Correlative study and analysis for hidden patterns in text analytics unstructured data using supervised and unsupervised learning techniques

E. Laxmi Lydia*

Computer Science and Engineering,
Vignan's Institute of Information Technology, India
Email: elaxmi2002@yahoo.com
*Corresponding author

S. Kannan

Citibank,
8 MARINA VIEW,
Asia Square Tower 1, 018960, Singapore
Email: Sathasivam.kannan@gmail.com

S. SumanRajest

Vels Institute of Science, Technology and
Advanced Studies (VISTAS),
Chennai, Tamil Nadu, India
Email: sumanrajest414@gmail.com

S. Satyanarayana

Raghu Engineering College,
Visakhapatnam, Andhra Pradesh 531162, India
Email: satyacomp@gmail.com

Abstract: Two-third of the data generated by the internet is unstructured text in the form of e-mails, audio, video, pdf files, word documents, text documents. Extraction of these unstructured text patterns using mining techniques achieve quick access to outcomes. Textual data available at online contains different patterns and when those huge incoming unstructured data enters into the system creates a problem while organising those documents into meaningful groups. This paper discusses document classification using supervised learning by focusing on the concept-based algorithm and also deals with the hidden patterns in the documents using unsupervised clustering technique and topic-based modelling for the analysis and improvement of systematic arrangement of documents by applying k-means and LDA algorithm. Finally, this presents comparative study and importance of clustering than classification for unstructured documents.

Keywords: text analytics; concept-based method; data representation; storage; latent Dirichlet allocation; LDA algorithm.

Reference to this paper should be made as follows: Laxmi Lydia, E., Kannan, S., SumanRajest, S. and Satyanarayana, S. (2020) 'Correlative study and analysis for hidden patterns in text analytics unstructured data using supervised and unsupervised learning techniques', *Int. J. Cloud Computing*, Vol. 9, Nos. 2/3, pp.150–162.

Biographical notes: E. Laxmi Lydia is an Associate Professor of Computer Science Engineering at Vignan's Institute of Information Technology(A). She is a big data analytics, online trainer for the international training organisation and she has presented various webinars on big data analytics. She is certified by Microsoft Certified Solution Developer (MCSD). She published 50 research papers in international journals in the area big data analytics and data sciences and she published ten research papers in international conference proceedings. She is an author for the big data analytics book and currently she is working on government DST funded project and she holds a patents.

S. Kannan has about 25 years of extensive professional experience in information technology industry in managing and leading wide range of program/project management function across industry, globally involving technology application systems development, IT transformation, enterprise integration, ERP system implementation, and data warehousing mainly for finance and supply chain domain. He is associated with Anti Money Laundering (AML) compliance domain in finance institution, practicing on data mining and data governance

S. SumanRajest is an editor of *International Research Journal of Management*, IT and Social Sciences and Journalist as well as courses in rhetoric and composition. All of his writing, including his research, involves elements of creative non-fiction. He is also interested in the intersections of creative writing and digital media. He has published an essay on creative non-fiction in several journals and anthologies, including the Creative Writing and Creative Writing Pedagogies in the Twenty-First Century including 4C' education. He is also on the editorial board of various journals. He has published widely in Literacy Studies and Rhetoric and Composition. Currently, he works in journalism and editor for newspapers, magazines, Freelancer, and various journals in Scopus Indexed, ISI, and UGC, etc.

S. Satyanarayana completed his PhD in Neural Networks and Bi-Directional Associative Memory in 2015 from Andhra University. He has huge experience of 15 years in academics and Administration of Engineering Education. His research area is data science and machine learning algorithms. He is focused learner on new technologies using python. He finished several certifications in various courses from NPTEL (India) and Coursra.

1 Introduction

The term 'text' is referred to as words, sentences, paragraphs. Unstructured data is been studied in terms of revealing hidden patterns in a huge amount of information. Unstructured data deals with automated information that does not have any data models. Most of the information retrieval approaches provide text-indexing techniques for handling unstructured data. Even though text mining and text analytics have, the same problems they strive to extract meaningful text. Individually they both perform different

techniques. Bringing together both text analytics and text mining usually, tend to higher performance when compared to a single approach. Techniques implemented in text mining are natural language processing (NLP), sentiment analysis, entity extraction, and categorisation. Some of the general basic steps for text mining are information retrieval, pre-processing steps like cleaning, segmentation, tokenisation, removal of stop words and punctuations, stemming, converting text to lowercase, grammatical tagging, creation of text corpus and text-document matrix. Text analytics is preceding steps for the text mining. It involves modelling, training, and evaluation and visualising of models after obtaining term document matrix. Text mining deals with the cleaning of data and extraction of data whereas text analytics apply statistical and machine learning approaches to the extracted data from the text mining.

