Websites:
http://www.jofamericanscience.org
http://www.sciencepub.net

Emails:
editor@sciencepub.net
sciencepub@gmail.com

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

Big Data: Algorithmic Approaches to Topic Modelling Analysis for Social Media News Data

Bisallah H. I*, Olumide O*, Aminat A*

*Department of Computer Science, University of Abuja, Nigeria
hashim.bisallah@uniabuja.edu.ng

Abstract: Topic Modeling is a computational model that derives the latent theme from large collection of text data. In this paper we developed a topic model for Nigerian Newspapers social media news corpus to find the screened topics from the corpus. Topic modeling algorithms Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and three different machine learning approaches (Naive Bayes, K-NN and K-means) was implemented. The performance of topic modeling algorithms with machine learning approaches using the measures precision and recall was compared. Topic modeling algorithms with multiple topic distribution shows better for corpus in the social media data obtained.

[Bisallah H. I, Olumide O, Aminat A. **Big Data: Algorithmic Approaches to Topic Modelling Analysis for Social Media News Data.** *J Am Sci* 2019;15(7):61-69]. ISSN 1545-1003 (print); ISSN 2375-7264 (online). http://www.jofamericanscience.org. 7. doi:10.7537/marsjas150719.07.

Key words: Topic modelling, Latent Semantic Analysis, Naïve Bayes, K-NN, K-means, Social media, Latent Dirichlet Allocation.

Introduction

Over the last few years, people have migrated towards digital media. Close to 90% of the data are in unstructured form which compounds the problem at hand. Text analysis is one of the techniques used for deriving high quality of information from the text (D. M. Blei, A. Y. Ng, Mi. i. Jordan, 2003). Many text analysis tasks are involved in obtaining information about the digital data. It involves tasks like data retrieval, document clustering, concept/entity extraction, document categorization, sentiment analysis, topic modeling and visualization (Sebastiani, F, 2005).

Topic modeling approaches find the hidden topics (David M. Blei, John D. Lafferty, 2014). The most common methods involved in topic modeling are: Latent Semantic Analysis (LSA), probabilistic Latent Semantic Analysis (pLSA), and Latent Dirichlet allocation (LDA). LSA aims to discover meaning behind the words about the topics in the documents. pLSA is automated document indexing and is used in information retrieval. It identifies latent classes using an Expectation Maximization Algorithm. LDA discovers the unobserved groups from similar groups of data, wherein, the words are assigned to particular topics from documents. Tweets data from Nigerian news social media provides benchmark for research purpose. The tweets and website contents covers facts and people opinion, this gives a multiple topics to visualize.

This research paper has taken some selected Nigerian news media tweets handles dataset which contains different topics such as politics, sports and business. A comparative analysis of topic modeling algorithms with machine learning approaches were adopted. The results show that the topic modeling algorithms perform better than machine learning algorithms for finding latent information. Implementation is done using python as stated in the preceding section, and the performance is calculated using the measures of precision and recall (K.R. Bindu, Latha Paramesran, K.V. Soumya, 2015).

Section two delved on the methodologies of classification and clustering algorithms used for finding similar documents from a given corpus. An attempt to discuss the employability of Naive Bayes, K-NN classifier and K Means clustering for the social media news dataset was made. Section three delves on the various topic modeling approaches - Latent Semantic Analysis and Latent Dirichlet Allocation. Sections four, five and six described the implementation, results and performance analysis of various algorithms.

Machine Learning Approach

Naïve Bayes

The dataset contains the topics sports, politics, and business as class labels, hence an attempt to classify the corpus based on each category. The corpus is characterized into bag of words and it follows the Bayes theorem rule for further classification (F. Semastiana, 2002).

