A Case-Study on Topic Modeling Approach with Latent Dirichlet Allocation (LDA) Model

Abisheka Pon, C. Deisy, P. Sharmila

Thiagarajar College of Engineering, Madurai, TamilNadu, India

Corresponding author: Abisheka Pon, Email: abishekapon@gmail.com

In natural language processing, subject displaying is a sort of factual information models for identifying the points from an enormous assortment of corpus of records. Subject demonstrating is a sort of text-digging device for revelation of stowed away semantic designs in a text body. In the proposed model high layered text informational index named article.csv is handled to acquire primary themes or as often as possible happening subjects for our text information by giving the catchphrases of every point. In this work, a dataset of abstracts are collected from two different domain journals for tagging journal abstracts. The document models are built using Latent Dirichlet Allocation (LDA).Topics thus extracted can be used to get meaningful insights from the text data. In this paper LDA model is used to gives an extra analytical boost for the model. First the data is preprocessed as text data before giving the data to the model as the predictions. Then topic modeling is performed on the preprocessed data by integrating the framework of LDA topic modeling for more optimal classification of topics in the documents.

Keywords: Topic Modeling, LDA, Topics, Natural Language processing, Preprocessing.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 291–299. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-30

1 Introduction

Topic modeling is an unsupervised approach to classifying topics in documents. Data analytics plays a key role in making big data a reality, for group of manufactures of to get simple and accurate results by providing insights in the form of dashboards, graphs, forms, etc. The tokenized words on the left panel of the above image are the result of text preprocessing steps. These words will then be used to construct a model, which may include additional features (more than present in the image).Instead of using the tokenized words obtained from the Bag of Words vectorizer, we can divide the initial document-word matrix into two parts: one for the word per topic and one for the topic per document.

Latent Dirichlet Assignment (LDA)-based topic modeling is used to organise topics by calculating distribution probability over a set of words. Note that topic modeling is not the same as topic classification. Topic classification is a supervised learning approach in which a model is trained using manually performs data with predefined topics. Whereas Topic modeling is an unsupervised approach in which the given data is classified as topics.

After training, the model accurately classifies unseen texts according to their topics. On the other hand, topic modeling is an unsupervised learning approach in which the model identifies the topics by detecting the patterns such as words clusters and frequencies. The outputs of a topic model are:

(i) clusters of documents that the model has grouped based on topics and

(ii) clusters of words (topics) that the model has used to infer the relations.

This is schematically shown in figure 1. Some of the well-known topics modeling techniques are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), and Correlated Topic Model (CTM).

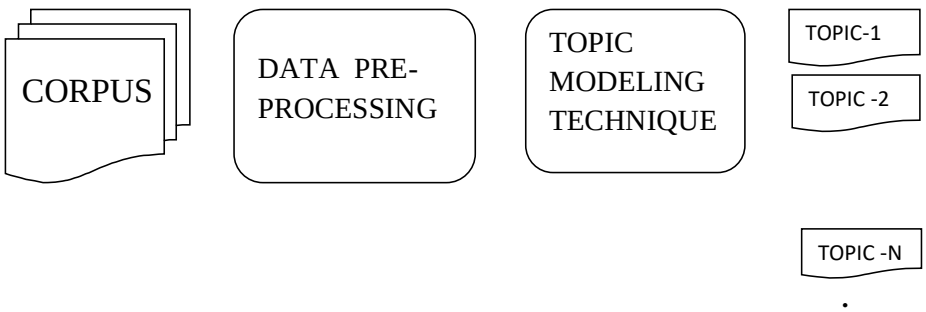

Fig. 1. Framework of Topic Modeling

2 Literature Survey

To break down Point displaying, a few papers like Pröllochs and Feuerriegel [2] proposed a model in which, analytics assumes a key part in making enormous information a reality, assisting associations with obtain straightforward and exact outcomes by giving experiences as dashboards, diagrams, structures, and so forth

Language Handling or Regular Language Handling is a difficult field of study in software engineering to oversee data; this commitment manages the issue of displaying the theoretical consequences of an informational collection. Hamed et al. [8] tracks down that Latent Dirichlet Algorithm (LDA) based theme displaying is utilized to bunch subjects by computing conveyance likelihood over a bunch of words. Note that theme demonstrating isn't equivalent to point characterization.