Unsupervised learning depletes more timing while training data. To overcome time complexity when training, unsupervised learning has provided special techniques to find hidden patterns within the data and tries to extract every term in the document and extend that to cluster concepts. Text mining aims to build document clusters for fast retrieval and access to unstructured data. Any text document endures some structured fields like title name, the name of the authors, date of publications, and a name of the publications and so on, likewise unstructured data also contain abstract and contents, i.e., commonly used similar terms in every document. Clustering a long ongoing truly believable technique differentiating every document according to their similarity measures and a new method Topic modelling is the practically new approach used to search persistent patterns characterised words within huge text data.

Document clustering tries to find out the similarity among existing documents. Cluster maintains a related topic in the corpus for fast tracking of the concepts and properties of the documents. A text document, however, needs to be categorised based on the terms used in the document. Therefore, for an effective use of clustering, each document is represented in the form of vectors containing non-negative values. It is represented using document-term matrix (DTM). A document is analysed by performing TF-IDF to find out the most frequently used terms and make several text transformations to extract exact term and calculate similarity measure.

2 Literature survey

Text mining is a process of illustrating knowledgeable information by using various techniques in various areas for efficient communication and search from unstructured data. When large databases need help to find the hidden patterns in data mining, then text mining points to intelligent text analysis. Therefore, we bring in these present-days require more productive techniques for retrieval and improved procedures for indexing. Feldman and Sanger first introduced text mining when he was working with analysis of the text that needs service from the machine. The appropriate relevant techniques used are Information retrieval, NLP in the environment of data mining, machine learning, and statistics.

Han and Kamber (2006) has introduced that text mining is like a slice in data mining. Text mining approaches like latent semantic analysis for decomposition of text data into text document matrix to obtain the latent class model, probabilistic latent semantic analysis in distinct steps of retrieval of information, filtering and so on. Latent Dirichlet allocation (LDA) for topic-based modelling, hierarchical LDA is similar to LDA

(Alghamdi and Alfalqi, 2015) with additional topic modelling related to a corpus, dimensionality reduction techniques, and classification techniques. Feldman and Sanger (2006) have represented advanced approaches while analysing unstructured data. They provided a generic text mining system using document fetching techniques, pre-processing tasks like categorisation and term extractions, pre-processed documents will be compressed, text mining discovery algorithms are performed for pattern identification and trend analysis. Browsing functionalities like simple filters, interpreters for query and search, visualisation tools. Furthermore, refinement techniques include suppression, ordering, pruning, generalisation, and clustering for all online documents. Shi et al. (2007) have described nonlinear dimensionality for text classification. Singh and Raghuvanshi (2012) illustrated the need for text mining retrieval systems in various research applicational fields to expose the hidden mixture of patterns in unstructured data. This provides meaningful retrieval from databases and challenges from the data mining in handling uncertain data example spellings, grammar mistakes, Syntax, and semantic analysis. Puri and Kaushik (2012) expressed similarity by considering features of different documents based on approaches for text classification using fuzzy similarity text categorisation, fuzzy association (Yang et al., 2009), fuzzy C-means, production rules are very much useful and applicable in different areas.

Jadhav (2014), mentioned topic models implementing generative models for sentiment analysis to improve performance and various applications generated by the use of topic models in distinct sciences.

Some have interpreted that intense growth in online textual information demanded more in classifying instances. An interesting solution is suggested by authorising topics to documents. Text categorisation is performed by arranging documents by setting predefined classes. Two phases like training and prediction are applied to the input documents. The problem with supervised text classification learning is overlapping. Classifier used here is Naive Bayes is known for its self-determination of assumptions and is extremely scalable.

Alghamdi and Alfalqi (2015), have expressed solutions to unclassified text data using different approaches in topic modelling in two divisions by showing its importance towards word patterns, how documents are interconnected with other documents and probability distribution related to terms. Latent semantic analysis, LDA, correlated topic model without considering time another division is the topic evolution model by considering time as a most important factor. Approaches it follows are topic over time, dynamic topic models, multiscale topic tomography, dynamic topic correlation detection and detecting topic evolution. They mainly focused on the interconnectivity among new topics and original topics within the documents and examining them.