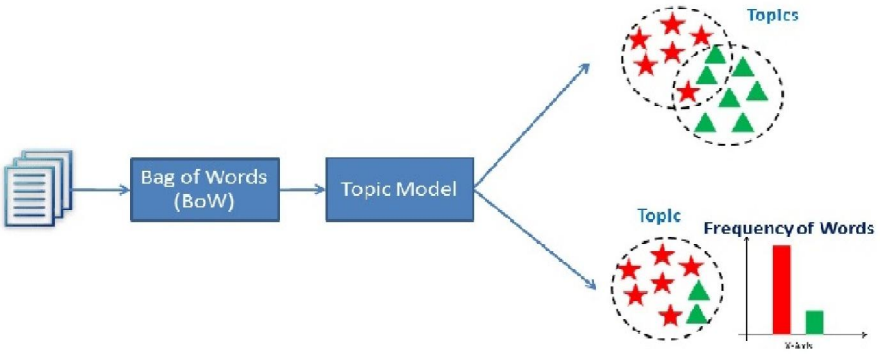

Fig 1. A set of bag of words (S) from a corpus is given as an input to the Naive Bayes classifier.

Let C denote the set of topics (C= {c1, c2})
where, c1 represents Topic1 and c2 represents Topic 2.
Naive Bayes classifier for Punch_web is as follows:
$P(c_i/w_d) = P(w_d/c_i)\ P(c_i) / P(w_d)$, i=1,2
Where,
$P(c_i/w_d)$ - Conditional Probability of topic for a given word
$P(w_d/_{ci})$ - Conditional probability of word for a given topic c_i.
$P(c_i)$ - Prior probability of topic
$P(w_d)$ - Probability of words

Naive Bayes pseudocode

Procedure:
Input: Set of bag of words {Term Document Matrix}
Output: $P(c_i/w_d)$ – conditional probability of class c_i given word w_d

1. For each c_i belongs to C,
   - find prior probability of $P(c_i)$
2. Count the words (w_d)
3. Find conditional probability of word w for given class c_i- $P(w_d/c_i)$
4. Compute $P(c_i/w_d) = P(w_d/c_i)\ P(c_i) / P(w_d)$

One major disadvantage of Naive Bayes approach is the requirement for fine-tuning the scale parameter *C (denotes the set of topics)* and it suits for binary classification problems rather than multidimensional category problem.

K-Nearest Neighbor (K-NN)

K-Nearest Neighbor is a classification algorithm which finds out closest neighbors among all categories in the document. It classifies new label by maximum number grouped under K neighbors. It segregates unlabeled data points into well-defined groups (C. Cortes, V. Vapnic, 1995).

Here K-NN is used for text categorization. From here the corpus training and the unknown samples are taken.

Initially, the term document matrix for training an unknown sample is calculated, and along with this the class label is given as an input to the K-NN algorithm.

Let X be the weighted document term matrix for training data

Y be the class labels

I be the Indices and

x be the weighted matrix for unknown sample

Set the value of K and find the distance between each x.

The pseudocode for K-NN:

Procedure
Input: X -Weighted document term matrix of training and unknown sample, Y- class label
Output: Class label Y_i for x
Steps:
1. For both training and unknown sample calculate the TF-IDF matrix.
2. Set k
3. For i =1 to n do
   Compute distance $d_i(X_i, x)$
4. end for
5. I[ ]=i of k smallest distances from $d_i(X_i,x)$
6. return majority label Y_i where each i belong to I

Now the majority of the given classes are classified under each category. The drawback in K-NN algorithm is the difficulty to compute K value. The classification time complexity is higher as compared to Naive Bayes classification - thereby incurs high computational cost.

K-Means Clustering

In K-means algorithm, for each cluster K, the centroid has to be defined and grouped under minimum distance of item and its centroid.

Here the dataset contains collection of documents on various topics. K-means results to form group of documents belonging to same topic.

TF-ID matrix is given as an input, the K value is set and the Euclidean distance of n data points to nearest cluster is determined.

The Pseudocode for K-means

Procedure:
Input: Weighted document term matrix (TF-IDF)
Output: Given corpus is grouped into K topics
Steps:
1. Calculate the TF-IDF matrix for given document.
2. Set the k value
3. Initialize the centroids
4. Employing Euclidean distance, grouping n points to nearest cluster.
5. For each cluster, calculate the centroid point.
6. Repeat steps 2, 3, 4 until no change in position of centroid is observed.