Discourse acknowledgment, sound to message interpretation, Subsequent to preparing, the model precisely characterizes inconspicuous messages as per their themes. Then again, theme displaying is an unaided learning approach in which the model recognizes the points by distinguishing the examples, for example, words bunches and frequencies referred to by Oberoi [12]. Subject displaying approaches are refined, strong methods utilized in normal language handling for theme disclosure and semantic examination. Latent Dirichlet Algorithm (LDA) based theme displaying is utilized to bunch subjects by computing conveyance likelihood over a bunch of words. Note that theme demonstrating isn't equivalent to point characterization.

Theme arrangement is a managed learning approach in which a model is prepared utilizing physically performs information with predefined subjects. Text mining on text archives is utilized in language handling to coordinate, track down designs, evaluate, and derive the outcome. Language Displaying is a piece of Normal Language Handling that decides the probability of a succession of words. Tong and Zhang [10] Points along these lines removed can be utilized to get significant experiences from the text information. In this paper LDA model is utilized to gives an additional an insightful lift for the model.

3 Topic Modeling

Topic modeling is the method of converting the high dimensional text data and to identify the main topics of our text data by giving the keywords of each topic. Topics might also be a supervised learning which gives an extra analytical boost to the resultant model. A few terms to know in this proposed process are:

- Document: Each text file (article in our case study)
- Corpus: Collection of all the documents
- Dictionary: Collection of mapping of each unique word to a unique index

In Natural Language Processing task, the text data is preprocessed before giving the data to the model as the predictions are as good as the data. In this paper, first given data is pre-processed and then perform topic modeling on the preprocessed data. The two results from the model in this case study are:

(i) Each article to be mapped to a unique topic

(ii) Each topic to have a set of keywords that make that particular topic

3.1 Ingredients To Achieve Topic Modeling

A. Gensim is an open-source python library used for unsupervised topic modeling and natural language processing. It is designed to handle huge collection of text.

B. LDA - Topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) which is use in this proposed paper are a bunch of regular language handling based models used to identify fundamental points in immense assortment of text. There are four types of Topic Modeling.

They are:

a. Latent Semantic Analysis (LSA): To produce semantic substance, LSA creates a vector portrayal of the text.

b. Probabilistic Latent Semantic Analysis (PLSA): To decide the fundamental semantic design of information, PLSA models related data in a probabilistic system.

c. Latent Dirichlet Allocation (LDA): LDA ascertains the closeness between the predetermined source archives and assembles the point circulations of each report.

In this paper, LDA is chosen for two reasons:

(i) It has a great result in the fields of Natural Language Processing.

(ii) It is the most popular statistical and probabilistic text model in the field of machine learning.

d. Correlated Topic Model CTM): CTM identifies subjects in a corpus using a normal logistic distribution.

Internals of LDA

(i) Select the number of topics (input parameter)

(ii) LDA randomly assigns each word to a topic number

(iii) Iterates through each document to calculate the following two probabilities:

Topics in documents: Probability of documents in each topic

Words in topics: Probability of words in each topic

(iv) Reassigns the words to make a topic number and document with the topic number based on the probability scores from step 3, '*number of passes*', an input parameter, times.

Gensim library provides a function for LDA Model which processes the above-mentioned steps with little or no instructions. However, there are various parameters used to obtain the desired output.