Liu et al. (2016) has explained all the existing applications in bioinformatics related to topic modelling by using different machine learning methods in the field of NLP. They concentrated on the terminology of 'document-topic-word' by using different toolkits like Gensim, TMT, MALLET, and some open-source packages from David Blei's Lab at Columbia University. Shotorbani (2016) investigated the various applications based on different classification and categorisation techniques in machine learning using concept based datasets.

Barde and Bainwad (2017) has presented the most highly essential approaches in topic modelling with respect to terms and concepts using standard modelling tools having

a bag of words and training models. They also specified applications related to tags, text categorisation, filtering of information, extractions of keywords and similarity measures.

Laxmi Lydia and Ramya (2018) and Laxmi Lydia et al. (2018) has described document clustering using unsupervised classification using the k-means algorithm in text mining by using similarity measures and text mining using NLP by extracting features using non-negative matrix factorisation.

In this paper, we have proposed an automatic document clustering procedure that uses-means, a discriminative model differentiates the contents and sets as cluster group and topic-based modelling using LDA a generative model which really knows how the words are generated and written in the document based on the probability (Agarwal et al., 2015).

3 Methodology

3.1 Document analysis and its representation

Every text document has a structure with some patterns. These text documents contain characters, numeric values, and symbols. To handle the large size of data that carries digital repositories and internet data simple techniques unable to handle it, therefore modern techniques using information retrieval has helped to perform processing over linear algebra and vector scale model. Data is shaped into the matrix to deal with noise and uncertain data and by performing vector functions. Every word in the document is treated as an individual and placed in the form of orthogonal.

- These documents are mostly represented using the vector space model, which is compatible with identifying and dealing with statistical calculation using frequencies of the terms to set up text information for retrieving high-quality information, NLP. Documents represented in this model are known as vectorising text.
- It stores the number of terms existing within the document and the frequency of each term.
- Representation of text in the form of queries some of the pre-processing steps in the vector space model are known as 'NLP pipeline', it includes
- Extraction of every individual word known as 'tokens'. This process is known as tokenisation.
- It removes every functional word from the document known as stop words removal.
- It extracts words to the base known as stemming.
- For improvement of query process IR consistently builds an inverted index.

Bag of words: Bag of words contains the count of each word, i.e., how many times a word is repeated. Based on the count of the word, term popularity increases. It is a very much advantageous process for similarity measure.

DTM: DTM is defined to determine the text document in the matrix form using terms or keyword or concepts within the document. Every term is calculated by its weight, or rank known as term weighing. When we try to form a DTM, functions like term frequency, document frequency and inverse document frequency are calculated to term

frequency-inverse document frequency. Documents should calculate the frequency of every term using term-frequency. Any term having more weight in the document is given by document frequency. Frequent words that are more relevant are defined by using inverse document frequency. Finally, all the sub-linear words obtained by calculating *tf* and *idf* are combined.

Suppose D is a matrix having a set of documents (m), i.e., $D = (d_1, d_2, \ldots, d_m)$ and V be the vocabulary having words (n), i.e., $V = \{w_1, w_2, \ldots, w_n\}$ known as document-term and if the rows of document matrix are ordered by the terms w_i and columns by the documents d_j. Known as term-document then, term-document matrix is given by D_{ij}.

This process is very simple for query language and dot product because the ranking reduces by using 0's and 1's. One of the major drawbacks of this model is that data is high in dimension and requires appropriate and proper representation. This model is poor in handling lexical uncertainty.

To overcome the ambiguity and variability problem in vector space model, concept-based representation, latent semantic analysis through SVD which reduces high dimensionality and maintains closer meaningful grammatical connectivity between terms? This will change the representation of documents, terms, and similarity slightly.

For classifying or clustering similarity among documents using document-document similarity. Likewise, we can also have term-term similarity for clustering document terms.

3.2 Supervised text document classification

Classification maps to have data with different predefined classes. They can be either single-label classification or multi-label classification. Text classification is an interesting category of accomplishing multiple topics from the document for easy selection of features maintaining predefined classes. Supervised learning has a problem of overlapping, so every class is separated as a group.