K-means cluster formation has a drawback regarding the number of clusters to be formed since this number is based on the initial K value assumed by the user (Likas, Aristidis, Nikos Vlassis, Jakob J. Verbeek, 2003).

Topic Modelling Approaches

Latent Semantic Analysis (LSA)

LSA is a technique used to extract and infer the relationship between the related words in a given context. Building a LSA does not involve any human-computer interaction or user-defined dictionaries, grammars, semantic networks etc (Landauer, T. Foltz, D. Laham, 1998). LSA receives the unprocessed data as input, parses them to get terms (unique character strings). These processed strings are separated into meaningful groups such as sentences or paragraphs.

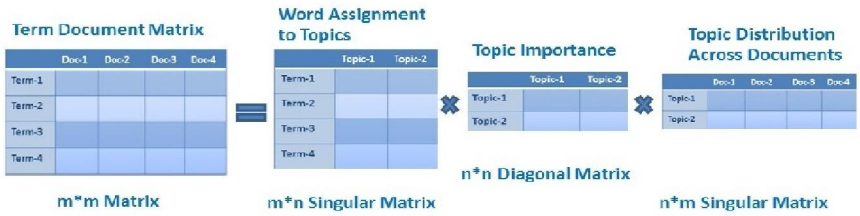

Fig 2. Steps involved in LSA Algorithm

- **Text Matrix**

Corpus has been represented in matrix format which specifies the frequency of the word for each term present in the document. Column [1..n] specifies the document and row 1 specifies the terms. Corpus is represented as bag of words using which TF-IDF matrix is created.

- **Generating TF-ID (Term Frequency-Inverse Document Frequency)**

TF-ID which describes the number of occurrence of word in a document. It is used for removing stop words.

Tf-id= (Ni, j / N*j) * log (D/Dj)

Ni, j – No. of times word i appear in document in j

N*j – No. of total words in document j

D - No. of document (Represented in number of Column)

Dj- No. of document in which words i appears.

- **Generating SVD Matrix**

Within corpus, terms (words) are used to build a library, which contains set of words to ensure text matrix. The TF-ID matrix is then decomposed into SVD. SVD is a numerical method which shows the relationship between the words/phrases and sentence. When the input matrix A is given it is decomposed into three matrices where U represents the term Eigen vector matrix, Σ represents the diagonal matrix and V^T represents the document Eigen value matrix.

$\mathbf{A = U \Sigma V^T}$

- **Distance / similarity:**

Cosine similarity is used to find the distance between the document and terms (K.P.N.v. Satyasree, J.V.R Murthy, 2012). If two documents are given their similarity is calculated between two column vectors of cosine angle. Maximum similarity is shown when the angle value is small which indicates cosine values will be high (Vikasthada, Vivekjaglan, 2013).

Latent Dirichlet Allocation (LDA)

LDA is an unsupervised statistical model that shower out words with certain probabilities. It is similar to probabilistic latent semantic analysis (pLSA), but LDA uses Bayes estimation and pLSA uses maximum likelihood estimation (David Blei, M Andrew, Y.Ng, Michael I. Jordan, 2003). Let M be the number of documents, α be the Dirichlet distribution of topics, β be the vocabulary matrix, t be the topics in a corpus.

Procedure:

Input: Number of document M
Number of topics t
β Vocabulary matrix

Output: Topic probability distribution for each word in document.

Steps:

1. Choose the topic distribution α
2. Assign each word W in a document d to one of the t topics.
3. For each word W in a document d
   - For each topic calculate P(Topic t | Document d)
   - Calculate P(word W | Topic t)
4. The selection word W for a topic t is depends on the distribution of β Vocabulary words.
5. Finding the posterior probability
   P(Topic t| Word w,α,β) = P(Word W |Topic t,β).P(Topic t | Document d).P(Document d | α)

The pseudocode for LDA

Implementation using Python

Dataset Description

The dataset being used here was obtained from the different web portals of the newspaper platforms.

Also, data was mined using from social media, the social media hashtags belonging to this newspapers.