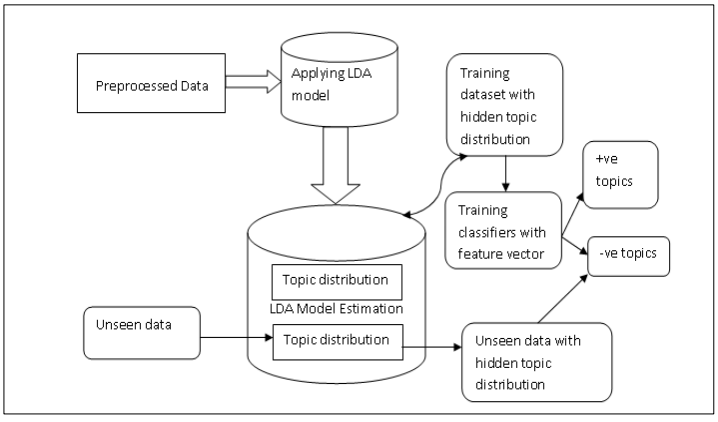

Fig. 2. Workflow of topic modeling with LDA model : Datasets are processed through LDA model to perform topic distributions by classifying it into positive and negative topic related to the datasets

3.2 Preprocess the data

Preprocessing text data is not a single-step process as it contains repetitive words in which the process have to go through lots of cleaning. Here the dataset named article.csv is preprocessed to obtain the density of document with length of words and length of title.

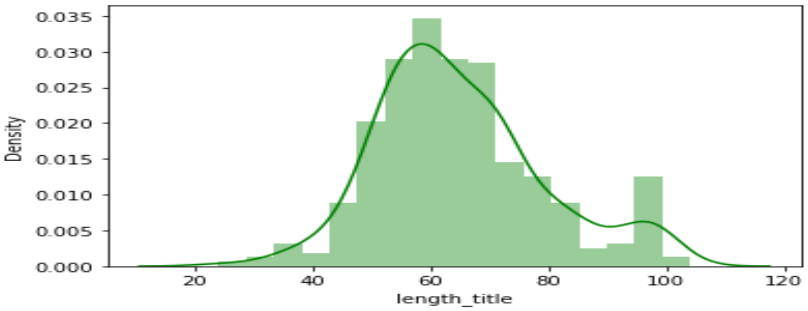

Fig. 3. Plot diagram of Dataset article.csv in the form of length title with the density of words in the documents

The preprocessing is done by the following methods

- **Lowering the case of text** Lowering the case of all the words helps to reduce the dimensions by decreasing the size of the vocabulary.
- **Removing any punctuation marks** will help to treat words like 'hurray' and 'hurray!' in the same way.
- **Stopwords** are commonly found words in a language, such as 'the', 'a', 'an', 'is'. It can be removed because they won't provide any valuable information for our analysis. Removing stop words will reduces the dimensionality of the data. The default function provided for preprocessing are as follows:

  (i) strip_tags(),

  (ii) strip_punctuation(),

  (iii) strip_multiple_whitespaces(),

  (iv)strip_numeric(),

  (v) remove_stopwords(),

  (vi)strip_short(),

  (vii) stem_text()

Data preprocess is done by using the above default filters on the entire text data Creation of dictionary and corpus. This process will create a dictionary and bag of word corpus to pass as input to the model.

Dictionary: Collection of all the unique words.

bow_corpus: Each word collection of data is converted into bag of words. Bag of Words is a transformation of words in the document into a vector by using a dictionary of unique words to get the frequency of each word. Now the dataset is preprocessed as below

