet distribution with parameter γ.

b) Choose another multinomial distribution φd for a document such that $(d \in \{d_1, d_2, d_3, \dots d_m\})$ from a Dirichlet distribution with parameter δ.

c) For a word w_n $(n = 1,2, \dots N_d)$ ina document d,

i) Select a topic z_i $\textit{from } \varphi d$

ii) Select a word w_n $\textit{from } \phi t$

The probability of observed data D is computed and maximized using Eq. (1):

$$p(D|\gamma,\delta) = \prod_{d=1}^{M} \int p(t_{d_m}|\delta) * \left(\sum_{n=1}^{N_d} p(Z_{dn}|t_{d_m}) p(w_d|Z_{dm,n},\phi) p(\phi|\gamma)\right) dt_d \phi$$

$i = Number\ of\ topics$
$\delta = parameter\ of\ the\ Dirichlet\ prior\ on\ the\ per-document\ topic\ distribution$
$\gamma = parameter\ of\ the\ Dirichlet\ prior\ on\ the\ per-topic\ word\ distribution$
$t_{d_m} = is\ the\ topic\ distribution\ for\ document\ m$
$\phi_t = is\ the\ topic\ distribution\ of\ topic\ t$
$Z_{dm,n} = is\ the\ topic\ for\ the\ n^{th}\ word\ in\ document\ m$
$w_d = is\ the\ specific\ word$

LDA retrieves the latent topics in whole corpus. To overcome the challenges of interpreting the large scale of data, visualizing the topics can be performed using pyLDAviz in python.

On the other side LSA is computed by the following [14] LSA works with the application of SVD along with the combinations of terms. Applying SVD on TDM (Term-Document Matrix) is what defines LSA. It makes more than mere word co-occurences analysis.

$$M = U \times \sum \times V^T$$

IV. EXPERIMENTAL RESULT ANALYSIS:

A.Data:

The data collected is the customer reviews on the products bought and the experience of shopping with the online market Amazon. The number of documents taken are the reviews provided by each customer in the form of text. It is a structured form of data. The total number of customer reviews collected are 28333. The data extracted consists of 11 columns out of which the column "Customer Review over the Amazon Product for the month of 2019 May" is considered for analysis.

B. Determine TF-IDF

The term document matrix is the vector representation of words present in the documents, where rows represents the words of each document and the columns represent the documents. The values of the matrix are given in the form on 0 or 1 representing the non-existence and existence of the terms in their documents respectively.

C. To Calculate TF-IDF[15]:

Multiplication of TF (a local component) with a IDF (global component), and normalizing the final documents to unit length. Following equation for non-normalized weight of term i in document j in a corpus of D documents is given by Eq. (2)

$$weight_{i,j} = frequency_{i,j} * \log_2 \frac{D}{Document_{frequency_i}} \quad (2)$$

D.Sentiment analysis

The level of sentiment can be calculated by using Lexicon method which assigns the sentiment score for each text and calculates the average of the words in each text and assigns the scores. These scores are further alligned to the requirement as '-1' means negative, '0' means neutral and '1' means positive as shown in Fig.1. Below Fig.2 is the list of sample words that are assigned as negative and positive based on the Lexicon approach and the count of these words throughout the complete texts of customer reviews.

0	My Kindle Fire HD meets my reading and web needs!!
1	We purchased this as a replacement for our now-broken iPad. Though
0	Didn't want to go to the sites amazon wanted me to.
1	Love it and would recommend it for the young as the old
1	It all of a sudden has stoped charging I have only used it a few times in
-1	I have enjoyed the new Kindle,however, was extremely disappointed th
1	For the price, you can't beat this tablet. It is fast and basic.
1	Good features, good cost. Recommend to anyone looking for tablet.

Fig.1 Assigning Sentiment score

Negative Word	Score	Positive Word	Score
hard	274	great	10000
disappointed	266	good	5906
issue	189	easy	3025
dead	181	like	2296
complaints	173	loves	2216

Fig.2 Positive and Negative Sentiment Words and Score

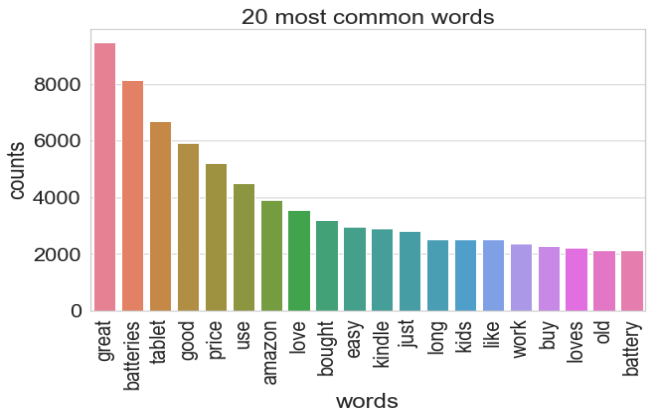

Fig.3 Frequency of Top 20 words

Above Fig.3 represents the frequency of top 20 words in the corpus. Looking at the words and frequency it is identified that positive inclination of reviews in collected dataset.