Preprocessing

In text mining techniques, pre-processing plays a significant role. Preprocessing is the initial step in text mining (Vikasthada, Dr. Vivekjaglan, 2013). Here, the first step is to convert unformatted data to plain text files. We then remove all numbers, signs, symbols, non-English letters, stop words, convert all English letters to lowercase and perform stemming (K. R. Bindu, Latha Parameswaran, K. V. Soumya, 2015).

Python-Packages

With this, a new latent semantic space can be constructed over a given document-term matrix. To ease comparisons of terms and documents with common correlation measures, the space can be converted into a text matrix of the same format as text

matrix (Deerwester, S. Dumais, S. Furnas, G. Landauer, R. Harshman, 1990).

The library functions used for LSA are: -

- Numpy (np)
- Pandas (pd)
- Matplotlib (plt)
- Sklearn (skl)

Results

Results: Machine Learning Approaches

Here, we have implemented various machine learning classification and clustering algorithms for finding similar documents from a corpus. We experimented by giving input in two different forms: first the corpus contained same topics, and second the corpus contained mixed topic represented as business, politics and sports.

K-NN Classification Algorithm

Fig 3a show the ggplot of K-NN algorithm of different topics in which document 4 belongs to politics group, document 1 belongs to sports group, document 5 belong to economy group and remaining documents belong to business group. Fig 3b shows the ggplot of K-NN algorithm of similar topic in which the documents are classified into business group except the documents 1 and 4.

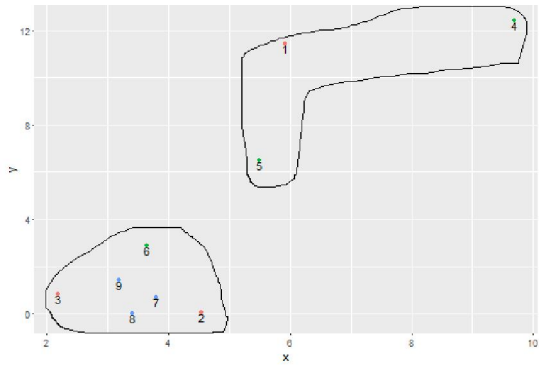

Fig 3a. ggplot of K-NN algorithm of mixed topic (K=3)

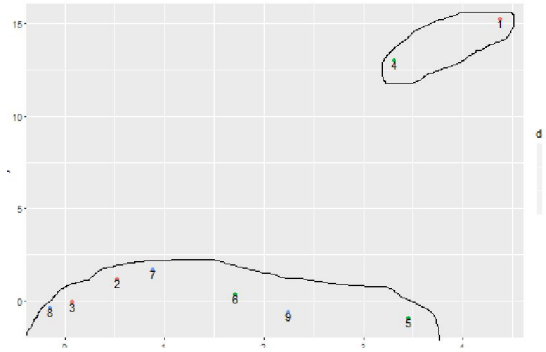

Fig 3b. ggplot of K-NN algorithm of similar topic (k=3)

Naive Bayes Classification Algorithm

Fig 4 shows the top 10 word probability values under two classes topic 1 and topic 2, where each row denotes the word, column 1 denotes class 1 (business group) and column 2 denotes class 2 (sports group).

	1	2
1	0.414177	0.585823
2	0.495673	0.504327
3	0.354197	0.645803
4	0.884772	0.115228
5	0.830915	0.169085
6	0.835195	0.164805
7	0.250543	0.749457
8	0.830915	0.169085
9	0.503366	0.496634
10	0.644118	0.355882

Fig 4. Result: Naive Bayes Algorithm

K-means Clustering

Fig 5 shows the results of K- means clustering algorithm with K=4. The four diffferent clusters are represented in different colours, grouped under category of politics, sports, business and entertainment.

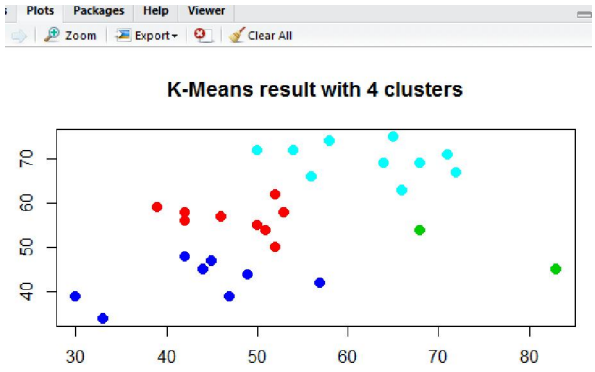

Fig 5. K-means Plot

Results: Topic Modeling Approaches

Latent Semantic Analysis

Fig 6a and 6b show graphical representation of the corpus which contains mixed topics and similar topics of corpus respectively.