Number of unique words in initial documents: 18975

Number of unique words after removing rare and common words: 2720

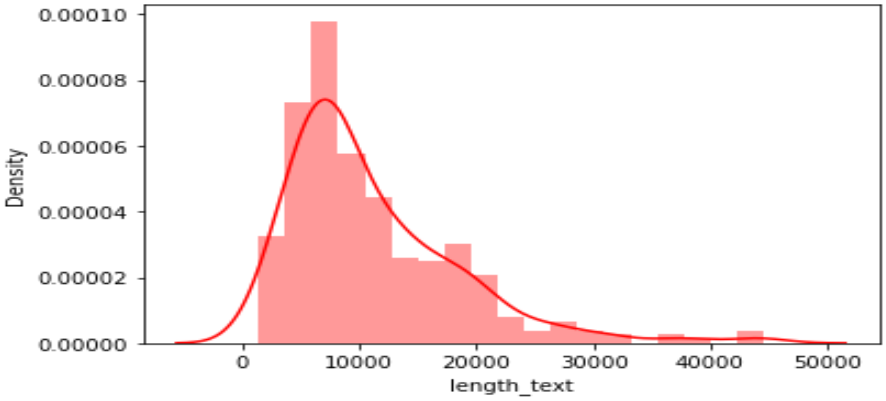

Fig. 4. Plot diagram of processed data to find the length of text with the density of words in document

3.3 Latent Dirichlet Allocation (LDA)

The word **'Latent'** means yet-to-be-found' or hidden topics from the documents. **'Dirichlet'** indicates LDA's assumption that the distribution of topics in a document and the distribution of words in topics are both Dirichlet distributions. '**Allocation'** is the distribution of topics in the document. LDA is an unsupervised text modeling technique for topic mining in NLP.

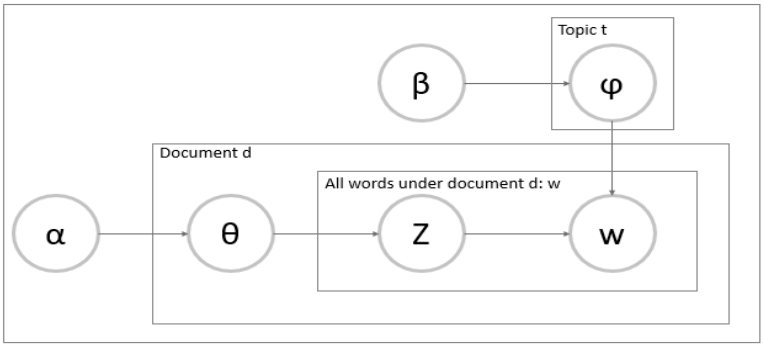

Fig. 5. α is the Dirichlet parameter. is the topic parameter, 1:K are topics, where k is a word distribution, is the Document-topic distribution parameter, d,k is the topic proportion for topic k in document d, and z is the word-topic assignment parameter. The topic assignment for the nth word in document d is zd,n, the observed word is w, and the nth word in document d is wd,n: Blue print of LDA Graphical MODEL Diagram:illustrate how the document is classified as topics and words by Dirichlet distribution and multinominal distribution.

LDA Algorithm:

LDA Algorithm: Generative Algorithm for LDA
Step1: Input: Dataset, K topics, Hyper parameter α and β
Step 2: for All topics k =1,K **do**
// the probability distribution over words for each topic
Step 3: sample mixture words ϕk = Dirβ;
Step 4: end
Step 5: for all documents m = 1,M **do**
// proportion of topics for each document
Step 6: sample mixture proportion θm = Dirα;
// Length of documents in the corpus is normally distributed
Step 7: sample document length Nm = Poissonξ ;
Go to step 4
Step 8: for all words n = 1, Nm in document m **do**
// assign the topic to each word
Step 9: sample topic index Zm,n = Multθm;
// identify the word identity from a probability distribution over words
Step 10: sample term for wordsWm,n =MultϕZm,n ;
Step 11: end
Output: Topic files, Document Topic distribution

Formula for LDA model

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right) \quad (1)$$

Fig. 6. A Graphical distribution of LDA model machine Equation

The datasets are processed through the LDA model to classify it into Dirichlet and multinominal distributions where the first 2 parts represents the Setting of the machine and the last 2 parts are the Gears of the machine to classify the given documents as topics and frequency of topics in the document.

4 Visualization of Results

Data visualization is the concept of placing a visual context so that you can visualize patterns, trends, and correlations discovered using the pyLDAvis feature. pyLDAvis is an open-source Python library that aids in the analysis and creation of highly interactive visualizations of LDA clusters. In this article, we'll look at how to make Topic Modelling Clusters visualisations with LDA and pyLDAvis. Whereas Figure 8 depicts the distribution of topics Each word's frequency is represented by a blue bar, and the local frequency is represented by a red bar.. Each bubble represents a subject. The size of each topic indicates the frequency of occurrence of the topic and the similarity between the topics in a given document.