3.2.1 Term-based classification

Term-based classification focuses on the terms in the documents predefining classes using dictionary terms.

Preparation of training data: It is completely a supervised learning with a pre-defined classification for training data. Documents mutually model corpus for training. A series of steps for pre-processing is performed and convert them to matrix form calculating terms by using vector representation. A bag of terms is needed to classify using the specified classifier. Following are the steps followed by term-based classification:

Step 1 Unknown text documents are taken as input to classify.

Step 2 Preparation of training data using text corpus to extract term documents.

Step 3 Pre-process the text corpus.

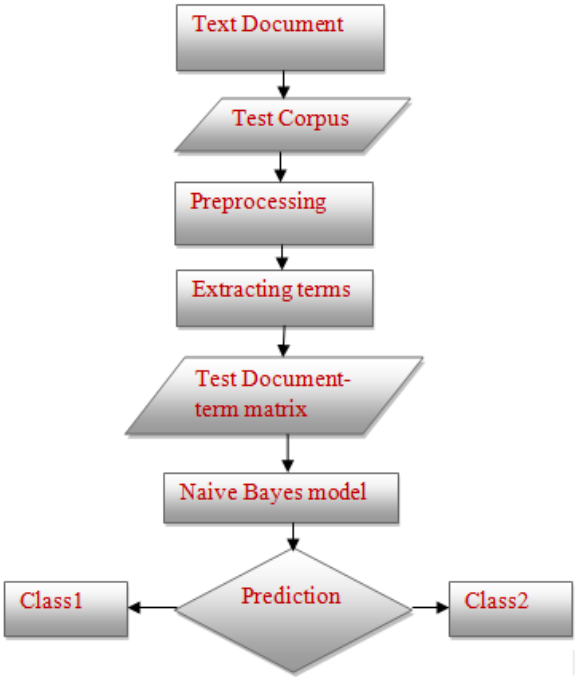
Step 4 Extract every term from the pre-processed test corpus and form term-document matrix.

Step 5 Test DTM.

Step 6 Apply classifier (Naive Bayes model) because it uses less training data related to remaining supervised models.

Step 7 Finally, the prediction is performed based on the categories.

Figure 1 Flowchart for general term-based classification (see online version for colours)



The flowchart in Figure 1 describes the process of predicting document using term-based classification whether the input document is classified as class1 or class2 based on training data by extracting terms.

Although there are quite advantageous in term-based classification, few failures were recognised during this learning process.

- When handling more number of training documents, it automatically leads to a large number of terms. Ever after performing pre-processing operations, we can still find inappropriate terms. This may fall down to reduce inaccuracy. Very much sensitive in size.
- Document vector representation is not explicitly equal to the actual document.
- Takes cost for training.

Even after pre-processing, the trained data is not completely clean and still contains some inappropriate irrelevant or generic terms, which will reduce the precision of the classification process.

3.2.2 Concept-based classification

Due to the existence of insufficiency in the term-based method, an alternative approach has been involved in the generation of dictionary words, i.e., concept-based learning. Similar to the term-based classification it resides in supervised learning, but it is distinguished by considering a bag of concepts than terms. Here every concept is referred to a class label having alternative class labels too. It is based on simple knowledge organisation system (SKOS) for thesaurus and this includes concepts, definitions, relations for defined semantics and labels.

Preparation of training data:

Step 1 Training data is provided through requesting queries to the generated dictionary.

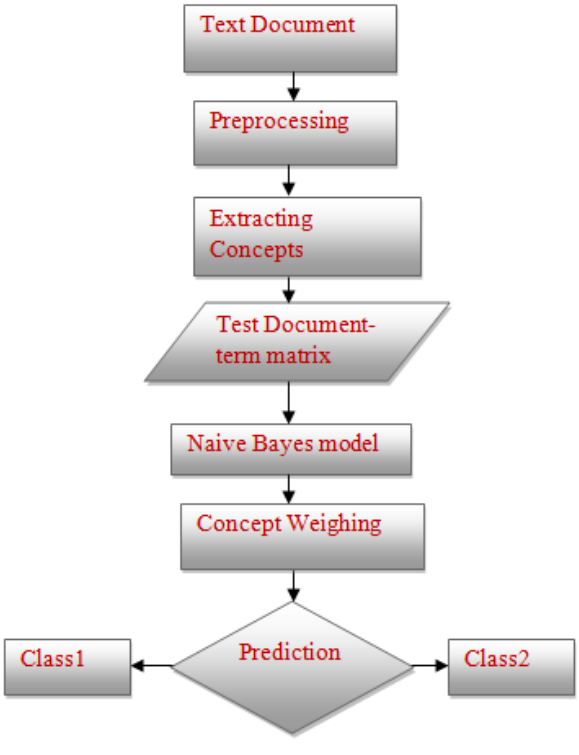
Step 2 Document classification is performed using the conceptual level. The most preferred classifier is Naive Bayes. It helps in calculating probabilities to build classes.