Fig 6a: Plot of LSA algorithm for mixed class

Fig 6b: Plot of LSA algorithm for similar class.

By comparing the results of classification algorithm, LSA gives better classification results when compared to K-NN and Naive Bayes algorithm.

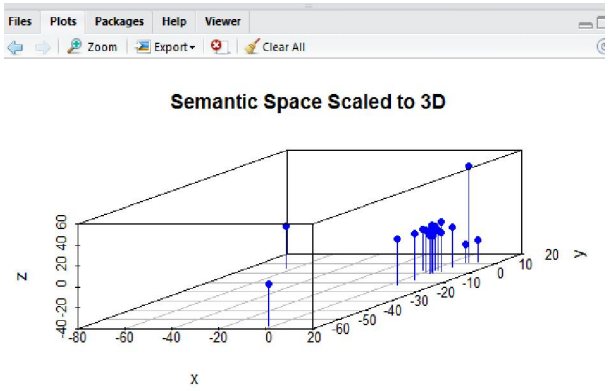

Fig 6a. Semantic Space Scaled to 3D-1

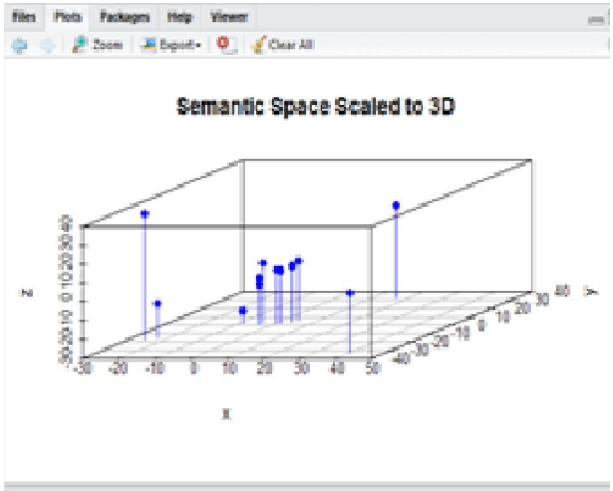

Fig 6b. Semantic Space Scaled to 3D-2

Latent Dirichlet Allocation

Fig 7 shows LDA results for clustering of two topics. Topic 1 and Topic 2 belong to categories of business and sports respectively.

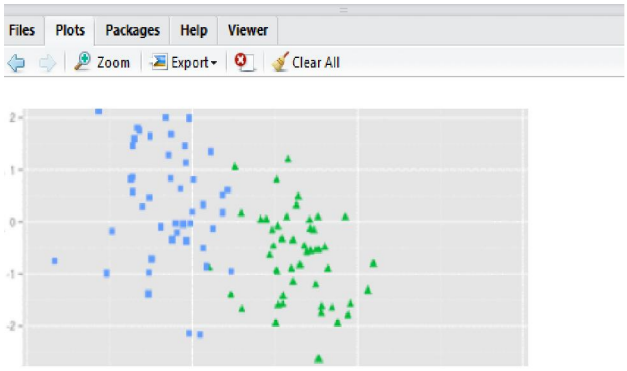

Fig 7. Plot of LDA algorithm

Fig 8 depicts the topic probability of each word, where, the rows represent the words (top 10 words are displayed here) and columns represent the topics (Topic 1-business and Topic 2-sports).

	1	2
1	4.810342e	9.980759e
2	2.651256e	9.989395e
3	1.430945e	1.430945e
4	8.402940e	8.402940e
5	8.404830e	9.996638e
6	1.903630e	9.992385e
7	1.297372e	1.297372e
8	6.906241e	6.906241e
9	1.242780e	1.242780e
10	9.997361e	6.597684e

Fig 8. LDA algorithm result

Performance Analysis

In this research paper, topic search tells if whether the given topic is relevant or not to the document. The performance analysis for the same is done by evaluating the measure of precision and recall.

