Topic Identification in Article using Latent Dirichlet Allocation (LDA)

Bryan Josafat Ferdinand   
*School of Computer Science*  
*Bina Nusantara University*Bandung, Indonesia  
[bryan.ferdinand@binus.ac.id](mailto:bryan.ferdinand@binus.ac.id)

Nabiilah Putri Aviarta   
*School of Computer Science*  
*Bina Nusantara University*Bandung, Indonesia  
[nabiilah.aviarta@binus.ac.id](mailto:nabiilah.aviarta@binus.ac.id)

Harura Cendekia Hasanasultoni   
*School of Computer Science*  
*Bina Nusantara University*Jakarta, Indonesia  
[harura.hasanasultoni@binus.ac.id](mailto:matthew.munthe@binus.ac.id)

*Abstract*— Topic identification used mainly in a text processing based research. In topic identification, there are a lot of methods that can be used to be processed. One of the method is called LDA or Latent Dirichlet Allocation. LDA is the most common method to be used in processing a topic identification with its many functions such as generalizing each word or classifying it by most frequently used word to determined its main topic.

Keywords—Topic Identification, Text Processing, NLP, LDA, Topic Centering

# Introduction

The quantity of articles produced by humans nowadays is growing exponentially every year. In 2023, there are roughly 2 to 3 million online news articles released on the web, not including research papers, journals, printed news, and many other articles.The growth could be a challenging obstacle for many people to find the appropriate article for their needs. Topic modeling is a relevant technique in Natural Language Processing that can be used to extract information from a set of documents or text to discover their topics.Topic modeling often uses LDA (Latent Dirichlet Allocation) as its algorithm. LDA is a generative statistical model, allowing us to observe the similar parts from the observed data. In a simple term, LDA is able to determine the topic from a document by collecting words from the document itself. The goal of this research is to find the topic of any articles, this allows people to spend less of their time rummaging through a bunch of irrelevant articles and making.

# Related Work

Because of the abundance of articles there are on the dataset, it will be difficult for people to find a topic that is relevant and accurate from the document they inputted. There is where Topic identification method came in to help find any topic related to the document given by the users. We all know that a topic identification is based on a text processing which is usually more easy to be interpreted. But, because of its limitation, topic identification still need lots of development and increasement in its method and functionality on identifying a topic in a corpus, for example to find a topic from a news article that uses two different languages [1], and also to identify a topic using any language that isn't in the database of the default languages in the program [12]. Topic modeling can be used in a lot of situations, as example for spam filtering on a comment section [7], or even we can develop a topic modeling by using encyclopedic [4] or even use a correlation [5] to make an automatic Topic identification modeling that can be used for simplifying any problem system. By using Topic identification method, it will help to cut lots of time on finding any main problem on any researches that is needed to be find in a short amount of time which is really crucial if we needed the result as soon as possible.

# Methods

## Dataset

In this project we use a dataset from Kaggle that contains an ID, Abstarct, and also a title of the article. In this dataset, it contains eight thousands nine hundreds and ninety set of articles that includes a 6 different topics that is a Computer Science, Physics, Mathematics, Statistics, Quantitative Biology and Quantitative Finance. The dataset was made by Blesson Densil with the title of “Topic Modelling for Research Paper’ and it is updated around 3 years ago.

## Algorithm and preprocess

Latent Dirichlet Allocation (LDA) is a popular Topic Modeling technique for extracting Topics from a given corpus. The term latent means hidden or hidden, which means that the topics we extract from the data are also “Hidden Topics” in LDA. Dirichlet allocation is after Dirichlet distribution and processing. The Dirichlet process was named by the German mathematician Peter Gustav Lejeune Dirichlet. The Dirichlet process is a probability distribution where the range of this distribution itself is a set of probability distributions.

The LDA model works with the first few steps providing the number of topics searched for in the document set, then each document in the document set is represented in the form of topic probabilities, each topic in the document set is represented as a word probability distribution, then LDA will find the probability distribution of topics and words that best to represent each document, the result of the distribution of the best words and topics to represent each document can be visualized. The steps can be seen in Figure (1.0).

|  |
| --- |
| A diagram of a diagram  Description automatically generated with medium confidence |
| **(1.0)** Diagram LDA model |

We built the LDA model using the genism library. We use 1000 rows of data randomly from the dataset and only use abstract columns. In the initial preprocessing step we do lemmatization and remove stopwords from the data. After the data is lemmatized and the stopwords are removed using WordNetLemmatizer(), word\_tokenize(), and set stopwords in English using stopwords() from the nltk library stopwords here are useful for not lemmatizing the words included in stopwords, then we make a dictionary of data that we have pre-processed using Dictionary() from the gensim library, then the dictionary that has been created is filtered by filtering words that appear in fewer than 5 documents or more than 50% of documents, after the filter process for bow\_corpus uses doc2bow which comes from dictionary on the genism library. After the dictionary is formed, we train the LDA model with a number of topics 5. After training, take the topic from the model that has been formed, then output the results from the topic obtained. The results can be seen in Figure (2.0). after that combine the various titles that have been processed together then use WordCloud() to visualize it. Image visualization results (2.1). do a word cloud of topic plotting for each abstract of the results of plotting images (2.2).