E.*Applying and visualization LDA*

LDA with the variants of TF-IDF weights is applied for each document term matrix using the LDA model and LSA model function that are inbuilt in gensim models. PyLDAvis is the function used to visualize the results of LDA topics. Here the number of topics taken are $i = 5$.

Analyse os carriedout for this model using 28333 reviews in the form of text data.

Each figure below explains resluting different LDA model after the application of different weighting schemes of TF-IDF [12] and these different models are taken from gensim models from the attribute called 'Smartirs'. These combination of weigting scheme is applied and through which we get different LDA models showing variations in the probability of existence of words with respect to the topic and the document in which the topic exists. The combinations of weights are taken from the gensim library through gensim model the attribute 'smartirs' uses different combinations of weights such as shown in Table 1 and applied to diffrent Topic models .

TABLE 1
TF-IDF weighting Schemes for Topic Models

TF-IDF Model	TF weight	IDF Weight	Document Normalization	Figure (LDA)	Figure (LSA)
Basic model	No. Of terms	No. Of docs	cosine	Fig. 4	Fig. 8
npc	Raw	Probabilistic idf	cosine	Fig. 5	Fig. 9
dpc	Double log	Probabilistic idf	cosine	Fig. 6	Fig. 10
lfc	Logarithm	idf	cosine	Fig. 7	Fig. 11