Fig 9a and 9b show the performance analysis of classification algorithms using precision and recall statistics. It is clearly shown that LSA outscores the machine learning algorithm.

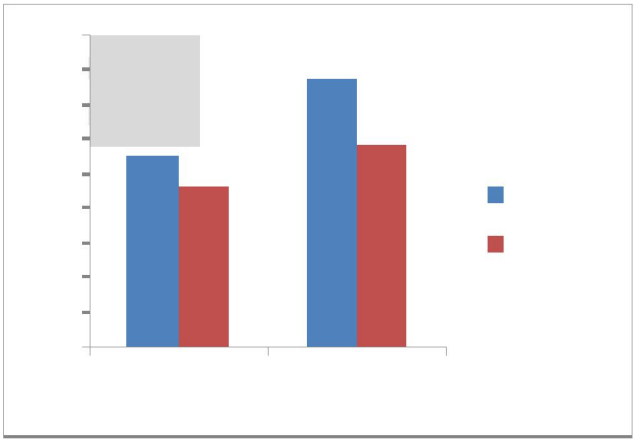

Fig 9a. Comparison: Classification algorithm for mixed topic

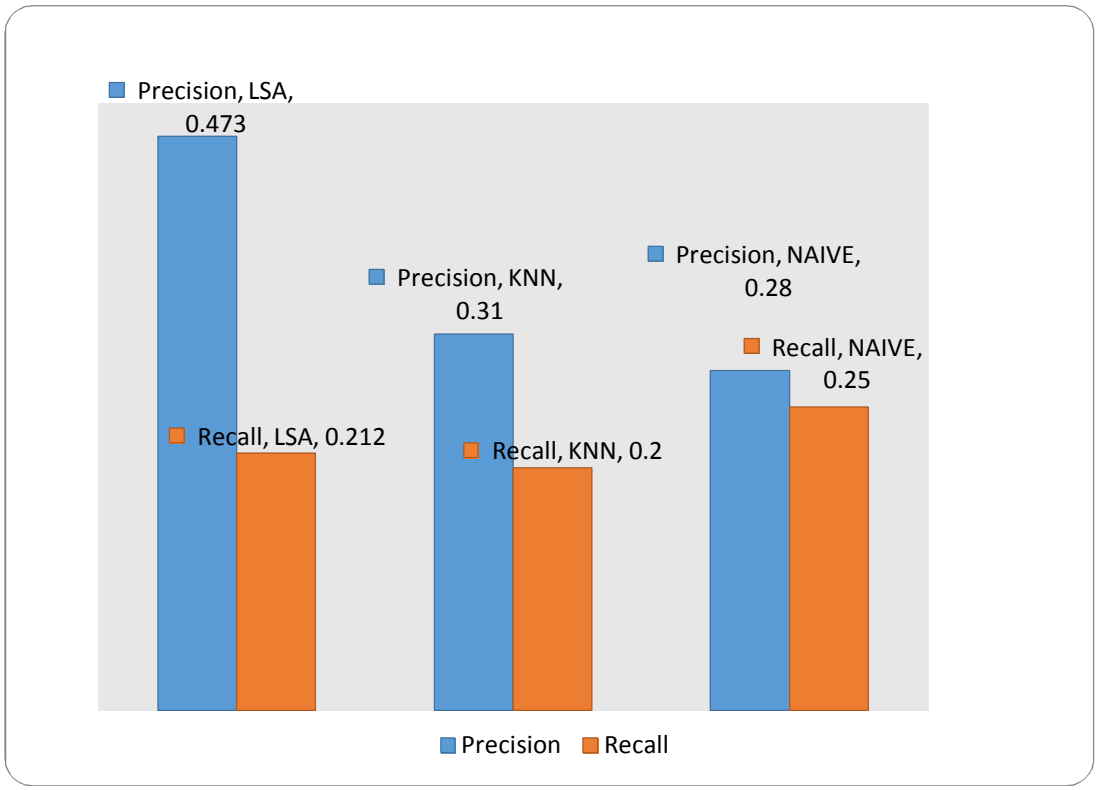

Fig 9b. Comparison: Classification algorithm for similar topic

Fig 10a and 10b show the performance analysis for clustering algorithms using the measures of precision and recall.