|  |
| --- |
|  |
| **(2.0)** Topic Result |

|  |
| --- |
| A close up of words  Description automatically generated with low confidence |
| **(2.1)** Wordcloud Visualization Results |

|  |
| --- |
| A picture containing text, font, screenshot, graphics  Description automatically generatedA picture containing text, font, screenshot, graphics  Description automatically generated  A picture containing text, font, screenshot, graphics  Description automatically generated  A screenshot of a computer  Description automatically generated with low confidence  A picture containing text, font, screenshot, graphics  Description automatically generated |
| **(2.2)** Wordcloud Visualizations from some of the abstract on the dataset. |

We create simple\_preprocess from the genism library and stopwords from the nltk library to remove the stopwords with the stopwords we use ['from', 'subject', 're', 'edu', 'use'] then these words will be matched against what dataset if the dataset contains these words then the words will be discarded, form a corpus of word data that has removed the stopwords, form a model with the number of topics 10 using genism.model.LdaMulticore(). Perform visualization using the pyLDAvis library by entering the LDA model, corpus, and dictionary which are the LDA model itself to see the words that appear frequently in the dataset and perform correlation analysis between words. The visualization results can be seen in figure (3.0).

|  |
| --- |
|  |
| **(3.0)** pyLDAvis Visualization Result |

## Experiment Result

From the results of the visualization in figure (3.0) it can be seen that the 3 parts of the circle that are close to each other, each circle that is close to each other means that words in the topic often appear and have a correlation. From the results we got, there are 3 topics that have words that rarely appear and are not correlated with each other, while the other 7 topics are divided into 3 sections, each of which is correlated, such as topics 1, 3, and 5, in this section there are the words model, data, result. and approaches that often arise from these 3 topics and the correlation between the words is quite close. For topics 8, 10 only have words in common with a small number and correlation. For topics 2 and 6, this is quite correlated and the words that frequently appear on the two topics are quite similar..

# Conclusion

Topic modeling using the Latent Dirichlet Allocation (LDA) method is a powerful approach for analyzing and understanding textual data. LDA helps in uncovering hidden topics within a collection of documents, providing insights into the content and structure of the text. By reducing the dimensionality of the text data, LDA facilitates exploratory analysis, document clustering, and content recommendation based on relevant topics.

The application of LDA in topic modeling enables researchers and users to gain a deeper understanding of large text collections, identify patterns and trends, and make meaningful discoveries. It aids in information retrieval, document organization, and content recommendation systems. LDA provides a framework for exploring and extracting valuable insights from textual data, making it a widely used method in various domains, including natural language processing, information retrieval, social media analysis, and content analysis.

However, it is important to note that LDA has certain limitations. It assumes that each document is a mixture of topics, which may oversimplify the complexity of real-world text. It also requires the specification of the number of topics in advance, which can be challenging and subjective. Additionally, LDA does not consider the temporal dynamics of topics or the semantic relationships between words.

Overall, topic modeling using the LDA method is a valuable tool for understanding and organizing textual data. It enables researchers and analysts to extract meaningful insights, discover hidden patterns, and facilitate decision-making in various applications. As the field of natural language processing and text analysis continues to advance, LDA and other topic modeling techniques are likely to evolve and be further enhanced to address the challenges and complexities of analyzing textual data.

##### References

[1]   A.Srivastav and S.Singh, “Proposed Model for Context Topic Identification of English and Hindi News Article Through LDA Approach with NLP Technique”, India : 14 Aug.2021

[2]  G.Manias, A.Mavrogiorgou, A.Kiourtsis, D.Kakomitas, and D.Kyariazis, “Real-Time Kafka-Based Topic Modeling and Identification of Tweets”, China : Shanghai, 25 Jan.2022

[3]  H.S.Baghdadi and B.R.Malancon, “An Automatic Topic Identification Algorithm”, Malaysia : 2021

[4]   K.Courser, R.Mihalcea and W.Moen, “Using Encyclopedia Knowledge for Automatic Topic Identification”, USA : Jun.2009

[5]  S.Slater, R.Baker, M.Almeda A.Bowers and N.Heffernan, “Using Correlational Topic Modelling for Automated Topic Identification in Intelligent Tutoring system”, USA : 2017.

[6]   R.Adaikkalavan, L.Elkhalifa, and Y.Aslandogan, “Topic Identification Through Ontology-based Concept Generalization”, USA : 2003.

[7]   C.Rădulescu, M.Dinsoreanu and  R.Potolea, “Identification of spam comments using natural language processing techniques”, Romania : 2014

[8]  K.Nokkaew, and R.KongkaChandra, “Keyphrase Extraction as Topic Identification Using Term Frequency and Synonymous Term Grouping”, Thailand : 2018

[9]  C.May, F.Ferraro, A.McCree, J.Wintrode, D.Garcia-Romero, and B.V.Durme, “Topic Identification and Discovery on Text and Speech”, Portugal : Lisbon.2015

[10]  V.Stoyanov and C.Cardie, “Topic Identification for Fine-Grained Opinion Analysis”, Manchester : 2008.

[11]  X.Wang, Y.Wang, W.Zuo and G.Cai, “Exploring Social Context for Topic Identification in Short and Noisy Texts”, China : 2015.

[12]  M.Abbas and K.Smaili, “Comparison of Topic Identification methods for Arabic language”, Bulgaria : Borovets.2005.