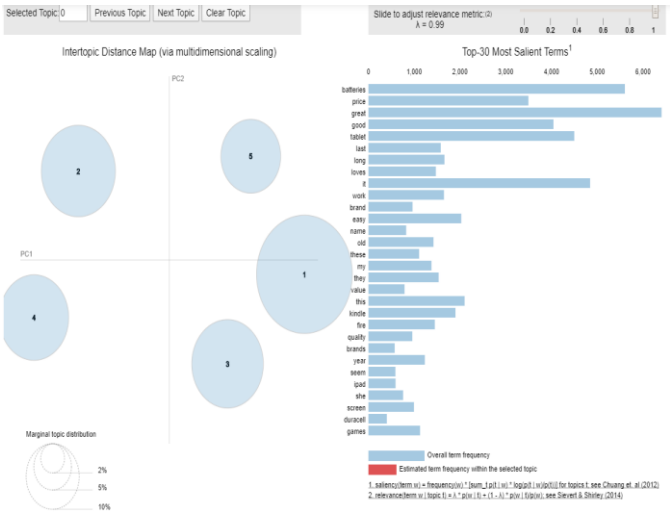

Fig.4. Basic TF-IDF model with LDA visualization with i =5 topics

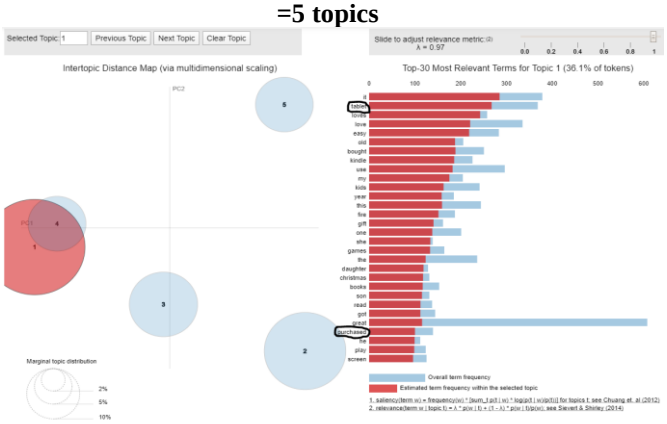

Fig.5. 'npc' combination of weights applied to LDA

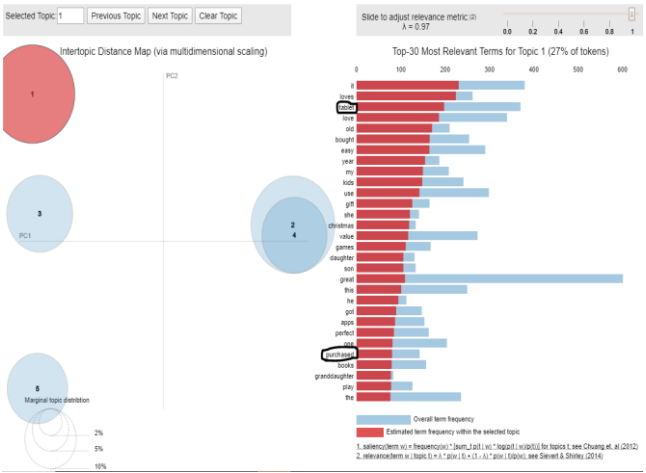

Fig.6 'dpc' combination of weights applied to LDA

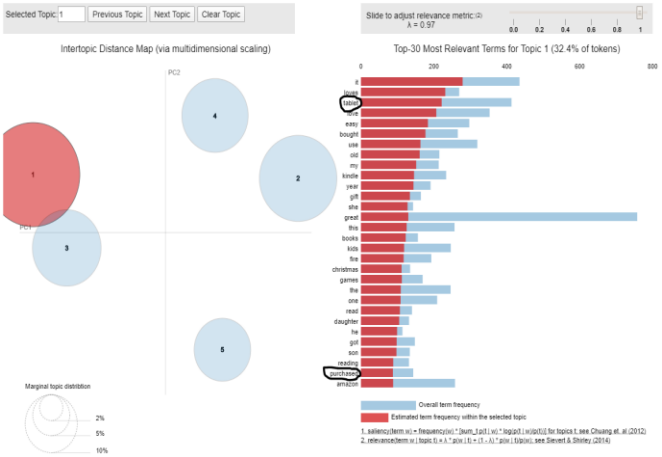

Fig.7 'lfc' combination of weights applied to LDA

Observing the red bar in the above Fig.4 to Fig. 7 for the words which are circled have the variations in the estimated term frequency with the selected topic. Hence this could clearly prove that the variations of the weighting schemes of each term in the document will vary the outcome of the required analysis. Thus, the application of different weights will vary the performance of the topic models. Similarly, is the case with LSA clustering, results are shown as Fig.8 to Fig.11.

```
[(0,
  '0.357*"batteries" + 0.317*"it" + 0.273*"tablet" + 0.262*"great"
0.114*"kindle" + 0.105*"like" + 0.103*"bought" + 0.102*"device" + 0
y" + 0.087*"easy" + 0.086*"last" + 0.085*"get" + 0.084*"time" + 0.0
*"these" + 0.063*"much" + 0.062*"my" + 0.061*"store" + 0.060*"old"
0.053*"brand" + 0.052*"app" + 0.052*"really" + 0.052*"play" + 0.051
 (1,
  '-0.736*"batteries" + 0.309*"tablet" + 0.294*"it" + -0.135*"last"
5*"the" + -0.086*"they" + 0.086*"device" + 0.085*"this" + -0.082*"r
061*"duracell" + 0.061*"loves" + 0.060*"one" + 0.059*"old" + -0.057
*"kids" + -0.051*"brands" + 0.051*"my" + 0.046*"play" + 0.046*"you"
r" + 0.036*"google" + -0.034*"box" + 0.034*"also" + -0.034*"seem" +
```

Fig. 8 Basic TF-IDF applied on LSA and the variation of weights assigned to each term

```
0.100*"batteries" + -0.084*"it" + -0.077*"loves" + 0.07
ght" + -0.049*"so" + -0.046*"love" + -0.046*"year" + -0
"this" + -0.036*"daughter" + -0.036*"one" + -0.036*"qua
d" + -0.025*"gift" + -0.025*"christmas" + -0.025*"he" +
 0.021*"long" + -0.020*"grandson" + -0.020*"size" + -0.

ey" + -0.259*"last" + 0.256*"product" + -0.246*"long" +
29*"seem" + -0.108*"brands" + 0.072*"loves" + 0.067*"ta
-0.049*"duracell" + -0.044*"battery" + -0.044*"cheaper"
 + -0.038*"much" + -0.038*"like" + -0.038*"longer" + 0.
 -0.031*"energizer" + 0.029*"kids" + 0.029*"daughter" +
```

Fig.9 'npc' combination of weights applied on LSA

```
 0.160*"batteries" + 0.160*"tablet" + 0.136*"use" +
" + 0.102*"get" + 0.101*"fire" + 0.101*"battery" +
 0.088*"buy" + 0.088*"they" + 0.087*"long" + 0.085*
ore" + 0.073*"device" + 0.073*"really" + 0.073*"bet
"best" + 0.067*"ve" + 0.064*"that"'),

"tablet" + -0.178*"brand" + -0.177*"long" + -0.176*
117*"games" + -0.116*"work" + 0.114*"apps" + 0.108*
+ 0.094*"my" + -0.093*"price" + 0.092*"she" + 0.084
3*"read" + -0.068*"ve" + -0.068*"energizer" + -0.06
9*"basics" + -0.059*"used" + 0.059*"son" + -0.058*"
```