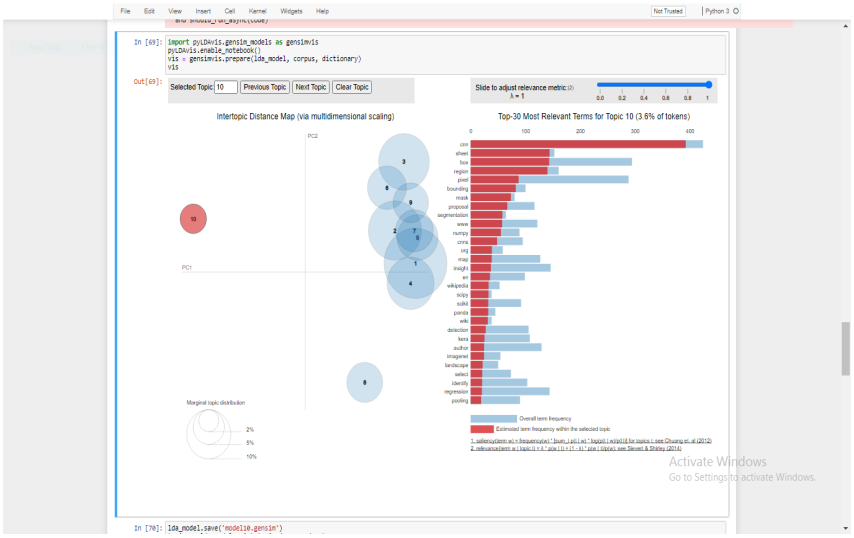

Fig. 8. pyldavis topic visualization of number of topic 10

LDA has been conventionally used to find topics from in text data.

5 Conclusion and Future Work

The goal of proposed paper is to find short descriptions of smaller samples from a collection whose results could be extrapolated to a larger collection while preserving the basic statistical relationships of relevance and apart from detecting topics in text; LDA has been used in Bioinformatics, harmonic analysis for music, and even image object localization. In the future, an upgraded framework of the proposed model will be applied to other corpora in order to get practical insights for further improvement of the suggested framework of LDA topic modelling with various classifiers to achieve more optimal classification results.

References

[1] Thielmann, A. et al. (2020). Unsupervised Document Classification integrating Web Scraping, One-Class SVM and LDA Topic Modeling. *Journal of Applied Statistics,* https://doi.org/10.1080/02664763.2021.1919063

[2] Pröllochs, N. and Feuerriegel, S. (2020). Analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. *Information & Management*, 57(1): 103070.

[3] Qiang, J. et al. (2020). Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*, arXiv:1904.07695.

[4] Ayitey, W. (2010). A Simple Approach to Strategic Management, Methodist Book Depot Ltd. https://www.researchgate.net/publication/279958992

[5] Cheng, C. and Havenvid, M. (2017). Investigating strategy tools from an interactive perspective. *The IMP Journal*, 11(1): 127–149.

[6] Laamanen, T., Mantere, S. and Vaara, E. (2018). Strategy Processes and Practices: Dialogues and Intersections. *Strategic Management Journal*, 39.

[7] Dess, G. G., Eisner, A.B. and Lumpkin, G.T. (2008). Strategic Management: Text and cases, 4th ed., *McGraw-Hill/Irwin, New York, NY*.

[8] Jelodar, H. et al. (2018). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications,* 78: 15169-15211.

[9] Ponweiser, M. (2012). Latent Dirichlet allocation in R. *Vienna University of and Economics*.

[10] Tong, Z. and Zhang, H. (2016). A text mining research based on LDA topic modeling. In *International Conference on Computer Science, Engineering and Information Technology*, 201–210.