Step 3 Results for the queries are selected as features of the concepts that distinguish a class by using entry concept selection and sub-tree extraction. This generates a label and word frequency matrix.

Step 4 Concept weighting to check how relevant the concepts are in documents.

Step 5 Concepts of similarity and dissimilarity.

Figure 2 Flowchart for general concept-based classification (see online version for colours)



The flowchart in Figure 2 describes the prediction of documents using concept-based classification by maintaining thesaurus and using concept weighting for document similarity. Advancement in suggested higher accuracy lies a part deeper in the larger dictionary set for extracting relevant concepts and all concepts that are both limited and unlimited. Term-based and concept-based aggregates with various queries. Therefore, no pre-processing is needed in concept-based because of relevant terms that are already present in the corpus.

3.3 Unsupervised document clustering

Unsupervised techniques when compared to the supervised techniques, where data needs to be trained at a greater extent due to high quality and when data is not trained appropriately results with inaccurate outcomes. In order to deal with such issues, the data analyst's have opted to choose unsupervised techniques. Use of unsupervised methodologies, the unrevealed sequence of text documents have been revealed without training them. In this paper, two most frequently and analysed unsupervised learning techniques have been introduced especially for clustering and topic modelling. Clustering is a mechanism of organising every document to its appropriate and similar contents. Most significantly implemented a new method named topic modelling is used to discover repeated existing words among large texts.

Text mining introduces a way to select terms and term clustering to find similar terms and replacing them with centroids by using clustering k-means partitioning algorithm. This algorithm spontaneously groups the similar documents compared to other clusters.

3.3.1 K-means

The k-means algorithm is identified in larger datasets it specifies the user to declare the number of document clusters by considering the minimum distances by verifying centroid. These distances are computed based on the term matrix. Following are the basic considerations for performing clustering algorithms

- Documents are distributed between predefined clusters randomly.
- The position of each cluster is identified related to the centroid.
- Distances are calculated among the documents as well as centroids.
- Based on the distances every document will be assigned to the closest centroid, each centroid forms a cluster.
- The procedure continues until all the documents are arranged to the clusters.

Following are the defined equations for the k-means algorithm:

Suppose documents are arranged in a cluster with the *centre* as (w) and the *centroid* as (u) then

$$\vec{u}(w) = \frac{1}{w} \sum_{\vec{x} \varepsilon w} \vec{x} \tag{1}$$

The distance from every vector is summed from the centroid and squared is given by

$$RSS_k = \sum_{\vec{x} \in w_k} \left| \vec{x} - \vec{u}(w_k) \right|^2 \tag{2}$$

$$RSS = \sum_{k=1}^{k} RSS_k \tag{3}$$

where *RSS* is defined as the residual sum of squares, w_k is defined as documents with k clusters, $\vec{u}$ is described as the documents with the centroid in cluster w_k and $\vec{x}$ is described as the document vector in k clusters, it can also be defined using DTM. The main purpose of using K-means is to achieve minimum RSS again the updated centre of clusters.

3.3.2 Topic modelling using LDA

LDA is described as a combination of topics with probabilities in the documents generally termed as a probabilistic method. Words are estimated with a probability distribution and try to generate a particular structure in the documents. Mostly used for topic modelling. It helps to find the latest approaches to explore and hypothetical excessive expository of texts.

The two basic presumptions of this model are as follows:

- There will be only a fixed sequence of words in a defined dictionary known as topics.
- It maintains different probabilities for every document topics with respect to the corpus.

Using LDA, the probability of every topic is estimated by the following equation

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d)\, p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d \tag{4}$$

where D defines the document corpus, M defines the total number of documents, N defines the total number of words in the documents, θ_d defines the document-level variable, α defines the Dirichlet level parameter, β defines the Dirichlet level parameter, z_{dn} and w_{dn} defines the word-level variables.