Fig.10 'dpc' combination of weights applied on LSA

Retrieval Number: D7669049420/2020©BEIESP
DOI: 10.35940/ijeat.D7669.089620
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

```
 0.330*"batteries" + 0.170*"work" + 0.151*"value
et" + 0.093*"it" + 0.092*"well" + 0.087*"works"
 + 0.065*"kids" + 0.062*"far" + 0.062*"this" + 0
.052*"excellent" + 0.049*"nice" + 0.048*"old" +
043*"fire" + 0.042*"so" + 0.041*"best"'),

0.201*"price" + 0.197*"love" + 0.189*"tablet" +
.118*"kindle" + 0.116*"this" + 0.112*"games" + 0
4*"books" + 0.084*"christmas" + 0.081*"apps" + 0
" + 0.064*"like" + 0.063*"grandson" + 0.060*"tim
ze" + 0.052*"works" + 0.052*"we" + 0.051*"set" +
```

Fig.11 'lfc' combination of weights applied on LSA

V. CONCLUSION

This paper shows that by applying different weighting schemes to the words could improve the performance of the topic models by removing the irrelevant terms and rising the weightage of those terms that are more required. This study shows use of only four types of TF-IDF weigting schemes, one normalization method and two topic models. As a part of future work this can be extended by applying diffrent weighting scheme along with various normalization methods. Applied on other topic models such as HDP (Hierarchical Dirichlet Process), ESA (Explicit Semantic Analysis), N-NMF (Non -Negative matrix factorization) and PLSA (Probabilistic LSA) [7] [15] as well as other Natural languge processing tasks such as Sentiment anlysis and opinion mining.

REFERENCES

1. N. Singh and M. Devi, "Document representation techniques and their effect on the document Clustering and Classification: A Review," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 5, pp. 1780–1784, 2017.
2. TextBlob, "TextBlob Documentation," TextBlob, p. 1, 2020.
3. W. Zhang, T. Yoshida, and X. Tang, "TFIDF, LSI and multi-word in information retrieval and text categorization," Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern., pp. 108–113, 2008.
4. D. Gefen, J. E. Endicott, J. E. Fresneda, J. Miller, and K. R. Larsen, "A guide to text analysis with latent semantic analysis in r with annotated code: Studying online reviews and the stack exchange community," Commun. Assoc. Inf. Syst., vol. 41, no. 1, pp. 450–496, 2017.
5. T. Hofmann, "Probabilistic latent semantic indexing," Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 1999, vol. 51, no. 2, pp. 50–57, 1999.
6. S. W. Kim and J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," Human-centric Comput. Inf. Sci., vol. 9, no. 1, 2019.
7. D. M. Blei, A. Y. Ng, and M. T. Jordan, "Latent dirichlet allocation," Adv. Neural Inf. Process. Syst., vol. 3, pp. 993–1022, 2002.
8. O. Kolchyna, T. T. P. Souza, P. Treleaven, and T. Aste, "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination," no. July, 2015.
9. J. Hockenmaier, "Lecture 7: Topic Models," Engr-Courses.Engr.Illinois.Edu, no. Spring, 2013.
10. J. T. Medler, "The types of Flatidae (Homoptera) in the Stockholm Museum described by Stål, Melichar, Jacobi and Walker," Insect Syst. Evol., vol. 17, no. 3, pp. 323–337, 1986.
11. P. Monish, S. Kumari, and C. Narendra Babu, "Automated Topic Modeling and Sentiment Analysis of Tweets on SparkR," 2018 9th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2018, pp. 1–7, 2018.
12. D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec," Inf. Sci. (Ny)., vol. 477, pp. 15–29, 2019.
13. S. Li, J. Li, and R. Pan, "Tag-weighted topic model for mining semi-structured documents," IJCAI Int. Jt. Conf. Artif. Intell., pp. 2855–2861, 2013.
14. D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," ICWSM 2010 - Proc. 4th Int. AAAI Conf. Weblogs Soc. Media, pp. 130–137, 2010.
15. C. R. Association for Computational Linguistics. Meeting (45th : 2007 : Prague, R. E. Association for Computational Linguistics., P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol. - Vol. 1, pp. 142–150, 2007.

AUTHORS PROFILE

S. Sai Manasa Bala, currently studing M.Sc in Applied Mathematics,.from M. S. Ramaiah University of Applied Sciences, Bangalore. Her research interests are in the areas of big data analytics, Applied Mathematics for data science with an emphasis on statistical analysis/modeling, machine learning.

Santoshi Kumari graduated with a BE in computer science engineering (CSE), in 2009. She received her Mtech degree in software engineering in 2011. She is working as an assistant professor in the department of CSE, M. S. Ramaiah University of Applied Sciences, Bangalore. She is currently pursuing her PhD degree in the area of data science and NLP. Her research interests are in the areas of big data analytics, with an emphasis on data mining, statistical analysis/modeling, machine learning, and social media analytics.