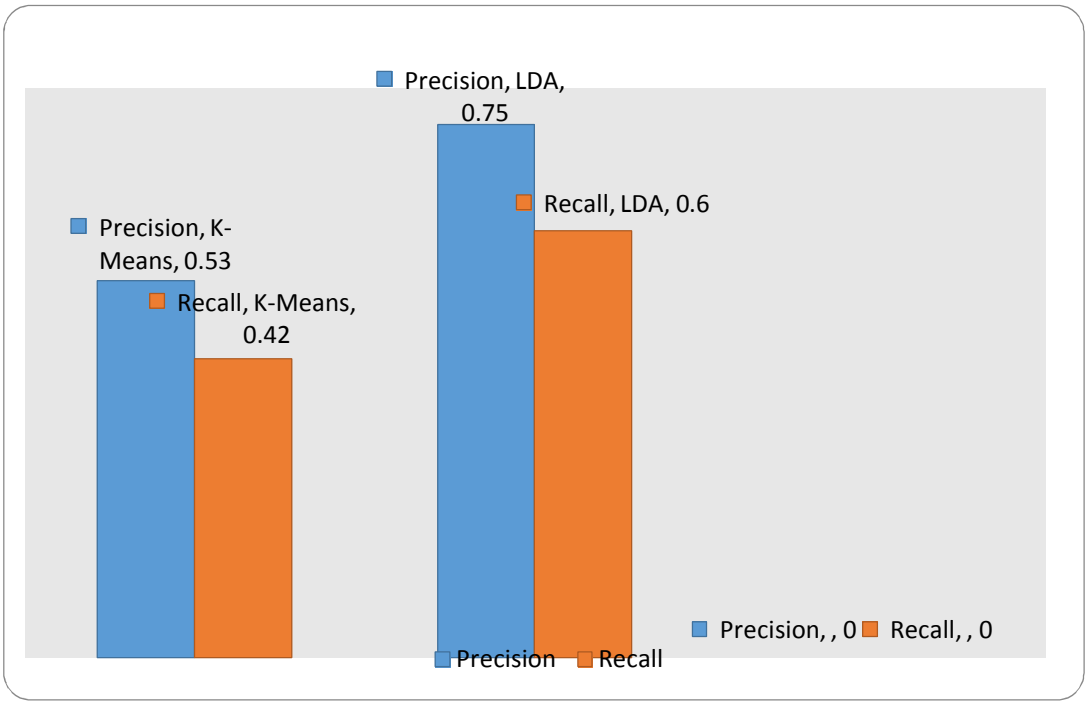

Fig 10a. Comparison: Clustering algorithm for mixed topic

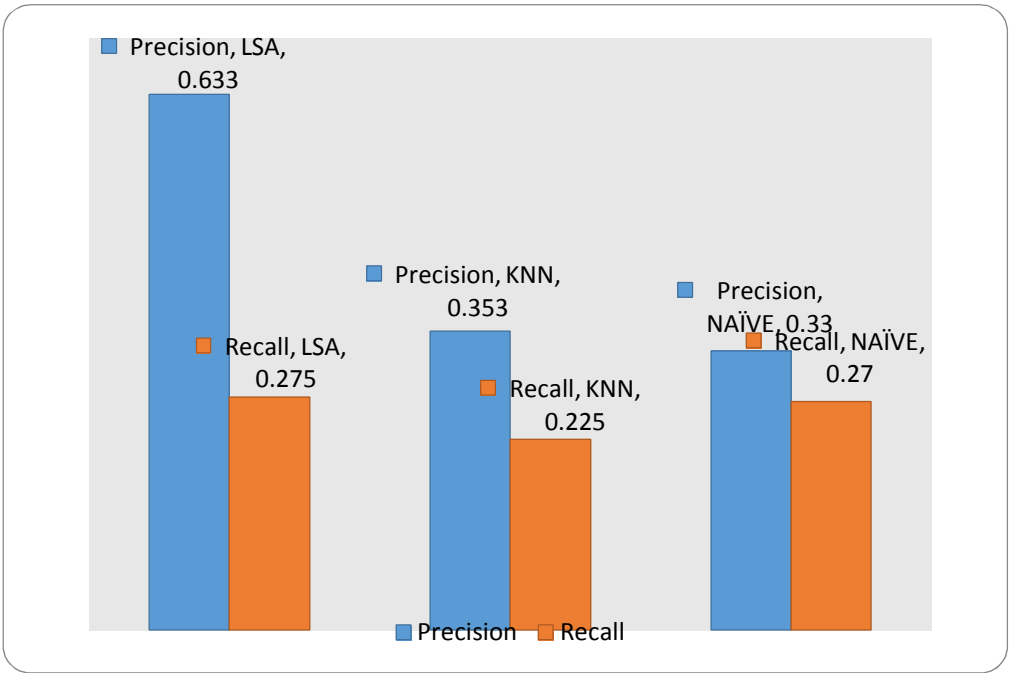

Fig 10b. Comparison: Clustering algorithm for similar topic

Conclusion

Having applied various machine learning approaches (Naive Bayes, K-NN and K-means), Topic Modeling approaches (Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), from the comparative analysis made, it has been observed that the topic modeling approaches find the hidden topics and relationship between words and documents with multiple probability distributions.

References

1. B. Grun, K. Hornik, "A Phyton Package for Fitting Topic Models", Journal of Statistical Software pp. vol. 40, May 2011.
2. C. Ramasubramanian, R. Ramya, "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013.
3. D. M. Blei, A. Y. Ng, Mi. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, pp. 993-102, 2003.
4. David M. Blei, John D. Lafferty, "Topic Modeling", Journal of Machine Learning Research, Volume 3, pp. 993-1022, Jan 2014.
5. David Blei, M. Andrew Y. Ng, Michael I. Jordan. "Latent Dirichlet Allocation." Journal of Machine Learning Research 3, Jan 2003.
6. Deerwester, S. Dumais, S. Furnas, G. Landauer, R. Harshman Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science 41(6), pp. 391– 407, 1990.
7. F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, 34(1), 2002.
8. Feinerer, K. Hornik, D. Meyer, "Text Mining Infrastructure in R", Journal of Statistical Software, 25(5):54, ISSN 1548-7660, March 2008.