The necessary procedural steps for IDA are given below

Step 1 Choose topics are defined.

Step 2 Every document is determined with topics containing terms.

Step 3 Topic identification and representation are performed. Similarly, various distinct words to related topics.

Step 4 Calculate ratio among words within documents, which are already assigned to the topics by dealing with entire documents.

Step 5 Now, again assign every word to the newly existed topics. These steps are performed for multiple repetitions to accomplish the optimal result.

3.4 Comparative study on term-based and concept-based on the supervised method with k-means and LDA clustering on unsupervised method

Table 1 Comparative study on classification techniques of term-based and concept-based and clustering techniques k-means and LDA clustering in text mining

Supervised method	Term-based classification	Term-based classification focuses on the terms in the documents predefining classes using dictionary terms. A bag of terms is needed to classify using the specified classifiers.
	Concept-based Classification	Concept-based classification considers a bag of concepts than terms, every concept is referred to a class label having alternative class labels and uses classifiers.
Unsupervised method	Clustering with LDA	LDA combines topics with probabilities in the documents generally termed as a probabilistic method.
	Clustering with K-means	Introduces a way to select terms and term clustering to find similar terms. Replacing them with centroids by using clustering, k-means partitioning algorithm.

4 Result analysis

For experimental analysis, we have considered 30 text documents with 10 defined topics. Software specifications applied are R 3.4 version on Linux environment. A package named 'topic models' is taken and data with a defined set of topics.

Figure 2 explains the existence of a topic in documents by using LDA approach, document1 contains topic 3, document 2 contains topic 9, document 3 contains topic 8, and document 4 contains topic 7, respectively.

Figure 3 R using data frame defining documents and topics as rows and columns (see online version for colours)

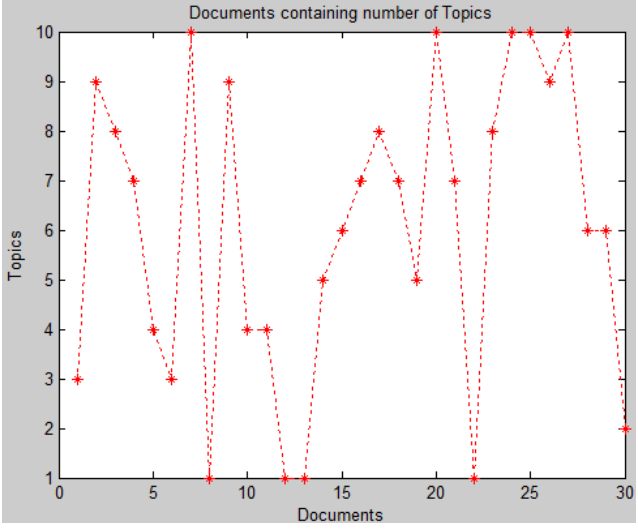
```
> toptopics
   document topic
1         1     3
2         2     9
3         3     8
4         4     7
5         5     4
6         6     3
7         7    10
8         8     1
9         9     9
10       10     4
11       11     4
12       12     1
13       13     1
14       14     5
15       15     6
16       16     7
17       17     8
18       18     7
19       19     5
20       20    10
21       21     7
22       22     1
23       23     8
24       24    10
25       25    10
26       26     9
27       27    10
28       28     6
29       29     6
30       30     2
```

Figure 3 explains the probability of each document in topic models relating to defined topics and the ten topics are known by using terms (Ida).

Figure 4 Probability of documents and considered topics (see online version for colours)

```
> #To see which documents belong to which topic with the highest probability in topic models
> topics(lda)
 [1]  3  9  8  7  4  3 10  1  9  4  4  1  1  5  6  7  8  7  5 10  7  1  8 10 10
[26]  9 10  6  6  2
> #To see the the topics generated from all the documents
> terms(lda)
  Topic 1   Topic 2   Topic 3   Topic 4   Topic 5   Topic 6   Topic 7   Topic 8
"dukakis"  "soviet"   "saudi"    "bank"   "baker"    "fire"   "study" "roberts"
  Topic 9  Topic 10
"soviet"      "two"
```

Figure 5 Graph showing document similarity with topics using MATLAB (see online version for colours)



This defines that among 30 documents, we can easily find the similarity of documents having similar topics.