9. K. P. N. V. Satyasree, Dr. J V R Murthy, "Clustering based on Cosine Similarity Measure", International Journal of Engineering Science and Advanced Technology, 2012.
10. K. R. Bindu, Latha Parameswaran, K. V. Soumya, "Performance Evaluation of Topic Modelling Algorithms with an application of Q & A Dataset," International Journal of Applied Engineering Research, ISSN 0973-4562, Vol. 10, No.73, pp. 23-27, 2015.
11. Landauer, T. Foltz, D. Laham, "Introduction to Latent Semantic Analysis". In: Discourse Processes 25, pp. 259–284, 1998.
12. Likas, Aristidis, Nikos Vlassis, Jakob J. Verbeek. "The Global k-means Clustering Algorithm, Pattern Recognition 36.2, 2003.
13. Menaka S, Radha. N, "Text Classification using Keyword Extraction Technique", International

Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013.
14. Sebastiani, F., Text categorization. In Zanasi, A., editor, Text Mining and its Applications, pages 109–129, Southampton, UK: WIT Press, 2005.
15. Srividhya, R. Anitha, "Evaluating Preprocessing Techniques in Text Categorization", International Journal of Computer Science and Application, Issue 2010.
16. Vikasthada, Dr. Vivekjaglan, "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm", International Journal of Innovations in Engineering and Technology, 2013.
17. Jui-Feng Yeh, Chen-Hsien Lee, Yi-Shiuan Tan, Liang-Chih Yu, "Topic Model Allocation of Conversational Dialogue Records", Journal of Machine Learning Research, 2015.

7/20/2019