In this graph, we can see that topic1 named 'Dukakis' is related to four documents, i.e., 8, 12, 13, and 21, Likewise, documents 7, 20, 24, 25, 27 are related to topic 10 named 'two'. Defining topics, we can easily find the existence of those topics in documents. To know most similar documents less number of important topics should be defined because if topics are more we get more documents to classify.

5 Conclusions

This paper has undergone a study and analysis by classifying and arranging text documents using supervised and unsupervised learning. Supervised classification using term-based and concept-based with the optimal and reliable preparation of training text data. In order to attain an even more powerful optimal solution unsupervised learning techniques like k-means clustering and topic modelling using LDA has given an easier

way to 'mine' documents by discovering hidden patterns. Use of LDA has shortened the process by grouping similar documents by using topics.

Acknowledgements

This work is financially supported by the Department of Science and Technology (DST), Science and Engineering Research Board (SERB) under the scheme of ECR. We thank DST, SERB for the financial support to carry the research work.

References

Agarwal, V., Thakare, S. and Jaiswal, A. (2015) 'Survey on classification techniques for data mining', *International Journal of Computer Applications*, Vol. 132, No. 4, pp.13–16, DOI: 10.5120/ijca2015907374

Alghamdi, R. and Alfalqi, K. (2015) 'A survey of topic modeling in text mining', *International Journal of Advanced Computer science and Applications*, Vol. 6, No. 1, DOI: 10.14569/ijacsa.2015.060121.

Barde, B.V. and Bainwad, A.M. (2017) 'An overview of topic modeling methods and tools', *International Conference on Intelligent Computing and Control Systems*, IEEE, ISSN: 978-1-5386-2745-7/1.

Feldman, R. and Sanger, J. (2006) 'The text mining handbook: advanced approaches in analyzing unstructured data', *Computational Linguistics*, Vol. 4, No. 1.

Han, J. and Kamber, M. (2006) *Data Mining Concepts and Techniques*, 2nd ed.,, University of Illinois, Urbana-Champaign.

Jadhav, N. (2014) 'Topic models for sentimental analysis: a literature survey', in *Proceedings Semantic Scholars*, Jadhab, Topic, MF, pp.1–11.

Laxmi Lydia, E. and Ramya, D. (2018) 'Text mining with Lucene and Hadoop: document clustering with updated rules of NMF (non-negative matrix factorization)', *International Journal of Pure and Applied Mathematics*, Vol. 118, No. 7, pp.191–198, ISSN: 1314-3395.

Laxmi Lydia, E., Govindaswamy, P., Lakshmanaprabu, S.K. and Ramya, D. (2018) 'Document clustering based on text mining K-means algorithm using Euclidean distance similarity', *Jour. of Adv. Research in Dynamical & Control Systems*, Vol. 10, No. 2-Special Issue, pp.208–214.

Liu, L., Tang, L., Dong, W., Yao, S. and Zhou, W. (2016) 'An overview of topic modeling and its current applications in bioinformatics', *SpringerPlus*, Vol. 5, No. 1, p.1608.

Puri, S. and Kaushik, S. (2012) 'A technical study and analysis on fuzzy similarity-based models for text classification', *International Journal of Data Mining & Knowledge Management Process*, March, Vol. 2, pp.1–15, doi: 10.5121/ijdkp.2012.2201.

Shi, L., Zhang, J., Liu, E. and He, P. (2007) 'Text classification based on nonlinear dimensionality, reduction techniques, and support vector machines', *Third IEEE International Conference on Natural Computation*, Vol. 1, pp.674–677 [online] http://blog.thedigitalgroup.com/supervised-leraning-for-text-classification.

Shotorbani, P.Y. (2016) *Text Mining Techniques for Analyzing Unstructured Manufacturing Data*, Graduate Council of Texas State University in Partial Fulfillment of the Requirements for the Degree of Master of Science with a Major in Technology Management, August.

Singh, P.D. and Raghuvanshi, J. (2012) 'Rising of text mining technique: as unforeseen-part of data mining's', *International Journal of Advanced Research in Computer Science and Electronics Engineering*, May, Vol. 1, No. 3, ISSN: 2277-9043.

Yang, Q., Chen, W. and Wen, B. (2009) 'Fuzzy ontology generation model using fuzzy clustering for learning evaluation', *IEEE International Conference on Granular Computing (GRC)*, pp.682–685.